

RESEARCH ARTICLE

# MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data

Daniel H. Huson<sup>1,2\*</sup>, Sina Beier<sup>1</sup>, Isabell Flade<sup>3</sup>, Anna Górska<sup>1,4</sup>, Mohamed El-Hadidi<sup>1</sup>, Suparna Mitra<sup>5</sup>, Hans-Joachim Ruscheweyh<sup>1</sup>, Rewati Tappu<sup>1</sup>

**1** Center for Bioinformatics, University of Tübingen, Tübingen, Germany, **2** Life Sciences Institute, National University of Singapore, Singapore, **3** CeMeT GmbH, Tübingen, Germany, **4** IMPRS 'From Molecules to Organisms', MPI for Developmental Biology and University of Tübingen, Tübingen, Germany, **5** Norwich Medical School, University of East Anglia, Norwich, United Kingdom

\* [Daniel.Huson@uni-tuebingen.de](mailto:Daniel.Huson@uni-tuebingen.de)



**OPEN ACCESS**

**Citation:** Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. (2016) MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol* 12(6): e1004957. doi:10.1371/journal.pcbi.1004957

**Editor:** Timothée Poisot, Université de Montréal, CANADA

**Received:** January 28, 2016

**Accepted:** April 29, 2016

**Published:** June 21, 2016

**Copyright:** © 2016 Huson et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All source code is available here: <https://github.com/danielhuson/megan-ce>.

**Funding:** The authors received no specific funding for this work.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

There is increasing interest in employing shotgun sequencing, rather than amplicon sequencing, to analyze microbiome samples. Typical projects may involve hundreds of samples and billions of sequencing reads. The comparison of such samples against a protein reference database generates billions of alignments and the analysis of such data is computationally challenging. To address this, we have substantially rewritten and extended our widely-used microbiome analysis tool MEGAN so as to facilitate the interactive analysis of the taxonomic and functional content of very large microbiome datasets. Other new features include a functional classifier called InterPro2GO, gene-centric read assembly, principal coordinate analysis of taxonomy and function, and support for metadata. The new program is called MEGAN Community Edition (CE) and is open source. By integrating MEGAN CE with our high-throughput DNA-to-protein alignment tool DIAMOND and by providing a new program MeganServer that allows access to metagenome analysis files hosted on a server, we provide a straightforward, yet powerful and complete pipeline for the analysis of metagenome shotgun sequences. We illustrate how to perform a full-scale computational analysis of a metagenomic sequencing project, involving 12 samples and 800 million reads, in less than three days on a single server. All source code is available here: <https://github.com/danielhuson/megan-ce>

## Author Summary

Microbiome sequencing projects continue to grow rapidly, both in the number of samples considered and sequencing reads collected. With MEGAN Community Edition (CE), we provide a highly efficient program for interactive analysis and comparison of such data, allowing one to explore hundreds of samples and billions of reads. While taxonomic profiling is performed based on the NCBI taxonomy, we provide a number of different functional profiling approaches such as SEED, eggNOG, KEGG, and a new InterPro2GO

classification scheme. MEGAN CE also supports the use of metadata in the context of principal coordinate analysis and clustering analysis.

This is a *PLOS Computational Biology* Software paper.

## Introduction

In microbiome analysis, 16S rRNA amplicon sequencing [1] is often used when a high-level analysis of taxonomic content suffices, and/or computational resources are limited. However, metagenomic shotgun sequencing allows a more detailed analysis of taxonomic composition and also provides a detailed functional analysis of a microbiome [2].

Individual samples usually involve millions of DNA reads and a typical project may involve hundreds of such samples [3]. An important step in the analysis of such data is the alignment of the reads against a protein reference database such as NCBI-nr [4] or InterPro [5]. The authors of [6] compared 250 million DNA reads from permafrost samples against the KEGG database [7] (containing less than 10 million sequences) using BLASTX [8] and this reportedly took 800000 CPU hours at a supercomputer center [9].

We recently published a new alignment tool called DIAMOND [10] that is able to align short metagenomic sequencing reads against the NCBI-nr database at 20000 times the speed of BLASTX without loss of sensitivity. This makes it possible to analyze large metagenome samples with little computational effort. For example, alignment of the permafrost data against the NCBI-nr database (containing over 60 million reference sequences) takes about one day on a single server with 32 cores.

Once all metagenomic reads of a sample have been aligned against a reference database, the next task is to then determine the taxonomic and functional content of the microbiome samples. This poses a number of computational challenges. First, how to process the large number of items (billions of reads and alignments) so as to support their efficient analysis? Second, how to provide user-friendly tools that allow interactive inspection and analysis of the data? Third, how to host and serve this data in a straightforward manner?

This paper addresses all three challenges, based on our new software MEGAN CE, which is a major rewrite and substantial extension of our MEGAN 4 metagenome analyzer tool [11]. This software performs taxonomic and functional analysis of reads. It also facilitates the interactive exploration and comparison of metagenomic samples. MEGAN uses a compressed, indexed file format (called RMA) to store reads, alignments, as well as taxonomic and functional classification information for a given sample. While such files can be produced interactively using MEGAN CE, we also provide a command line tool called `blast2rma` to compute such files on a server. Alternatively, in case that DIAMOND is used to compute alignments, we also provide a command line tool called `Meganizer` that can be run on a diamond file so as to perform taxonomic and functional binning of the reads in the file. The resulting information is appended to the file, together with additional indices required to efficiently access reads via taxonomic or functional classes. Meganized diamond files can be directly opened in MEGAN CE without any further processing and they are roughly the same size as the corresponding uncompressed fastq files.

Previous versions of MEGAN [11] required that files are present on the computer on which the software is running. While this remains possible with MEGAN CE, we also provide a new

program called MeganServer (manuscript in preparation) that serves RMA and meganized diamond files over a local network or the internet.

By integrating DIAMOND, MEGAN and MeganServer into a single, streamlined pipeline, we provide a straightforward and fast solution for microbiome analysis, facilitating the analysis of hundreds of samples and billions of reads on a single server in a matter of days. Any given sample is represented by only two or at most three files, namely the initial compressed fastq file obtained from a sequencer and either a meganized diamond file, when using DIAMOND, or an alignment file followed by an RMA file, when using some other alignment tool. In both cases, the resulting files contain all aligned reads, alignments and classification details. They can be stored on a server and made accessible through the MeganServer software, see [Fig 1](#).

Two of the main goals of computational analysis of metagenomic data is to determine the taxonomic content of each sample, i.e. which organisms are present, and to estimate the functional capacity of the sample, i.e. which genes are present. This can be addressed by assigning sequencing reads to taxa and functional categories, based on their alignments to a reference database, in a process called binning.

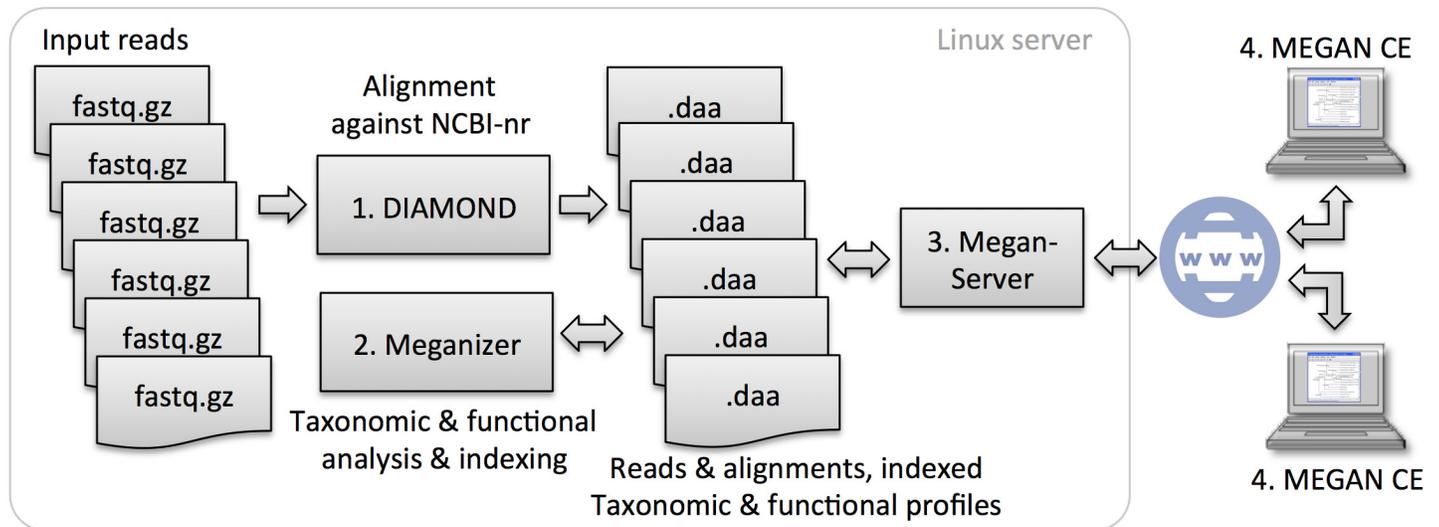
By default, MEGAN CE performs taxonomic binning by assigning reads to nodes in the NCBI taxonomy using the LCA algorithm [12]. MEGAN CE supports a number of different classification systems for the binning of reads by function. A novel *InterPro2GO* analyzer uses a metagenome GO-slim [13] to classify InterPro families [5] and is based on files publicly available from EBI. MEGAN CE offers a SEED analyzer based on the concepts of subsystems and functional roles [14] and an *eggNOG* viewer based on the *eggNOG* extension of COGs [15]. In addition, MEGAN CE provides a legacy KEGG [7] viewer, based on files downloaded from KEGG in 2011.

One can easily execute principal coordinate analysis (PCoA) and cluster analysis using a number of different ecological indices and methods, and also compute standard alpha diversity indices.

In MEGAN CE, we offer a gene-centric approach to sequence assembly. The user can request to have all reads assigned to any given taxonomic or functional node assembled and output as contigs. This calculation is performed on-the-fly (manuscript in preparation) from within MEGAN, requiring no additional software or major calculations.

With DIAMOND, Meganizer, MeganServer and MEGAN CE, we provide a complete and highly-efficient solution for performing metagenome analysis. To illustrate the speed and sensitivity of our pipeline, we report on the computational analysis of a set of 12 human gut metagenomic samples, consisting of 800 million HiSeq reads [16]. From beginning-to-end, it took only 67 hours (wall-clock) on a single server, to align all reads against the NCBI-nr database (downloaded February 2015, approximately 64 million protein sequences) and then to perform taxonomic and functional analysis, using *InterPro2GO*, SEED, *eggNOG* and KEGG, involving 620 million reads and nearly ten billion alignments. MEGAN CE and MeganServer provide easy access to the resulting files, allowing users to perform both high-level analyses using trees, charts or PCoA plots, or low-level analyses such as drilling down to individual organisms, genes, reads or alignments, on single or multiple samples. This data can be accessed using MEGAN CE by opening the default public instance of MeganServer, which is hosted at the University of Tübingen.

Other popular standalone taxonomic analysis tools include MetaPhlAn [17], MetaPhyler [18] and Kraken [19]. Another is QIIME [20], which was initially developed to analyze 16S rRNA sequences. Services such as MG-RAST [21] and the EBI metagenomic web service [13] allow users to upload their data so as to use provided computational facilities for taxonomic and functional analysis of metagenomic sequencing data. See [22, 23] for two recent comparisons of the performance of different approaches.



**Fig 1. The compressed fastq files in a metagenome or metatranscriptome sequencing project are (1) compared against a protein reference database such as NCBI-nr using DIAMOND. (2) Taxonomic and functional analysis is then performed on the diamond files using Meganizer. (3) The resulting meganized diamond files remain on the server and are accessed via the MeganServer software. (4) Researchers work interactively with the data using MEGAN CE.**

doi:10.1371/journal.pcbi.1004957.g001

## Design and Implementation

This paper introduces MEGAN Community Edition (CE), which is a major update of our MEGAN software [11]. This release contains a large number of new features and has been substantially rewritten so as to support the analysis of many samples (hundreds) and many reads (billions). This release includes a number of command line tools, in particular blast2rma, daa2rma and Meganizer, which can all be used to prepare input files for MEGAN CE.

In addition, we recommend the use of DIAMOND [10] for ultra-fast alignment of reads against NCBI-nr and MeganServer to allow web access to MEGAN files.

### RMA files

MEGAN CE analysis requires that sequencing reads are first aligned against a suitable reference database, such as NCBI-nr in the case of a protein-based analysis, or Genbank for a DNA-based analysis, or the Silva database [24], say, when aligning 16S rRNA reads. MEGAN CE can import reads and alignments in a number of different file formats and computes a compressed and indexed binary file in so-called RMA format that contains all reads, alignments, taxonomic and functional classifications. The file is indexed to allow quick access to reads and alignments by taxonomic or functional assignment. MEGAN CE also provides a command line program called blast2rma for computing RMA files from BLAST-like alignments. With MEGAN CE, we introduce a new version of the RMA format that requires much less disk space than previous versions.

### Meganizer

Our alignment program DIAMOND produces “diamond files” in a binary output format called DAA (“Diamond alignment archive”), from which both tabular and SAM format can be extracted. We provide a new program called Meganizer that analyses all reads present in a given diamond file, performs taxonomic and functional analysis of them, and then appends the resulting classifications and indices to the end of the diamond file. Meganizing a diamond file takes much less time than generating an RMA file and reduces the number of files that are created

during metagenome analysis. Indeed, using DIAMOND and Meganizer, each sample in a metagenome study is represented by only two files, namely the original compressed fastq file and the resulting meganized diamond file. This file is usually smaller than the corresponding RMA file.

## Taxonomic assignment

By default, MEGAN CE uses the naive LCA algorithm [12] to perform *taxonomic binning*. In this approach, each read is assigned to the “lowest common ancestor” node in the taxonomy that lies above all species for which the read has a significant alignment. The rationale here is that reads that align to widely conserved genes should be assigned to high-level taxa (such as the rank of Phylum), whereas reads that align to a gene that is specific to a given type of organisms should be assigned to a more lower taxon (such as at the rank of Genus or Species). As a consequence, reads are binned across all taxonomic ranks.

The naive LCA algorithm provides a conceptually straight-forward and fast approach to taxonomic binning, running at a rate of over 100 million reads and 2 billion alignments per hour on a single server, as discussed below. However, it is less suited for purposes of taxonomic *profiling*, where the goal is to obtain an accurate estimation of the taxonomic content of a sample, see [22, 23]. One reason for the poorer performance of the naive LCA algorithm as a profiling tool is that it processes each read in isolation, independent of all other reads.

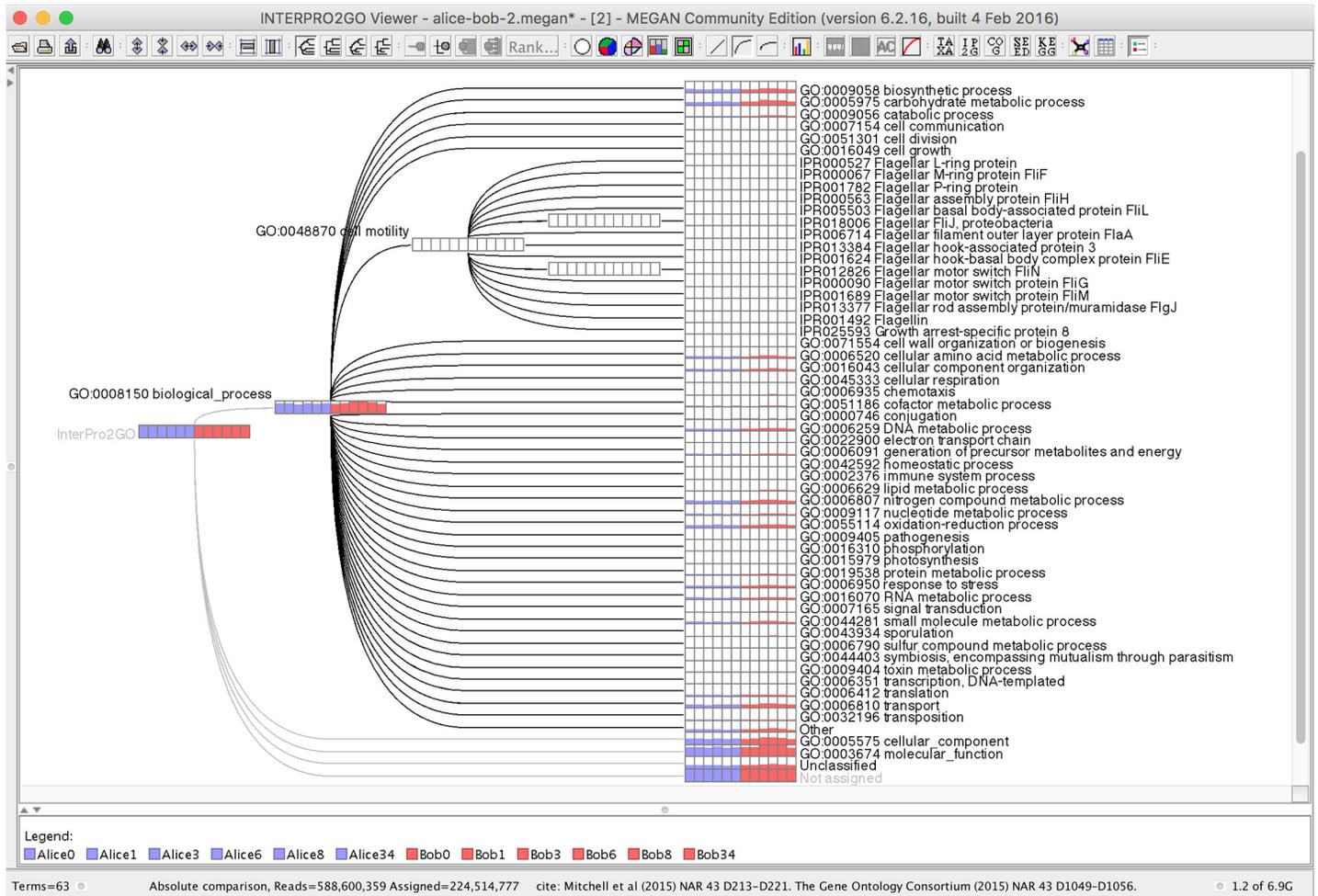
To address this issue, in MEGAN CE we provide an implementation of the *weighted LCA* algorithm, which was developed in the context of the 2013 DTRA Algorithms Challenge [25, 26]. The weighted LCA algorithm operates as follows. In a first phase, each reference sequence  $S$  is assigned a weight. This is the number of reads  $R$  that only align to  $S$  (or to other references as well, as long as they have the same species assignment as  $S$ ). Then, in a second phase, each read  $R$  is placed on the lowest node in the taxonomy that is above 75% (by default) or more of the total weight of all references to which  $R$  has a significant alignment. This improves the specificity of taxonomic assignment, but requires more time to run.

A *taxonomic profile* is usually calculated at a single specific taxonomic rank  $H$  and aims at providing of the number of reads attributable to each of the taxa at the given rank (such as a Genus-level profile). In contrast, both the naive and the weighted LCA algorithm assign reads across all ranks of the taxonomy. To summarize all counts at a fixed taxonomic rank, MEGAN CE provides a simple *projection algorithm*, which operates as follows. First, all reads that are assigned by the LCA algorithm to some taxon node  $t$  that lies *above* the desired rank  $H$  are pushed down to the children of  $t$  in proportion to the number of reads assigned on or below each of the the children. This is repeated until all reads have been pushed down to a node that lies in  $H$ . Second, all reads that are assigned by the LCA algorithm to some taxon node *below* the desired rank  $H$  are simply assigned to the ancestor node that lies in  $H$ .

## InterPro2GO viewer

The EBI metagenome service provide a hierarchical classification of reads by Gene Ontology assignment based on the metagenomic GO-slim [13] and a tabular mapping of reads to InterPro families [5]. Based on these concepts and data downloaded from <http://www.ebi.ac.uk/interpro> and <http://www.uniprot.org>, we have designed a novel InterPro2GO hierarchical viewer in which the top two tiers of nodes are based on the metagenomic GO-slim and all lower-level nodes represent InterPro families (see Fig 2).

Our new InterPro2GO classification has three first-tier nodes labeled “GO:0008150 biological process”, “GO:0005575 cellular component” and “GO:0003674 molecular function”, which represent the three domains of the Gene Ontology [27]. Below these, there are 84 second-tier nodes that provide a refined GO-based classification of function. A third tier of nodes



**Fig 2. The new InterPro2Go viewer.** High-level nodes represent the metagenomic GO-slim [13], whereas low-level nodes are based on InterPro [5]. Here we have uncollapsed the GO “biological process” domain node to show the second tier nodes attached below it. Each node is labeled by a bar chart representing the number of reads assigned to the node, or below it, for 12 different human stool samples [16]. In this example, 27.5% of 816 million reads are assigned to an InterPro family by MEGAN CE.

doi:10.1371/journal.pcbi.1004957.g002

represents all InterPro families that have an assignment to one or more GO categories. An InterPro family will give rise to more than one node in the graph, if it maps to different GO domains. An InterPro family that maps to multiple second-tier GO nodes in the same GO domain is placed below an “Other” node that is associated with the GO domain node. By construction, the resulting classification is a tree in which each InterPro family occurs up to three times, at most once below each domain node.

A fourth first-tier node labeled “Unclassified” is a catch-all node for all InterPro families that have not been assigned a GO assignment. The InterPro2GO classification in MEGAN CE has approximately 26 000 nodes in total.

### SEED viewer

The SEED [14] is a functional classification that is based on an assignment of genes to “functional roles”. These are grouped into “subsystems” of related functional roles that make up a metabolic pathway, a complex, or a class of proteins. MEGAN CE uses a significantly updated

SEED viewer that is based on files downloaded from SEED in November 2015. Reads are binned to functional roles.

### eggNOG viewer

MEGAN CE offers a new eggNOG viewer that is based on a classification of orthologous groups in which reads are binned to “clusters of orthologous groups” (COGs) and “non-supervised orthologous groups” that appear as leaves of the eggNOG classification [15].

### KEGG viewer

MEGAN CE also provides a functional viewer based on KEGG [7]. Here, KEGG orthologous groups (KO groups) are mapped to enzymes that appear in metabolic pathways. MEGAN CE ships with a legacy representation of KEGG that is based on files downloaded from the KEGG website in early 2011. For users in possession of a KEGG ftp license, a separate program called MEGAN UE (Ultimate Edition) provides tools for generating an up-to-date representation of KEGG in MEGAN UE.

### Mapping reference sequences

The classification of metagenomic reads depends on a classification of the reference sequences to which they align. MEGAN provides three mechanisms for determining the classification identity of reference sequences. First, MEGAN can scan for a classification tag in the header line of a reference. For example, a NCBI taxon id may be written as `tax|666`. Second, MEGAN supports the mapping of GI numbers to taxon or functional identifiers using a file-based index. Third, MEGAN supports the mapping of alpha-numerical accession numbers to taxonomic or functional identifiers using a file-based hash-table, in anticipation of NCBI’s plan to discontinue the use of GI numbers.

### Working with multiple samples and metadata

Any number of metagenome samples can be opened together in a single “comparison document”, using a dialog that allows one to source samples both from local disk and from any instance of MeganServer that MEGAN CE is currently connected to.

Metadata such as technical and clinical parameters, or physical and bio-chemical measurements, can be attached to individual samples. When such samples are combined into a single comparison document, then all associated metadata is merged into a single table. To facilitate working with metadata, MEGAN CE provides an interactive “sample viewer” that is based on a versatile spreadsheet that can be used to add or edit metadata, and also to select or group samples by metadata attribute values, or to set colors, labels, shapes by such values (see Fig 3). Metadata can be imported or exported using a standard CSV format that is compatible with other metagenome analysis tools.

Given a comparison document containing multiple samples, the sample viewer also allows one to extract individual samples, or to compute the “total biome” (i.e. the union of all samples) or the “core biome” (i.e. those taxa that appear in a given minimum percentage of all samples). One can also easily merge and compare samples based on the values of a selected attribute.

### PCoA, bi-plots and tri-plots

PCoA (principle coordinate analysis) is a standard tool used in the analysis of microbiome data [28]. MEGAN CE allows the user to perform such analysis in two or three dimensions, based on taxonomic or functional profiles. The program offers a number of different ecological

	#SampleID	sequencer	Antibiotic	Health	Size	Label	Gender	Timepoint
1	Alice0	iSeq2000	ciprofloxacin	good	21052127	A-0 -	male	0
2	Alice1	iSeq2000	ciprofloxacin	good	63156390	A-1 +	male	1
3	Alice3	iSeq2000	ciprofloxacin	good	21052129	A-3 +	male	3
4	Alice6	iSeq2000	ciprofloxacin	good	42104255	A-6 +	male	6
5	Alice8	iSeq2000	ciprofloxacin	good	63156380	A-8 -	male	8
6	Alice34	iSeq2000	ciprofloxacin	good	21052132	A-34 -	male	34
7	Bob0	iSeq2000	ciprofloxacin	good	42104259	B-0 -	male	0
8	Bob1	iSeq2000	ciprofloxacin	good	42104249	B-1 +	male	1
9	Bob3	iSeq2000	ciprofloxacin	good	63156377	B-3 +	male	3
10	Bob6	iSeq2000	ciprofloxacin	good	84208516	B-6 +	male	6
11	Bob8	iSeq2000	ciprofloxacin	good	63156385	B-8 -	male	8
12	Bob34	iSeq2000	ciprofloxacin	good	84208507	B-34 -	male	34

**Fig 3. Spreadsheet for entry and analysis of metadata associated with samples.**

doi:10.1371/journal.pcbi.1004957.g003

indices to be used in the calculation of PCoA plots, such as the euclidean distance, in which case the resulting plot is identical to the result of a PCA (principle component analysis), the Bray-Curtis distance [29], or the Jensen-Shannon distance, to name a few. MEGAN CE provides an implementation of a bi-plot, in which the taxa or functional groups that contribute the most to the variation shown in the PCoA plot are represented by vectors that indicate the direction of steepest increase (see Fig 4). MEGAN CE also provides a tri-plot, in which metadata that correlates the strongest with the changes shown in the PCoA plot are indicated by vectors.

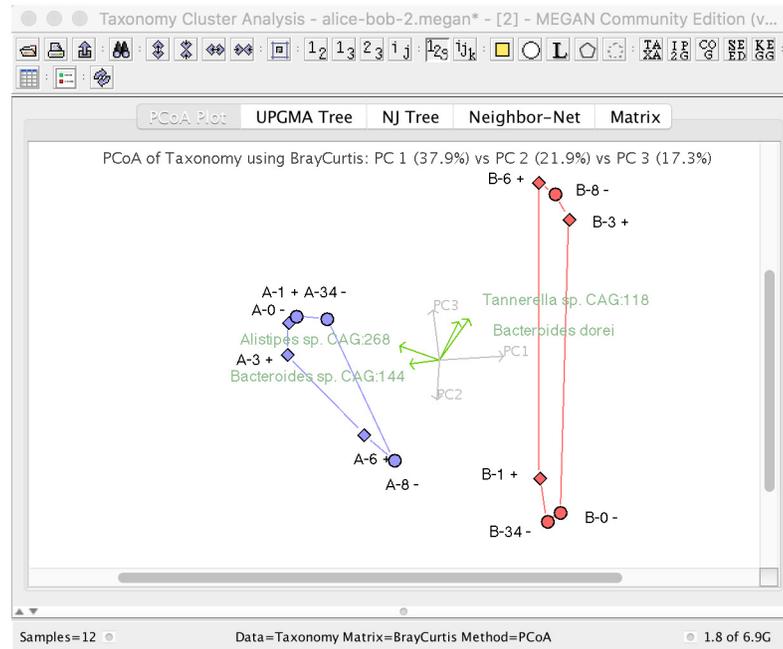
### Gene centric assembly

The aim of metagenome assembly is to stitch sequencing reads together so as to obtain longer stretches of contiguous sequence (contigs) of the genomes of the organisms present in a given microbiome. Metagenome assembly from Illumina reads is considered a difficult problem [30], although progress is being made [31]. In the past, one main reason for performing metagenome assembly has been to reduce the total amount of sequence that has to be aligned against a reference database, while another reason is so as to improve specificity. However, the introduction of DIAMOND has changed this equation and aligning all reads without an initial assembly step is now usually the faster option.

In MEGAN CE, we provide a gene-centric approach to sequence assembly that aims at assembling individual genes at the strain level, guided by protein alignments (see also e.g. [32]). The program can assemble all reads assigned to a specific node in a taxonomic or functional classification, allowing one to investigate the sequence variability of a given gene. This calculation is triggered interactively and is performed on-the-fly. Protein alignments to reference sequences are employed to infer DNA overlaps between reads, giving rise to an overlap graph, from which contigs are extracted (manuscript in preparation).

### Results

We illustrate the application of DIAMOND, Meganizer, MEGAN CE and MeganServer using public data from a recent study that we were involved in [16]. In this study, two healthy



**Fig 4. A PCoA analysis of 12 human gut samples [16] computed using species-level profiles and Bray-Curtis distances.** Samples are labeled by subject pseudonym, day 0–34 and whether antibiotics were taken (+) or not (-) on the given day. For both subjects, the plot clearly shows that the taxonomic profiles move further and further away from the original during the course of antibiotics, but then return back close to the original at the end of the study. The top five bi-vectors are also shown, labeled by species name.

doi:10.1371/journal.pcbi.1004957.g004

volunteers, with aliases Alice and Bob (although both are males) were administered a course of antibiotics (ciprofloxacin, days 2–6). Stool samples were collected on days 0, 1, 3, 6, 8 and 34, resulting in a total of 12 samples. The samples were subjected to shotgun metagenome sequencing (HiSeq2000, paired end, 100bp), resulting in 816 million reads (see Table 1).

The wall-clock time for the complete taxonomic and functional analysis of this data was 67 hours on a single server (32 cores, 512GB of memory). In more detail, DIAMOND alignment of all 816 million reads against NCBI-nr (downloaded February 2015) took 62 hours (wall-clock time), resulting in just under ten billion alignments. Meganization of the resulting diamond files took an additional 5 hours, running Meganizer in parallel on all 12 samples. The fastq files have a total (uncompressed) size of 199 Gb, whereas the meganized diamond files occupy 150 Gb. Loading the files into MeganServer took about one minute.

DIAMOND found one or more significant alignments against NCBI-nr for 75% (620 million) of the 816 million input sequencing reads. Taxonomic analysis assigned 71.7% (585 million) to a taxon. Functional analysis using InterPro2GO assigned 27.5% (224 million) to an InterPro family. Functional analysis using SEED assigned 10% (82 million) to a functional role in SEED. Functional analysis using eggNOG assigned 17.5% (134 million) to a COG or eNOG. Functional analysis using KEGG (downloaded July 2015) assigned 14% (114 million) of the total reads to a KO.

While the main focus of the study reported by [16] was to determine how the levels of antibiotic resistance genes change in the gut during the course of antibiotic treatment, in Fig 4 we present a PCoA analysis of all samples, computed using Bray-Curtis distances. For both subjects, we clearly see that the taxonomic profiles of their stool samples move away from the original as the treatment progresses. On day 34, that is, 28 days after the end of the course of

**Table 1. For twelve shotgun metagenome samples [16], we report (a) the number of reads, (b) wall-clock time required to align the reads against NCBI-nr using DIAMOND, (c) the number of matches obtained, (d) the number of reads that have at least one alignment and (e) the time required to run Meganizer to perform taxonomic and functional classification of all reads. The total wall-clock time is 67 hours on a single server with 32 cores.**

Sample	(a) Reads	(b) DIAMOND (s)	(c) Alignments	(d) Aligned reads	(e) Meganizer (s)
Alice 0	66 393 401	19 062	627 405 772	44 900 227	9 299
Alice 1	64 923 975	15 771	595 715 349	43 498 105	11 338
Alice 3	55 092 349	13 435	515 249 349	37 675 494	8 621
Alice 6	66 289 376	16 801	910 892 059	52 627 776	11 771
Alice 8	57 957 661	14 134	790 946 244	45 358 448	13 911
Alice 34	64 380 386	15 615	608 114 143	44 741 897	11 962
Bob 0	61 232 588	14 573	825 213 917	48 882 884	12 058
Bob 1	65 763 766	16 203	841 038 616	51 408 892	12 270
Bob 3	89 034 641	34 598	1 233 571 041	72 017 720	15 789
Bob 6	89 339 172	27 333	1 138 796 522	70 344 161	15 507
Bob 8	78 001 118	19 734	1 049 831 855	63 336 241	13 423
Bob 34	57 627 119	15 406	780 844 319	45 568 158	11 433
Total	816 035 552	222 665	9 917 619 186	620 360 003	Max: 15 789
Time		≈ 62 h			≈ 5 h

doi:10.1371/journal.pcbi.1004957.t001

antibiotics, in the case of Bob, the taxonomic profile is practically indistinguishable from his original profile, while in the case of Alice it has returned most of the way.

We also display the top five bi-plot vectors. The vectors point in the direction of those samples that have substantially higher reads counts for the given species. For example, *Bacteroides stercoris* CAG:120 points in the direction of samples A-6 and A-8, whereas the *Bacteroides dorei* vector is oriented toward B-3, B-6 and B-8. Interestingly, much of the clear separation of Alice’s and Bob’s samples is based on *Bacteroides* species, which is a difference that would be lost in a Genus-rank comparison.

## Availability and Future Directions

The programs MEGAN CE, DIAMOND and MeganServer can be downloaded here: <http://www-ab.informatik.uni-tuebingen.de/software>.

MEGAN Community Edition is free software under the GNU General Public License. All source code is available here: <https://github.com/danielhuson/megan-ce>. User support for the Community Edition is provided through a community website at <http://megan.informatik.uni-tuebingen.de>.

With MEGAN CE, DIAMOND, Meganizer and MeganServer, we provide a powerful suite of programs that allow researchers to explore and analyze microbiome sequencing samples interactively on a very large scale. Future work will continue to focus on speeding up the analysis and supporting the exploration of ever greater numbers of ever larger microbiome sequencing samples, fueled by the continuing decrease in the price of sequencing.

## Author Contributions

Conceived and designed the experiments: DHH. Performed the experiments: DHH SB IF AG MEH SM HJR RT. Analyzed the data: DHH SB IF AG MEH SM HJR RT. Wrote the paper: DHH AG. Suggested and tested software features: SB IF MEH SM HJR RT.

## References

1. Pace NR, Stahl DA, Lane DJ, Olsen GJ. Analyzing natural microbial populations by rRNA sequences. *ASM News*. 1985; 51:4–12.
2. Handelsman J, Rondon MR, Brady SG, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology*. 1998; 5:245–249. doi: [10.1016/S1074-5521\(98\)90108-9](https://doi.org/10.1016/S1074-5521(98)90108-9)
3. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*. 2015; 17(5):690–703. doi: [10.1016/j.chom.2015.04.004](https://doi.org/10.1016/j.chom.2015.04.004)
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2005; 1(33):D34–38.
5. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*. 2015; 43(D1):D213–D221. Available from: <http://nar.oxfordjournals.org/content/43/D1/D213.abstract>. doi: [10.1093/nar/gku1243](https://doi.org/10.1093/nar/gku1243) PMID: [25428371](https://pubmed.ncbi.nlm.nih.gov/25428371/)
6. Mackelprang R, Waldrop M, DeAngelis K, David M, Chavarria K, Blazewicz S, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*. 2011; 480(7377):368–371. doi: [10.1038/nature10576](https://doi.org/10.1038/nature10576) PMID: [22056985](https://pubmed.ncbi.nlm.nih.gov/22056985/)
7. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000 Jan; 28(1):27–30. doi: [10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27) PMID: [10592173](https://pubmed.ncbi.nlm.nih.gov/10592173/)
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 1990; 215:403–410. doi: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
9. Jansson J. Towards “Tera-Terra”: Terabase Sequencing of Terrestrial Metagenomes. *Microbe magazine*. 2011 July;.
10. Buchfink B, Xie C, Huson DH. Fast and Sensitive Protein Alignment using DIAMOND. *Nature Methods*. 2015; 12:59–60. doi: [10.1038/nmeth.3176](https://doi.org/10.1038/nmeth.3176) PMID: [25402007](https://pubmed.ncbi.nlm.nih.gov/25402007/)
11. Huson DH, Mitra S, Weber N, Ruscheweyh HJ, Schuster SC. Integrative analysis of environmental sequences using MEGAN 4. *Genome Research*. 2011; 21:1552–1560. doi: [10.1101/gr.120618.111](https://doi.org/10.1101/gr.120618.111) PMID: [21690186](https://pubmed.ncbi.nlm.nih.gov/21690186/)
12. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007 March; 17(3):377–386. Available from: <http://dx.doi.org/10.1101/gr.5969107>. doi: [10.1101/gr.5969107](https://doi.org/10.1101/gr.5969107) PMID: [17255551](https://pubmed.ncbi.nlm.nih.gov/17255551/)
13. Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, et al. EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*. 2014 01; 42(Database issue):D600–D606. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965009/>. doi: [10.1093/nar/gkt961](https://doi.org/10.1093/nar/gkt961) PMID: [24165880](https://pubmed.ncbi.nlm.nih.gov/24165880/)
14. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*. 2013; Available from: <http://nar.oxfordjournals.org/content/early/2013/11/29/nar.gkt1226.abstract>.
15. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*. 2012; 40(Database-Issue):284–289. doi: [10.1093/nar/gkr1060](https://doi.org/10.1093/nar/gkr1060)
16. Willmann M, El-Hadidi M, Huson DH, Schütz M, Weidenmaier C, Autenrieth IB, et al. Antibiotic selection pressure determination through sequence-based metagenomics. *Antimicrobial Agents and Chemotherapy*. 2015 Dec; 59(12):7335–45. doi: [10.1128/AAC.01504-15](https://doi.org/10.1128/AAC.01504-15) PMID: [26369961](https://pubmed.ncbi.nlm.nih.gov/26369961/)
17. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth*. 2012; 9(8):811–814. doi: [10.1038/nmeth.2066](https://doi.org/10.1038/nmeth.2066)
18. Liu B, Gibbons T, Ghodsi M, Pop M. MetaPhyler: Taxonomic profiling for metagenomic sequences. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2010. p. 95–100. Available from: <http://dx.doi.org/10.1109/bibm.2010.5706544>.
19. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 2014; 15:R46. doi: [10.1186/gb-2014-15-3-r46](https://doi.org/10.1186/gb-2014-15-3-r46) PMID: [24580807](https://pubmed.ncbi.nlm.nih.gov/24580807/)
20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. 2010 Apr; 7(5):335–336. doi: [10.1038/nmeth.f.303](https://doi.org/10.1038/nmeth.f.303) PMID: [20383131](https://pubmed.ncbi.nlm.nih.gov/20383131/)

21. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the Metagenomics RAST Server (MG-RAST) for Analyzing Shotgun Metagenomes. *Cold Spring Harb Protoc.* 2010 Jan; 2010(1):pdb.prot5368+. doi: [10.1101/pdb.prot5368](https://doi.org/10.1101/pdb.prot5368) PMID: [20150127](https://pubmed.ncbi.nlm.nih.gov/20150127/)
22. Peabody MA, Van Rossum T, Lo R, Brinkman FSL. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics.* 2015; 16(1):1–19. doi: [10.1186/s12859-015-0788-5](https://doi.org/10.1186/s12859-015-0788-5)
23. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports.* 2016; 6:19233. doi: [10.1038/srep19233](https://doi.org/10.1038/srep19233) PMID: [26778510](https://pubmed.ncbi.nlm.nih.gov/26778510/)
24. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *NAR.* 2013; 41:D590–D596. doi: [10.1093/nar/gks1219](https://doi.org/10.1093/nar/gks1219) PMID: [23193283](https://pubmed.ncbi.nlm.nih.gov/23193283/)
25. Servick K. Controversial Pentagon DNA Analysis Contest Names Champion; 2013. *Science Magazine.*
26. Buchfink B, Huson DH, Xie C. Metascope—Fast and accurate identification of microbes in metagenomic sequencing data. *arXiv; 2015.* arXiv:1511.08753.
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000 May; 25(1):25–29. PMID: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)
28. Goodrich JK, Riezi SCD, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a Microbiome Study. *Cell.* 2014; 158(2):250–262. doi: [10.1016/j.cell.2014.06.037](https://doi.org/10.1016/j.cell.2014.06.037) PMID: [25036628](https://pubmed.ncbi.nlm.nih.gov/25036628/)
29. Bray RJ, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr.* 1957; 27:325–349. doi: [10.2307/1942268](https://doi.org/10.2307/1942268)
30. Treangen T, Koren S, Sommer D, Liu B, Astrovskaia I, Ondov B, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biology.* 2013; 14(1):R2. Available from: <http://genomebiology.com/2013/14/1/R2>. doi: [10.1186/gb-2013-14-1-r2](https://doi.org/10.1186/gb-2013-14-1-r2) PMID: [23320958](https://pubmed.ncbi.nlm.nih.gov/23320958/)
31. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences.* 2014; 111(13):4904–4909. doi: [10.1073/pnas.1402564111](https://doi.org/10.1073/pnas.1402564111)
32. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome.* 2015; 3(1):1–13. doi: [10.1186/s40168-015-0093-6](https://doi.org/10.1186/s40168-015-0093-6)