BMC Bioinformatics

Research

Methods for comparative metagenomics

Daniel H Huson*[1], Daniel C Richter[1], Suparna Mitra[1], Alexander F Auch[1] and Stephan C Schuster[2]

Address: [1]Center for Bioinformatics ZBIT, Tübingen University, Sand 14, 72076 Tübingen, Germany and [2]310 Wartik Laboratories, PennState University, Center for Comparative Genomics, Center for Infectious Disease Dynamics, University Park, PA 1803, USA

Email: Daniel H Huson* - huson@informatik.uni-tuebingen.de; Daniel C Richter - drichter@informatik.uni-tuebingen.de; Suparna Mitra - mitra@informatik.uni-tuebingen.de; Alexander F Auch - auch@informatik.uni-tuebingen.de; Stephan C Schuster - scs@bx.psu.edu

* Corresponding author

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, **10**(Suppl 1):S12 doi:10.1186/1471-2105-10-S1-S12

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S12

Abstract

Background: Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification, and can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. Increasingly, there is a need for new ways of comparing multiple metagenomics datasets, and for fast and user-friendly implementations of such approaches.

Results: This paper introduces a number of new methods for interactively exploring, analyzing and comparing multiple metagenomic datasets, which will be made freely available in a new, comparative version 2.0 of the stand-alone metagenome analysis tool MEGAN.

Conclusion: There is a great need for powerful and user-friendly tools for comparative analysis of metagenomic data and MEGAN 2.0 will help to fill this gap.

Background

Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. Although it is clear that communities of microbes play a vital role in such systems, a more detailed understanding is only beginning to emerge. A main promise of metagenomics is that it will accelerate drug discovery and biotechnology by providing new genes with novel functions.

Currently, the key approach used in metagenomics is large-scale sequencing of environmental samples. The recent development of ultra-high throughput sequencing

technologies [1,2], which do not require cloning or PCR amplification, and can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects, see [3,4]. The analysis of such datasets is aimed at determining and comparing the biological diversity and the functional activity of different microbial communities.

Computationally, species identification relies on the use of reference databases or reference phylogenies that contain of sequences of known origin and gene function. The most prominently used databases are the NR and NT databases [5]. Unfortunately, substantial database biases toward model organisms present a major hurdle for metagenomic analysis, and in a typical metagenome dataset as much as 90% of the reads may exhibit no similarity to any known sequence. However, this problem is beyond the scope of this paper. Early 2007, our group released and published the first publicly available, stand-alone analysis tool for metagenomic data, called MEGAN [6,7]. We initially developed this tool to analyze the microbial community present in a sample of mammoth bone [8]. MEGAN takes as input the result of a BLAST [9] comparison of a set of metagenomic reads against one or more reference databases and produces as output a taxonomical analysis of the sample, obtained by assigning the reads to different nodes in the NCBI taxonomy using an "LCA-algorithm".

As an exploration tool designed and optimized to run on a laptop, MEGAN complements other systems and resources for metagenome analysis, which are offered in the form of databases, web portals and web services, such as [10-14].

MEGAN now has over 400 registered users working in many different biological labs around the world. It is routinely used at the Joint-Genome-Institute (JGI) both in quality control and also to provide initial analyses of newly sequenced datasets. Other users include researchers at the J.C. Venter Institute studying viral populations. In a recent publication [15], we demonstrate how to use the software for meta-transcriptomics, as well.

Increasingly, the emphasize of metagenome analysis is shifting from species and functional identification for individual datasets toward comparative analysis. This paper addresses the latter issue and provides solutions to questions such as: Given two or more metagenome datasets, how similar or different are their taxonomical and functional profiles? Are observed differences statistically significant? Have enough reads been sequenced, i.e. what is the current "rate of discovery" as a function of the number of reads sequenced? In the following section, we will discuss some new ideas for analyzing individual metagenome datasets. Then, we will focus on new comparative methods. Finally, we will illustrate the application of the methods in two comparisons, one comparing the contents of a human gut [16] with the contents of a mouse gut [17] and the other comparing a soil sample [18] with a recent marine sample [19].

The ideas presented in this paper are all quite simple and unsophisticated. The main merit of this work lies in the integrated implementation of the methods in the form of a very robust and user-friendly program, which is easily used by biologists. The implementation goes well beyond the hastily written "proof of concept" implementations that so often accompany method papers. We are currently beta-testing version 2.0 of the MEGAN software, which implements all ideas presented in this paper. The latest beta version can be obtained from our website at [20].

Methods

One goal of metagenome analysis is to determine the taxonomical content of a dataset [6,21]. There are two main approaches toward doing this.

The *phylogenetic approach* is based on carefully chosen genes that are believed to provide robust phylogenetic information [22,23], see [21,24]. When randomly-targeted sequencing is used, only a small fraction of the sequences will correspond to such phylogenetic markers [21,25]. Often, universal primers are employed to specifically target the phylogenetic markers. The DNA sequences obtained are usually aligned into precomputed reference alignments and placed into precomputed reference trees, using fast heuristics and then taxonomical placements are deduced from this.

The *taxonomical approach* places reads directly into the NCBI taxonomy, based on the similarity of the reads to sequences in one or more reference databases. As randomly sequenced reads will exhibit very different levels of evolutionary conservation, it is important to make use of all ranks of the NCBI taxonomy, placing more conserved sequences higher up in the taxonomy (i.e. closer to the root) and more distinct sequence onto nodes that are more specific (i.e. closer to the leaves, which represent species and strains). This can be done using the *LCA algorithm* and is the basis of the MEGAN program.

In summary, the LCA algorithm works as follows. A sequencing read is compared against a database of reference sequences, such as the NCBI NR database, and the taxon information of significant matches is extracted and mapped onto the leaves of the NCBI taxonomy. The leaves of the NCBI taxonomy represent different species and strains. The LCA algorithm computes the lowest common ancestor of all these hits, which will correspond to some

higher-order taxon, and will then assign the read to that taxon. In this way, species-specific sequences will be assigned to the leaves or specific taxa, whereas sequences that are conserved among different species, or that are susceptible to horizontal gene transfer, will be assigned to taxa of less-specific rank. See the original paper [6] for more details.

Both approaches have different advantages and drawbacks. The phylogenetic approach can use established phylogenies that are well understood and targeted sequencing provides much more informative data per sequencing run. However, a commonly acknowledged draw-back is that the "universal primers" employed may produce only a subset of the true spectrum of different sequences. On the other hand, random sequencing is often used in metagenomics to analyze the gene content of a community and then the taxonomical approach can make full use of the data and can be complemented by a phylogenetic approach.

Rate of discovery

One important question is whether the level of sequencing performed for a given sample is sufficient to capture the most abundant taxa. This can be addressed by plotting the *discovery rate* of a dataset, which is obtained by repeatedly selecting random subsamples of the dataset at 10, 20 ..., 90% of the original size, and then plotting the number of taxa predicted by the LSA algorithm, see Figure 1. This graph can be used to estimate (roughly) how many additional species are likely to be discovered if one were to increase the number of reads by a factor of two, say.

In this, to estimate the number of species, one might first consider counting the number of leaves of the taxonomy to which reads have been assigned. However, this number may be confounded by the presence of different strains and isolates. To avoid this problem, in our implementation in MEGAN 2.0 we use the number of *strongly supported* nodes as a proxy for the number of species. We say that a node v in the NCBI taxonomy is *strongly supported at level t*, where t is a small number (≈ 5), if v has been assigned t or more reads and no node below v has that property.

Functional assessment

In a functional analysis, the goal is to determine which types of genes are available at what relative levels of abundance. Such an analysis can be based on sequences obtained by random sequencing either of the genomic DNA in a metagenome, or (reverse transcribed) RNA. In the former case, the coding potential is analyzed, whereas in the latter case, the focus is on gene expression. A general strategy is to compare the reads against reference databases of gene sequences such as COG [26] and SEED [11].

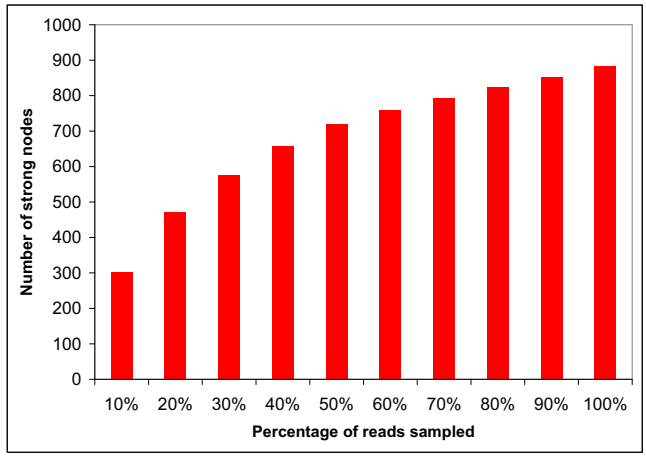

Figure 1
A discovery rate plot computed by MEGAN 2.0 for the mouse gut dataset. The *x*-axis represents the percentage of reads subsampled from the total dataset and the *y*-axis represents the number of strong nodes (with $t = 5$) computed by the LCA algorithm, approximating the number of identified species. The datapoint at $10 \times t$ % is based on t independent runs.

A number of sequences available in the NR database are annotated by COG [26] identifiers. Hence, after BLAST comparison of a metagenomic dataset with the NR database, a first analysis of the types of genes present in the dataset can be performed by extracting all COG identifiers from the BLAST hits and then summarizing the relative abundances of the different COG categories, see Figure 2.

Meta-data analysis

The result of a taxonomical analysis can be enhanced by using "meta-data" to summarize the identified species. For example, the "Prokaryotic Attributes Table" (obtainable from the NCBI website) lists attributes of microbes that describe their cellular features, environment, temperature, pathogenicity and relevance for diseases. A summary of an analysis based on such attributes is shown in Figure 3.

Taxonomy-guided capture of reads

Once a first analysis has been performed and reads have been assigned to taxa, it is often desirable to be able to identify and capture all reads that have been assigned to one part of the NCBI taxonomy, not only to a specific species, but also to a class, genus or other rank of the taxonomy. This is very useful, for example, when performing additional analysis such as determining the GC-content for a collection of taxa, or for sequence assembly purposes.

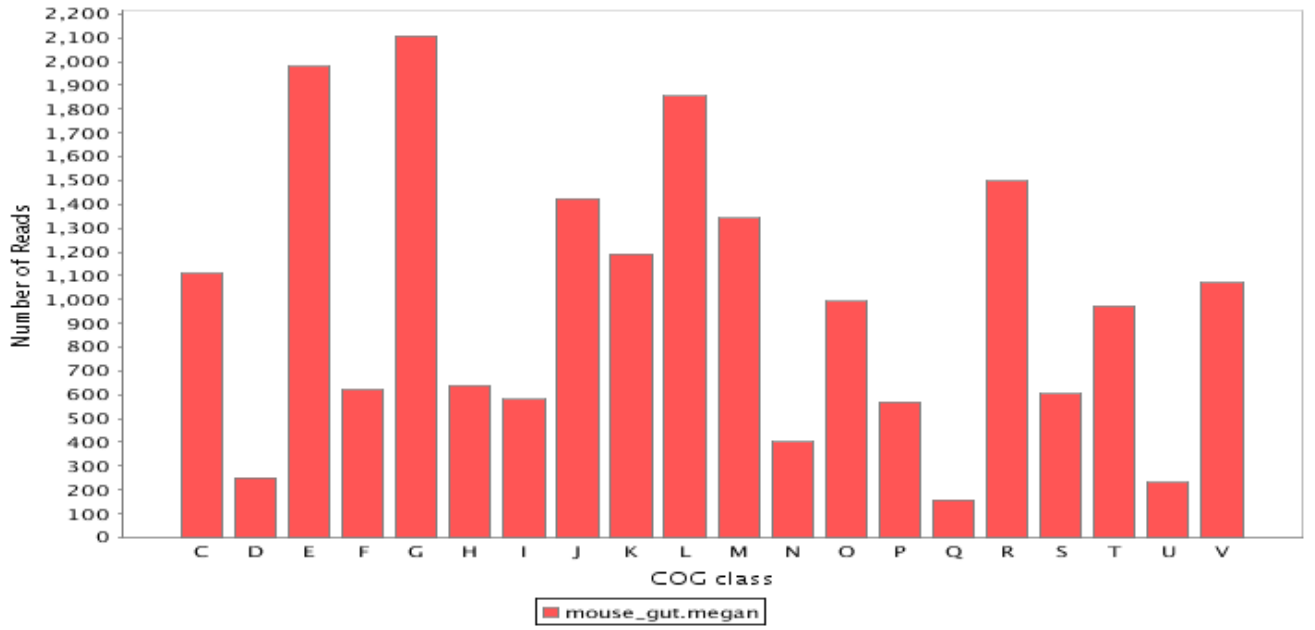

Figure 2
A classification of all COGs determined in the mouse gut sample.

Comparative visualization

In a comparative analysis, different datasets are brought together and compared for taxonomical and functional content. 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, as shown in Figures 4 and 5. In such a view, each node in the NCBI taxonomy is shown as a pie chart indicating the number of reads (normalized, if desired) from each dataset that have been assigned to that node. An important feature is the ability to interactively collapse or expand the presented tree at different levels of

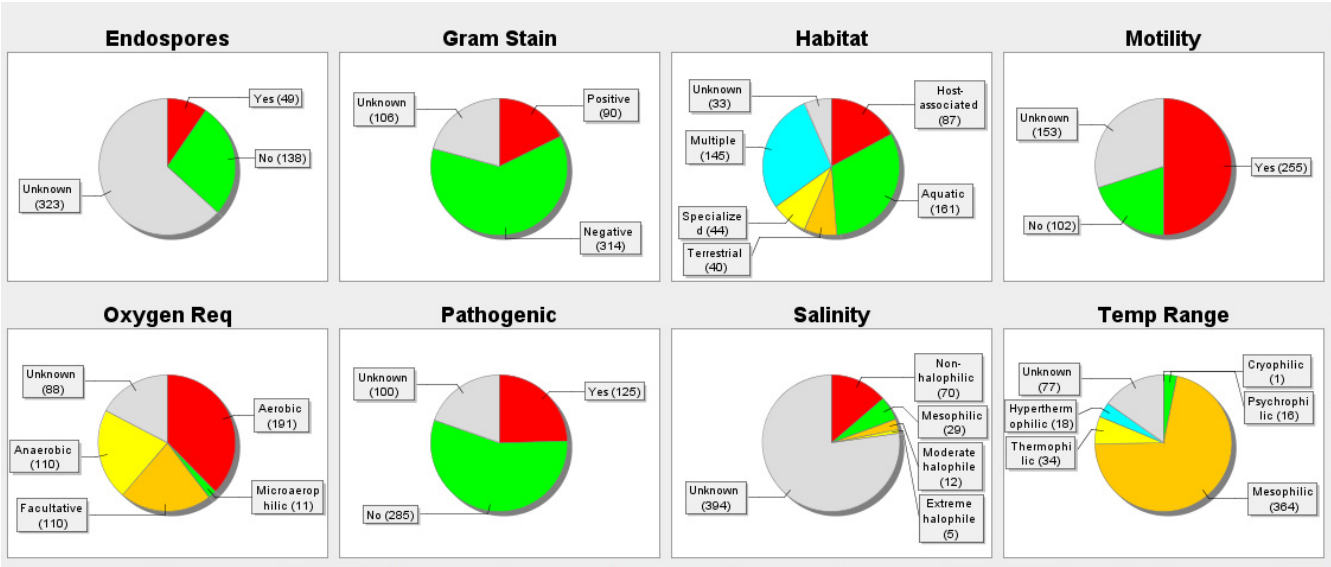

Figure 3
Summary of the microbial attributes of the soil dataset based on the NCBI's "Prokaryotic Attributes Table". In each pie chart, the number of classified species having the indicated property is displayed.

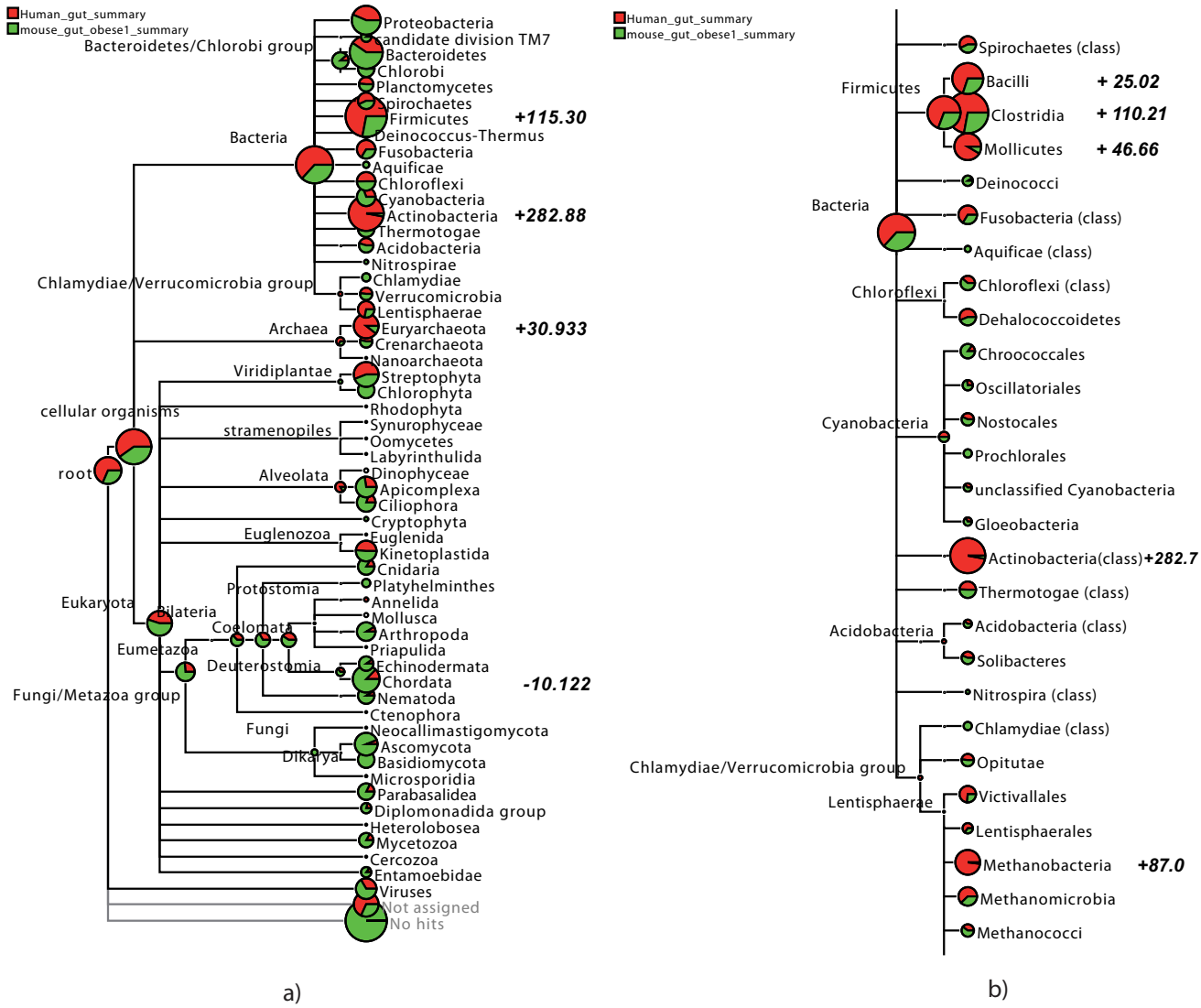

Figure 4
Two multiple-comparative tree views of a human gut metagenome [16] shown in red and a mouse gut metagenome [17] shown in green, as computed by MEGAN 2.0, using normalized counts. In (a), we show an overview of the taxonomy down to the phylum level, whereas in (b) we display a part of a class-level analysis. In bold we show the support values as listed in Table 1.

the taxonomy, so as to be able to start at a high-level view and then to drill down to a low-level comparison.

For publication purposes, the ability to interactively setup and generate different types of summaries using bar and pie charts, and also heat maps for many-way comparisons, are important. We are developing an interactive and fully customizable chart viewer for MEGAN 2.0 that allows one to extract a number of different comparisons directly from the multiple comparison tree view. For example, one can generate a bar chart summarizing the number of reads assigned at any desired rank of the NCBI taxonomy, see Figure 6.

Statistical significance

Comparative visualizations are useful to obtain an impression of how two datasets differ. For a more detailed analysis, one requires information on the statistical significance of observed differences, see Table 1 and 2. To this end, we have adapted a test developed for comparing curated subsystems in metagenomic data [27]. This test

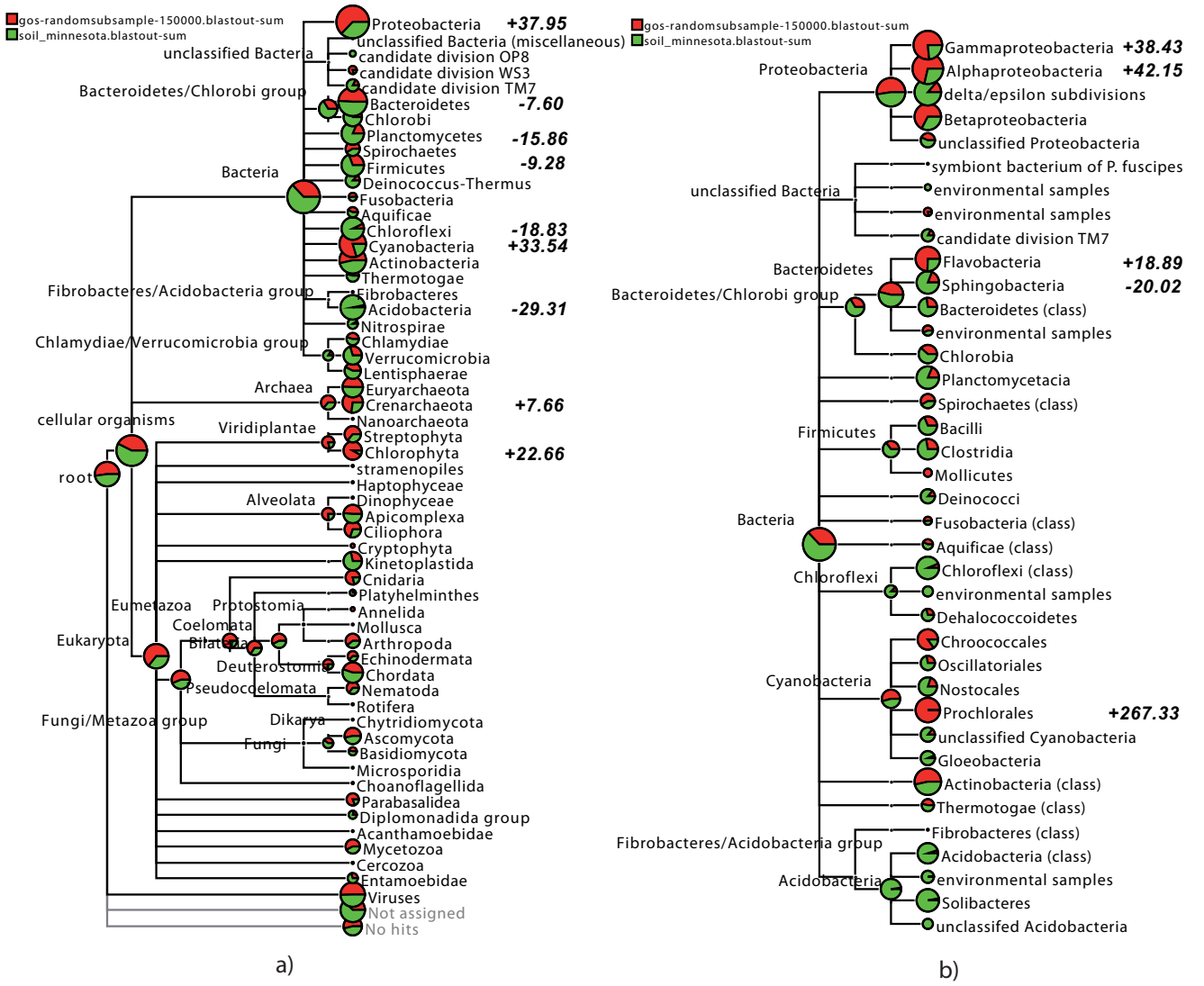

Figure 5
A multiple-comparative tree view of a soil metagenome [18] shown in green and a marine metagenome [19] shown in red, as computed by MEGAN 2.0. In (a), we show an overview of the taxonomy down to the phylum level, whereas in (b) we display a part of a class-level analysis. In bold we show the support values as listed in Table 2.

uses bootstrapping to determine for which subsystems a difference in counts is significant. This can be extended by defining a support value as the proportion of deviation given by $\frac{2|M-P_{50}|}{P_{95}-P_5}$, based on the average difference M of pairs of values sampled from the two different datasets and the percentile values P_x obtained by resampling from the *same* dataset. In MEGAN 2.0, it will be possible to apply this test to any level of the NCBI taxonomy.

Dealing with very large datasets

To be able to deal with ever larger, multiple datasets on a computer with a limited amount of main memory, MEGAN 2.0 can perform the analysis of any given dataset in a new *summary* mode, in which the analysis is performed on-the-fly and none of the read or match data are loaded into memory. A summary file obtained in this way describes only how many reads were assigned to each taxon, and thus the size of such a file is independent of the size of the original input dataset.

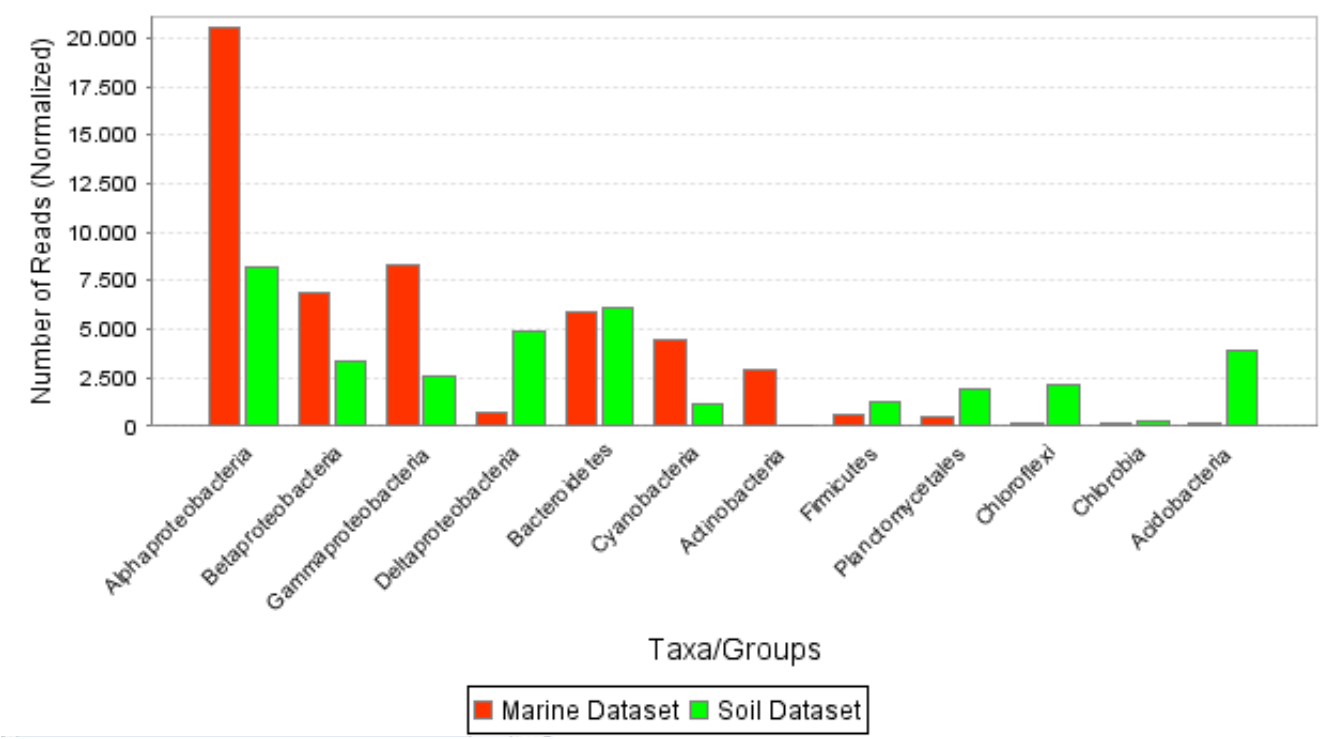

Figure 6
A summary of the comparison of the marine (red) and soil (green) datasets, generated at different taxonomical ranks.

In an ongoing study, we are using a beta version of MEGAN 2.0 to analyze datasets containing a million or more reads. As another example, a BLAST file generated for the soil sequences discussed in Section is 53 GB in size and can be parsed in less than 40 minutes on a laptop. Once parsed in this way, the data can then be saved in a summary format that can be reopened in seconds.

Results and discussion

In this section we will illustrate some of the methods described above, using a number of publicly available datasets. We first consider two recent metagenomic datasets, one taken from a human gut (approx. 145, 000 reads using Sanger sequencing) [16] and the other from the gut of an obese mouse (approx. 675, 000 reads using 454 sequencing) [17].

Using the mouse gut dataset, we show a discovery rate analysis in Figure 1. From this, we can estimate that doubling the number of sampled read sequences will only lead to the discovery of approximately 50 additional taxa. This result, therefore suggests that this particular metagenome consists of roughly 950 predominant taxa, a large majority of which are already identified using only half of the reads. This example illustrates that the assessment of the species discovery rate per number of reads may be highly beneficial for the design and economy of any project with unknown species composition. Cost savings

Table 1: Significant differences in the comparison of human gut and mouse gut metagenomes. The five most statistically significant differences in numbers of reads assigned to taxon classes in the comparison of a human gut [16] and obese mouse gut [17] metagenomes. A positive support (proportion of deviation) indicates that the difference is in favor of the human gut dataset, whereas a negative sign indicates the opposite.

	Comparison of human and mouse gut datasets				
Rank	1	2	3	4	5
Phylum level Support	Actinobacteria +282.88	Firmicutes +115.30	Euryarchaeota +30.93	Chordata -10.12	Ascomycota -6.96
Class level Support	Actinobacteria +282.70	Clostridia +110.21	Methanobacteria +87.0	Mollicutes +46.66	Bacilli +25.01

Table 2: Significant differences in the comparison of marine and soil metagenomes. The five most statistically significant differences in numbers of reads assigned to taxon classes in the comparison of marine [19] and soil [18] metagenomes. A positive support (proportion of deviation) indicates that the difference is in favor of the soil dataset, whereas a negative sign indicates the opposite.

Comparison of marine and soil datasets					
	1	2	3	4	5
Phylum level Support	Proteobacteria +37.95	Cyanobacteria +33.54	Acidobacteria -29.31	Chlorophyta +22.67	Chloroflexi -18.83
Class level Support	Prochlorales +267.33	Thermoprotei +82.36	Oligohymenophorea +52.36	Aconoidasida +50.36	Prasinophyceae +52.33

are likely to be realizable for any project that proves to have a much lower taxonomical diversity than assumed at the outset [28].

In Figure 4, we show a multiple-comparative tree view of the human gut and mouse gut metagenomes, using normalized counts. The analysis is based on a BLASTX comparison of all reads against the NR database. At first glance, there appears to be many nodes at the taxonomical rank of class for which the number of assigned reads differs substantially. However, using the described statistical test, we see that there are only a few statistically significant differences, listed in Table 1.

Because the two datasets were obtained using different sequencing technologies, it may be that some adjustments to the analysis will have to be made to account for the different read-length distributions of multiple data sets. This is ongoing work.

We now briefly discuss the five main differences identified in Figure 4 and Table 1. As expected, Actinobacteria are more dominant in the human gut, manifested through a high abundance of *Bifidobacterium longum*, *B. adolescentis* and *Collinsella aerofaciens ATCC 25986*. All three species are known to be normal inhabitants of the human intestine.

Also, Firmicutes are more dominant in the human gut, mostly in the form of Clostridia, Lactobacillales and Mollicutes. Clostridia and Lactobacillales can live in intestinal tracts of animals and humans, however it is not clear why the levels of abundance differ in the two datasets. The human dataset also contains *Eubacterium dolichum DSM 3991* whose presence has previously been established by its isolation from the human gut flora. *Mesoplasma florum* is considered a commensal strain in humans and an animal parasite. A striking contrast between the two datasets also seems to be the high abundance of Euryarchaeota/Methanobacteria. As previously reported, the main representative of this group is *Methanobrevibacter smithii*, a well-known archaeal inhabitant of the human gut, see [16,29].

In our experience, the class of Chordata is always problematic in this type of metagenomic analysis. This is most likely due to the high complexity and large sequence space covered by higher eukaryote and especially vertebrate genomes. This is further aggravated by database biases toward model organisms and the problem of false annotation of vertebrate genetic elements.

The amount of hits mapped to Ascomycota was significantly higher in the mouse gut probe, mostly reads assigned to yeast species like *Saccharomyces* and *Candida*. It is well known that these yeast species can be found in caeca of mouse [30] and rat [31]. As stated in [17], the mouse gut probe was extracted from its caecum, whereas the human probe was taken from distal gut.

Interestingly, the proportion of mouse gut reads that exhibit no hits to the NR database is much higher than for the other dataset. This probably reflects the different read lengths produced by the employed sequencing technologies (Sanger for the human gut sample, 454 for the mouse one). An additional potential explanation may be that there is a bias in NR database that favors human endosymbionts and parasites. A basic functional analysis of the mouse dataset can be obtained from the COGs present in the NR database. We show the result of such an analysis in Figure 2.

As a second example, we analyze a set of approx. 140, 000 reads extracted from a soil sample using Sanger sequencing [18] and then compare this to a small subset of approx. 145, 000 reads of the *Global Ocean Survey* dataset, [19] obtained using Sanger sequencing technology. The analysis is based on a BLASTX comparison of all reads against the NR database. In Figure 5 we show a multiple-comparative tree view of the two datasets.

We now briefly discuss some of the main differences summarized in Figure 5 and Table 2. Our analysis reiterates the well-known fact that soil metagenomes are significantly more complex than marine ones. However, this diversity is underrepresented in current reference data-

bases. Therefore, more reads are assigned to the proteobacterial phylum in the marine dataset than in the soil one, in particular *Pseudomonas mendocina ymp*, *Shewanella* (aquatic bacteria), and some unclassified gamma proteobacteria, such as marine gamma proteobacteria HTCC2080, HTCC2143 and EBAC20E09. Differences in the number of reads assigned to Cyanobacteria can be attributed to *Synechococcus* and *Prochlorococcus marinus* which both belong to the most abundant bacterial species in marine surface water [21].

Significantly more reads are assigned to Acidobacteria in the soil dataset, most mapping to *Solibacter usitatus Ellin6076*, a soil bacterium. However, since the Acidobacteria are a very divergent class of taxa, this discrepancy could be due to the low amount of currently sequenced species within this group. The fact that reads hitting Chlorophyta are more present in the marine dataset is due to the number of hits to Prasinophyceae, which are marine algae. The existence of fresh water variants may explain the small number of hits in soil. Reads that match Chloroflexi are found more often in the soil than in the marine dataset, in particular *Herpetosiphon aurantiacus ATCC 23779*, which was originally isolated from a lake in Minnesota, the same state from which the soil sample was taken. The fact that Thermoprotei are favored by the marine sample is due to reads assigned to *Nitrosopumilus maritimus SCM1*, which is a mesophilic (not thermophilic) salt-water bacterium. The groups Oligohymenophorea and Aconoidasida both belong to the phylum Alveolata comprising a very divergent group of unicellular eukaryotes, some of them are capable of photosynthesis. Accordingly, the marine dataset contains significantly more reads of these eukaryotic clades than the soil dataset. Interestingly, most hits within Aconoidasida belong to the taxon *Plasmodium falciparum*, the pathogen of malaria. Since it is known that *P. falciparum* possesses a chloroplast-like organelle which presumably was derived in a common ancestor of Apicomplexa [32], a possible explanation may be that these reads come from a marine species that is closely related to the Aconoidasida but not well represented in the NR database.

In Figure 3 we analyse the microbial attributes content of the soil dataset. Of 564 microbes identified in the dataset, 510 where found among the ≈ 1500 prokaryotes currently listed in the NCBI "Prokaryotic Attributes Table". Somewhat disappointingly, this profile of attributes differs only insignificantly from the one computed for the marine dataset (not shown), most likely due to database biases.

The comparison of the soil and marine datasets can be performed at different levels of the NCBI taxonomy and represented as bar charts, see Figure 6.

Conclusion

Comparative metagenomics is a fast growing field and novel tools are required to support comparative analysis of multiple metagenomic datasets. In this paper we have discussed a number of new techniques that address this issue. These will all be made available in a new version 2.0 of MEGAN.

We anticipate that a metagenomic project will routinely look at 5–10 different samples, each consisting of 100,000 or more reads. Once the data has been compared against appropriate reference databases, MEGAN 2.0 can be used for fast and user-friendly comparative analyses of datasets of this size, providing graphical support for the publication process.

A number of papers on new metagenome datasets that employ MEGAN as the primary analysis tool are in preparation. Future improvements of the program will include the use of the GO gene ontology [33] to classify functional content and the implementation of more statistical tools for comparing different datasets. MEGAN 2.0 is currently being incorporated into the CAMERA metagenomics web portal [13].

Availability

The datasets discussed are available at [34]. Installers for common operating systems for MEGAN 2.0 will be available at [20].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DHH, DCR and SM implemented the comparative methods for MEGAN 2.0. AFA, DCR and SM curated and analyzed the datasets. All authors participated in writing the manuscript. DHH and SCS supervised the study.

Acknowledgements

This work was funded in part by Deutsche Forschungsgesellschaft (DFG).

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S1

References

1. Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen YJ, Chen Z, Dewell SLD, Fierro J, Gomes X, Godwin B, He W, Helgesen S, Ho C, Irzyk G, Jando S, Alenquer M, Jarvie T, Jirage K, Kim JB, Knight J, Lanza J, Leamon J, Lefkowitz S, Lei M, Li J, Lohman K, Lu H, Makhijani V, McDade K, McKenna M, Myers E, Nickerson E, Nobile J, Plant R, Puc B, Ronan M, Roth G, Sarkis G, Simons J, Simpson J, Srinivasan M, Tartaro K, Tomasz A, Vogt K, Volkmer G, Wang S, Wang Y, Weiner M, Yu P, Begley R, Rothberg J: **Genome sequencing in microfabricated high-density icolitre reactors.** *Nature* 2005, **437(7057):**376-380.

2. Bentley D: **Whole-genome re-sequencing.** *Current Opinion in Genetics & Development* 2006, **16:**545-552.
3. **GOLD: GenomesOnLine Database** [http://www.genomesonline.org]
4. Bernal A, Ear U, Kyrpidesa N: **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.** *Nucleic Acids Res* 2001, **29:**126-127.
5. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D: **GenBank.** *Nucleic Acids Res* 2005, **1(33 Database):**D34-38.
6. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17(3**377-386 [http://dx.doi.org/10.1101/gr.5969107].
7. Huson DH, Auch AF, Qi J, Schuster SC: **Metagenome analysis using MEGAN.** In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference, Volume 5 of Series on Advances in Bioinformatics and Computational Biology* Edited by: *Sankoff D, Wang L, Chin F. Imperial College Press*; 2007:7-16.
8. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RDE, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC: **Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA.** *Science* 2006, **311(5759**392-394 [http://dx.doi.org/10.1126/science.1123360].
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.
10. Lozupone C, Hamady M, Knight R: **UniFrac – An Online Tool for Comparing Microbial Community Diversity in a Phylogenetic Context.** *BMC Bioinformatics* 2006, **7:**371.
11. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33(17**5691-5702 [http://dx.doi.org/10.1093/nar/gki866].
12. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavrommatis K, Ivanova N, Kyrpides N: **The integrated microbial genomes (IMG) system.** *Nucleic Acids Research* 2006:344-348.
13. Seshadri R, Kravitz S, Smarr L, Gilna P, Frazier M: **CAMERA: A Community Resource for Metagenomics.** *PLoS Biology* 2007, **5(3):**.
14. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J: **Phylogenetic classification of short environmental DNA fragments.** *Nucleic Acids Res* 2008, **36(7):**2230-2239.
15. Urich T, Lanzen A, Qi J, Huson DH, Schleper C, Schuster SC: **Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome.** *PLoS ONE* 2008, **3:**e2527 [http://dx.doi.org/10.1371/journal.pone.0002527].
16. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312(5778**1355-1359 [http://dx.doi.org/10.1126/science.1124234].
17. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006, **444(7122**1027-1031 [http://dx.doi.org/10.1038/nature05414].
18. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative Metagenomics of Microbial Communities.** *Science* 2005, **308:**554-557.
19. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific.** *PLoS Biol* 2007, **5(3**e77 [http://dx.doi.org/10.1371/journal.pbio.0050077].
20. **MEGAN website** [http://www-ab.informatik.uni-tuebingen.de/software/megan/]
21. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, Smith HO: **Environmental Genome Shotgun Sequencing of the Sargasso Sea.** *Science* 2004, **304(5667):**66-74.
22. Woese CR: **Bacterial Evolution.** *Microbiol Rev* 1987, **51:**221-272.
23. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311(5765**1283-1287 [http://dx.doi.org/10.1126/science.1123061].
24. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P: **Quantitative phylogenetic assessment of microbial communities in diverse environments.** *Science* 2007, **315(5815**1126-1130 [http://dx.doi.org/10.1126/science.1133420].
25. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: **Prediction of effective genome size in metagenomic samples.** *Genome Biol* 2007, **8:**R10 [http://dx.doi.org/10.1186/gb-2007-8-1-r10].
26. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278(5338):**631-637.
27. Rodriguez-Brito B, Rohwer F, Edwards RA: **An application of statistics to comparative metagenomics.** *BMC Bioinformatics* 2006, **7:**162 [http://dx.doi.org/10.1186/1471-2105-7-162].
28. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
29. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308(5728**1635-1638 [http://dx.doi.org/10.1126/science.1110591].
30. Wells CL, Johnson MA, Henry-Stanley MJ, Bendel CM: **Candida glabrata colonizes but does not often disseminate from the mouse caecum.** *J Med Microbiol* 2007, **56(Pt 5**688-693 [http://dx.doi.org/10.1099/jmm.0.47049-0].
31. Lambert R, Chassignol S, Sedallian A, Descos L, Martin F: **Influence of gastrectomy and by-passing of the stomach on the intestinal flora of the rat.** *J Pathol Bacteriol* 1967, **94:**183-189 [http://dx.doi.org/10.1002/path.1700940123].
32. Lang-Unnasch N, Reith ME, Munholland J, Barta JR: **Plastids are widespread and ancient in parasites of the phylum Apicomplexa.** *Int J Parasitol* 1998, **28(11):**1743-1754.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29 [http://dx.doi.org/10.1038/75556].
34. **Datasets of comparative analysis** [http://www-ab.informatik.uni-tuebingen.de/software/megan/comparative/]