Chapter 19

Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN

Caner Bağcı, Sina Beier, Anna Górska, and Daniel H. Huson

Abstract

Metagenomics has become a part of the standard toolkit for scientists interested in studying microbes in the environment. Compared to 16S rDNA sequencing, which allows coarse taxonomic profiling of samples, shotgun metagenomic sequencing provides a more detailed analysis of the taxonomic and functional content of samples. Long read technologies, such as developed by Pacific Biosciences or Oxford Nanopore, produce much longer stretches of informative sequence, greatly simplifying the difficult and time-consuming process of metagenomic assembly. MEGAN6 provides a wide range of analysis and visualization methods for the analysis of short and long read metagenomic data. A simple and efficient analysis pipeline for metagenomic analysis consists of the DIAMOND alignment tool on short reads, or the LAST alignment tool on long reads, followed by MEGAN. This approach performs taxonomic and functional abundance analysis, supports comparative analysis of large-scale experiments, and allows one to involve experimental metadata in the analysis.

Key words Metagenomics, Software, MEGAN, Taxonomic analysis, Functional analysis, Long reads

1 Introduction

Metagenomics is the study of microbiome samples, such as obtained from ocean water, soil, plant matter, or feces, say, using high-throughput DNA sequencing [1]. Metagenomic sequencing allows the study of microorganisms found in environmental samples without relying on culturing methods or prior knowledge of the composition of the community. With metagenomics, one can determine the taxonomic and functional content of samples.

While most metagenomic projects to date have used short read sequencing (next-generation sequencing), there is increasing interest in using long read sequencing technologies in this area. Long read technologies have been considered too expensive, difficult, or error-prone for application in metagenomics. However, this is changing and computational analysis methods designed for processing short reads now need to be modified to work well on long

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_19,

reads, so as to make good use of the ability of long reads to cover multiple genes.

A major computational challenge in metagenomics is the alignment of sequencing reads against a comprehensive reference database. Billions of reads can be aligned against a large protein reference database in reasonable time using high-throughput alignment tools such as DIAMOND [2]. Long reads require frame-shift aware alignment tools, such as LAST [3, 4], because insertions or deletions due to sequencing errors impact long reads, as discussed in Subheading 2.

In the following, we will first discuss how to perform basic alignment and analysis of short reads in Subheading 2.1 and long reads in Subheading 2.2. We will then show, in Subheading 3, how to compare large numbers of samples in MEGAN6 [5] and perform basic statistical analysis of the samples and their metadata. In Subheading 4 we briefly discuss the challenges we will have to face to further improve the analysis of data from environmental samples. Finally, in Subheading 4.1 we describe some additional resources available for using MEGAN 6.

2 Workflows for Metagenomic Analysis with MEGAN

The basic workflow for using MEGAN consists of two main steps: read alignment against a reference database and then import an analysis of the alignments in MEGAN. The aim of pipeline is to perform taxonomic and functional binning of the input reads.

The alignment can be performed using a number of different tools depending on the type of sequencing data and on the chosen database, its sequence type, size, and available computer power. For smaller databases more sensitive tools can be chosen such as MALT [6] or even BLAST [7]. These tools generally offer higher sensitivity at the cost of a longer runtime. For large datasets and databases, it is more suitable to choose an alignment tool such as DIAMOND or LAST. We use the NCBI NR database [8] with both of the latter tools, because it is the largest and most comprehensive protein database available today. NCBI NR contains 144.5 million protein sequences (August 2017).

2.1 Short Read Pipeline

We describe here the basic short read analysis pipeline as shown in Fig. 1. By default, we use DIAMOND to align reads against the full NCBI NR database.

Before running the pipeline, one can optionally perform preprocessing, that is, quality control, trimming, and filtering, of the raw reads. However, these steps usually have little impact on the results of the alignment-based analysis described in this document.

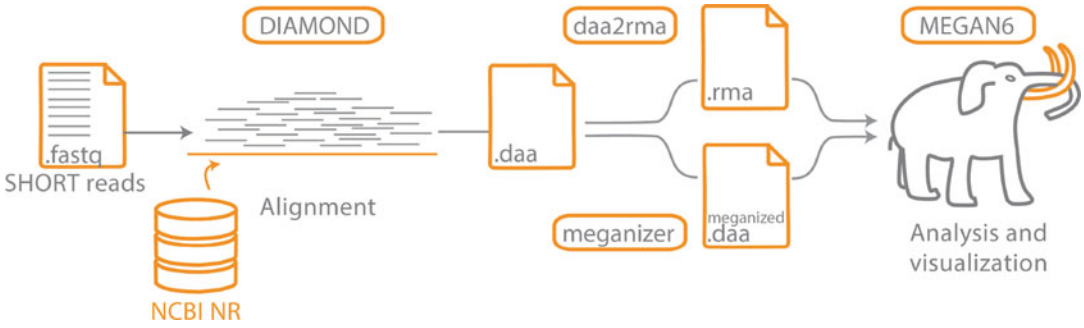


Fig. 1 Basic pipeline for short read analysis

2.1.1 Read Alignment with DIAMOND

DIAMOND uses double indexed alignment, which means both the reference database and the query are indexed for comparison. This leads to a large speedup especially for large queries and databases. Like BLASTX, DIAMOND uses the "seed and extend" method to find all matches between a query and the database. To further increase speed, DIAMOND utilizes spaced seeds, which are long seeds where only some positions are used for matching the seed. This leads to another increase of speed without decreasing sensitivity.

DIAMOND can be run either in fast or sensitive mode. Fast mode will run around 20,000 times faster than BLASTX on short reads and will be able to find 75–90% of all relevant matches that one would find with BLASTX, while sensitive mode provides a speedup of 2500× while recovering up to 94% of significant matches.

2.1.2 Taxonomic and Functional Classification with MEGAN6

DIAMOND can save alignments in a compressed format called DAA (DIAMOND alignment archive) format. DAA files can be imported into MEGAN6 in multiple ways. A small number of small DAA files can easily be imported interactively using menu items provided in MEGAN. For larger datasets and or many files, one should use the command-line tools provided with MEGAN. These include `daa2rma`, which will generate a RMA file as used by MEGAN from one or two (for paired reads) DIAMOND files and `daa-meganizer`, which analyzes a DAA file and then appends the result to the end of the file. Such "meganized" DAA files can then be opened directly in MEGAN. The latter approach is much faster and is more space efficient. However, to use paired reads all alignments have to be in the same file.

One can use the program `blast2rma` to process the output of a range of different alignment programs, such as BLAST.

During the processing of alignments for MEGAN, the reads will be assigned to nodes in the NCBI taxonomy and any functional classifications that have been configured in the import dialog or on

the command-line. Taxonomic binning of each read is done separately, by assigning it to the lowest common ancestor (LCA) of its significant matches. Matches can be filtered by multiple parameters, for example, e-value and bit-score, as well as sequence identity. Only matches passing those filters will be used to determine the LCA. It is also important to choose the minimum support (or minimum support percentage), the number or percentage of reads that must be assigned to a single taxon before it will be part of the final result. Reads assigned to a taxon that does not pass the minimum support filter will be pushed up the taxonomy until a taxon is found that passes the filter.

Functional binning is performed by mapping the NCBI database accessions for the matches of a read to identifiers of the selected functional classification. Mapping files are currently available for InterPro2GO [9, 10] (InterPro families embedded in a GO-based hierarchy), eggNOG [11], KEGG [12], and SEED [13].

2.1.3 *Investigation of the Results*

The resulting files can be opened and interactively investigated using the MEGAN6 graphical user interface. The first view when opening a file is always a hierarchical representation of the taxonomic composition of the sample. Selecting different nodes of this tree, the user can uncover further information on the reads mapped to the represented taxon. Selecting `Inspect Reads` on a node will open the Inspector Window, which displays the reads assigned to that node, as well as their alignments. This functionality can be used both in the Taxonomy Viewer, where nodes represent taxa, and in any of the Functional Viewers. Figure 2a shows an example of the Inspector Window.

Instead of just viewing a listing of the matches and alignments, it is also possible to select `Show Alignments`. This will open the Alignment Viewer (Fig. 2b), where for each of the database references with matches from the reads assigned to the selected node it is

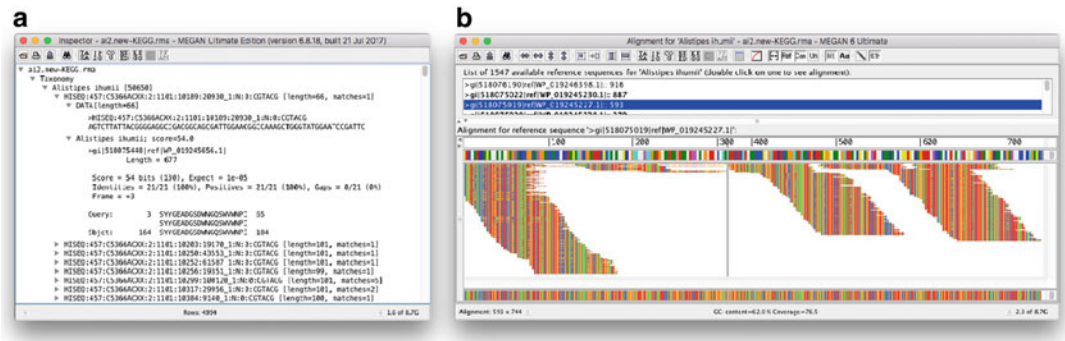


Fig. 2 (**a**) The Inspector Viewer showing some reads that have been assigned to *Alistipes ihumii*. (**b**) The Alignment Viewer showing reads aligned to a reference sequence

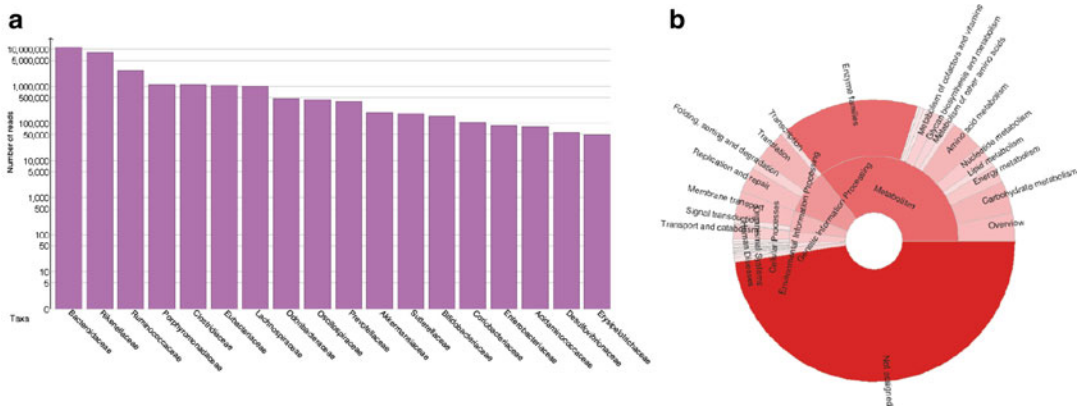


Fig. 3 (**a**) Bar chart of taxonomic assignments on family level, sorted by abundance. (**b**) Radial chart of functional assignments to KEGG for the same sample from [14]

possible to show the alignment of all of those reads on the reference. This can be useful, say, to determine how much of a reference gene is covered by reads.

Apart from being able to investigating taxonomic diversity, the advantage of using metagenomic sequencing to study an environmental sample is the ability to study the functional potential of the community. MEGAN currently provides four different functional classification systems for this purpose: InterPro & GO, eggNOG, KEGG, and SEED.

Each functional classification is displayed as a tree. The nodes of the tree can be investigated very much like the nodes of the taxonomic tree. Abundances can be visualized using different visualization options from simple bar charts over box plots and heat maps to radial tree charts drawn based on the abundances of the selected nodes. Two examples show charts that are shown in Fig. 3.

Alignments or reads matching a selected function can be exported to a text file or extracted to a new MEGAN document. This makes it possible to study only a part of a microbial community that is of particular interest. For example, if you select nodes associated with antibiotic resistance genes, you can determine which taxonomic assignment the reads assigned to antibiotic resistance genes have. An example of this is shown in Fig. 4.

If you want to study the full gene sequence of proteins found in your samples and be able to compare variants of those genes, it can be helpful to use gene-centric assembly [15]. Gene-centric assembly uses the alignments to reference proteins to assemble the matching reads. One can thus obtain the gene sequences from different organisms found in a sample for further analysis steps.

We will introduce more possibilities for studying the taxonomic and functional diversity of multiple samples in comparison in Subheading 3.

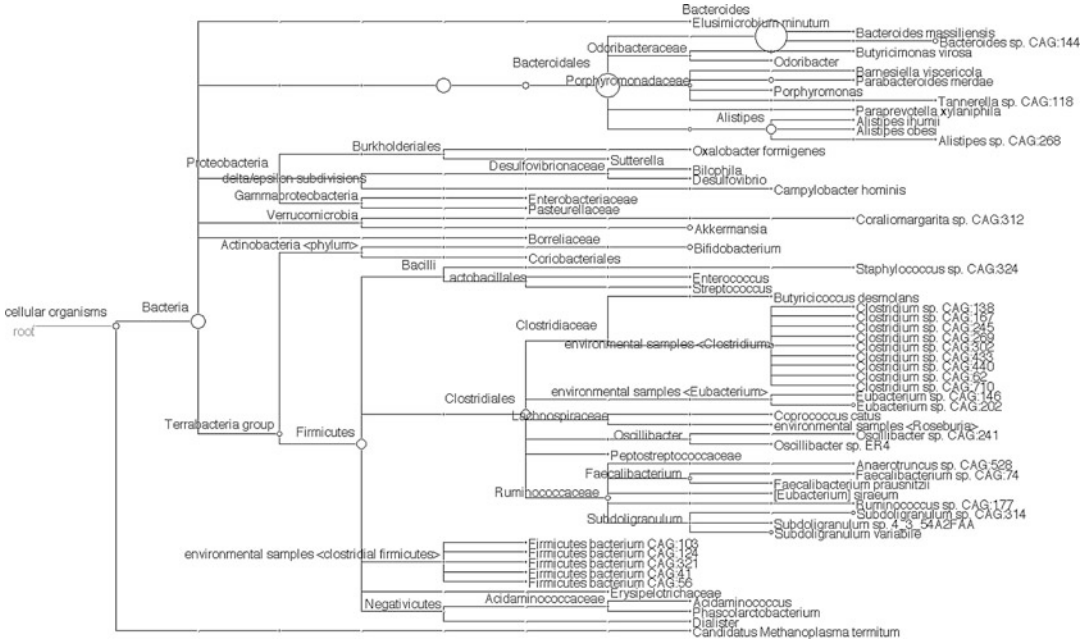


Fig. 4 Taxonomic assignment of reads from the day 0 sample for "Alice" from the ASARI [14] dataset which have been assigned to "resistance of fluoroquinolones" in the SEED hierarchy

2.2 Long Read Pipeline

As presented in the previous section, using metagenomic short reads, one can assembly gene sequences and obtain variants of a single gene using a gene-centric assembly, or of course use other assembly techniques. However, using short read data, it is very difficult to establish whether different genes are present in the same organism. We can connect the genes if they are found on a single DNA molecule with long sequencing reads, provided by third generation sequencing technologies such as PacBio [16] or Oxford Nanopore [17].

The PacBio and Nanopore devices can produce reads that are hundreds of thousands of bases long, with error rates of around 10%, say [17]. In contrast to short reads, which each can be safely assumed to overlap with only a single gene, long read will usually overlap or contain multiple genes. Hence, many popular short read alignment and analysis algorithms may require modification so as to take into account that a given read can align to multiple genes.

2.2.1 Long Read Analysis Pipeline

The basic long read analysis pipeline is analogous to the above described short read pipeline, and consists of the alignment and MEGAN analysis steps (Fig 5), but the details of the analysis pipeline as well as some components of MEGAN6 differ from the short read solution.

As described in the following, for long reads alignment is performed using LAST, processing of the alignments requires an additional step and MEGAN provides some modified algorithms for processing and visualizing long reads.

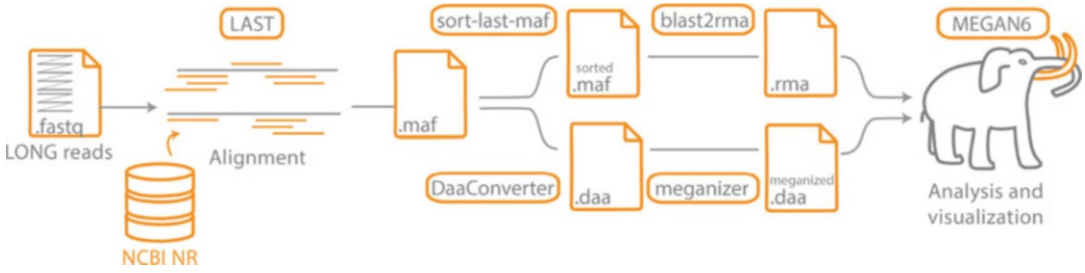


Fig. 5 Basic pipeline for long read analysis

```
Score = 86 bits (159), Expect = 7e-13
Identities = 34/37 (92%), Positives = 34/37 (92%), Gaps = 2/37 (5%)
Frame = -1
Query:  1080    EAVMVLSLDAEA\LVGYRE/KFPAWMDADRFEIKPRK   976
                EAVMVLSLDAEA LV YRE KFPAWMDADRFEIKPRK
Sbjct:  232     EAVMVLSLDAEA-LVRYRE-KFPAWMDADRFEIKPRK   266
```

Fig. 6 A frame-shift aware DNA-to-protein alignment produced by LAST

2.2.2 Alignment Using LAST

Third generation sequencing technologies produce much longer reads, with a higher error rate (approximately 10%, mostly insertions and deletions). Most DNA-to-protein aligners (such as BLASTX [7] or DIAMOND) translate the complete DNA query sequence in all six reading frames and then align the translated sequences against the protein database. Insertions or deletions in long reads cause a frame-shift and break translation-based alignments. LAST is a frame-shift aware aligner that incorporates single-base insertions or deletions into the alignment calculation. These are represented as "\" for forward-shifts and "/" for reverse-shifts, as shown in Fig. 6.

LAST, when used with large databases, such as NCBI-nr, splits the database into several volumes and indexes them individually. Similarly the large input files are loaded in separate volumes, and each volume of input is searched against each volume of the database. LAST, by default, generates output in MAF, "Multiple Alignment Format."

2.2.3 Taxonomic and Functional Classification of Long Reads

Because of processing both the query and database in different volumes and writing the output as soon as it is generated, the alignments for a single read appear in different parts of the MAF output of LAST. MEGAN processes alignment files line-by-line, identifies all alignments of a single read, and then assigns that read to a taxonomic and/or functional class. The unordered structure of LAST output prevents MEGAN from doing this. Thus, MAF files produced by LAST must be sorted before they are imported to MEGAN. For this task, MEGAN provides a command-line script, called *sort-last-maf*.

Alternatively, the user can use *DAA_Converter* (available at http://github.com/BenjaminAlbrecht84/DAA_Converter), which converts a given MAF file to a DAA file. This has several advantages, including space compression and faster processing. Additionally, the output of LAST can directly be piped into *DAA_Converter* which will then convert the output into a DAA file as LAST continues to operate. The trade-off when using *DAA_Converter* currently is that the alignments are filtered out with the default settings in MEGAN6 and resulting DAA file only has the alignments that would pass the filter, making it impossible to change filtration parameters without running LAST again once the conversion is done.

Similar to short reads, these long read MAF and DAA can then be imported into MEGAN and each read will get assigned to a taxon and/or functional class(es) of any provided functional hierarchy. The filtration based on bit-score of alignments work differently for long reads. In case of short reads, the alignments are filtered globally—only those that are within top 10% (by default) of the best-scoring alignment are taken into account. For long reads, this filtration is applied to each "gene" separately, as one long read can contain many different genes along its length. The alignments that overlap significantly (>90% by default) are grouped into *segments*, denoting different genes, and each interval is then processed individually in the filtering step.

The LCA algorithm to assign reads to taxonomic classes is also modified for long reads. As there are multiple genes on a single long read, and each of them may be conserved in different clades of the taxonomic tree, the *naïve LCA* is usually uninformative. Instead long reads are assigned to the most specific taxon that covers more than a fixed percentage (>80% by default) of every base pair that has an alignment. This algorithm assigns reads specifically to lower levels of taxonomy as long as they cover a gene which has low level conservation, other taxa gets lower percentages of coverage. Functional classification of long reads does not necessarily assign each read into one functional class, instead reads are assigned to the functional class of best-scoring alignment in each *segment*, thus each segment is assigned to one function and one read can be assigned to multiple different functional classes.

2.2.4 Investigation of the Results

The first view the user gets when a long read dataset is loaded in to MEGAN6 is identical to that of a short read dataset; however, there are some underlying differences and several investigation modes designed specifically for long reads.

Due to a large variability of read length of long reads [18], it is impractical for MEGAN to report number of reads assigned to class as a mean of abundance. Using the raw read length is also not feasible for Nanopore technology as reads tend to have "head"

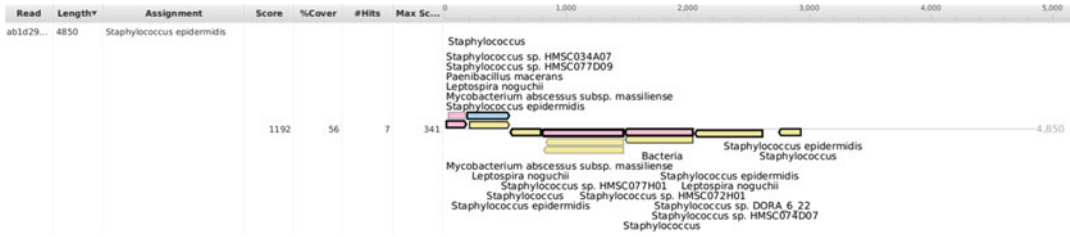


Fig. 7 Long Read Inspector in MEGAN6. The read is drawn as a line in the middle and the protein alignments are drawn as arrows on their corresponding positions and strands on the read

and "tail" regions composed of random bases [19] (Fig. 7 shows a read whose tail region has no significant alignment to any protein in the database). Thus, the default mean of reporting the abundance for a particular taxon or functional class in long read pipeline is the number of aligned bases.

The number of alignments on a long read can easily exceed hundreds and complicates the Alignment Viewer and the Inspector features of MEGAN6. In order to simplify the investigation of alignments on the reads, MEGAN6 offers a *Long Read Inspector* window (Fig. 7), accessible via right-click on any of the nodes in the main view. This inspector draws reads as horizontal lines and alignments as arrows on their corresponding positions. The names of taxa or functional classes are also linked to these alignment arrows.

The Inspector Window helps particularly in the case of suspicious assignments. Figure 8a shows the inspector view for a read that was assigned to *Trichuris trichiura*, a human parasitic whipworm, in a sample of known mixture of microorganisms [20]. A closer inspection to Fig. 8a lets us see that, although the read is spanned by several alignments from *Escherichia coli*, it is assigned to *T. trichiura* because the total length of alignments to *T. trichiura* is longer than 80% whereas it is below that for *E. coli* and all other competing taxa.

For further analysis of such suspicious assignments, MEGAN6 offers a remote BLAST function, in which selected reads are aligned against a selected database (such as the nucleotide collection—NCBI nt) on the NCBI website and the resulting assignments are captured, processed, and presented in a new MEGAN document. In Fig. 8b, we see that our "suspicious" read is assigned to *E. coli*, which was in the known mixture of microorganisms, based on remote NCBI-BLAST against NCBI nt.

Similar to exporting alignments and reads as explained in the previous section, these can also be exported in general feature format (GFF) for downstream analysis. This provides a simple way of obtaining the annotation, especially for long reads and contigs. The annotations exported to the GFF files contain the accessions of

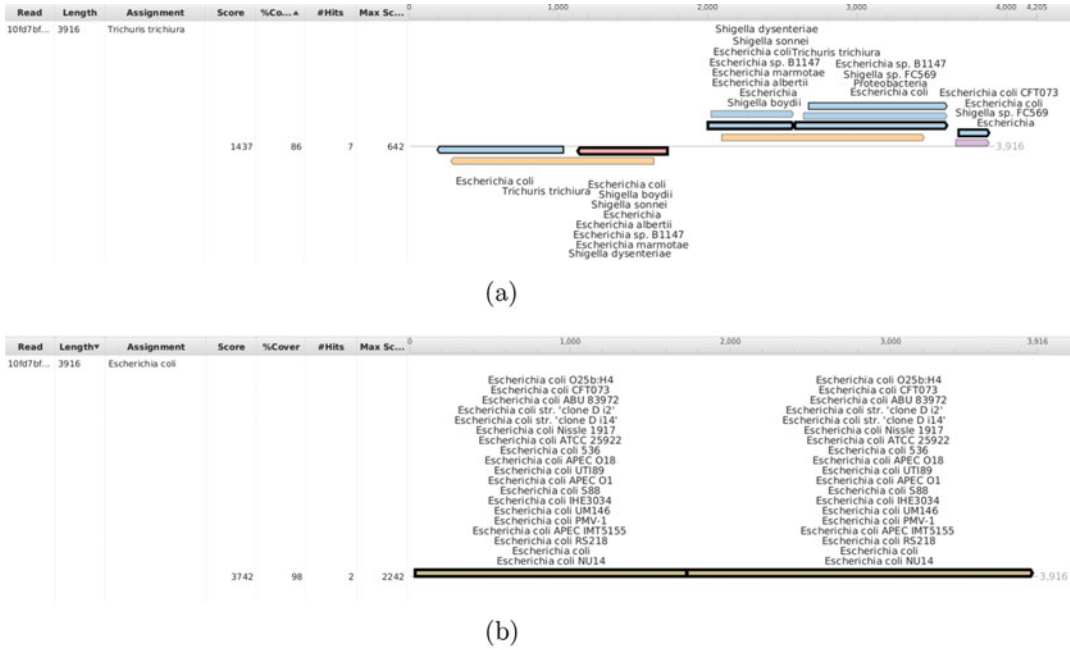


Fig. 8 MEGAN6 offers a remote BLAST functionality, namely "BLAST on NCBI," which can be used for suspicious assignments. (**a**) Long Read Inspector view for a read assigned to *Trichuris trichiura*, based on protein alignments against NCBI nr. (**b**) Long Read Inspector view for the same read as in (**a**), assigned to *Escherichia coli*, after searching it against nucleotide collection of NCBI using the remote BLAST functionality of MEGAN6

references and their corresponding taxonomic and/or functional mappings depending on which mapping files were used during importing the dataset into MEGAN.

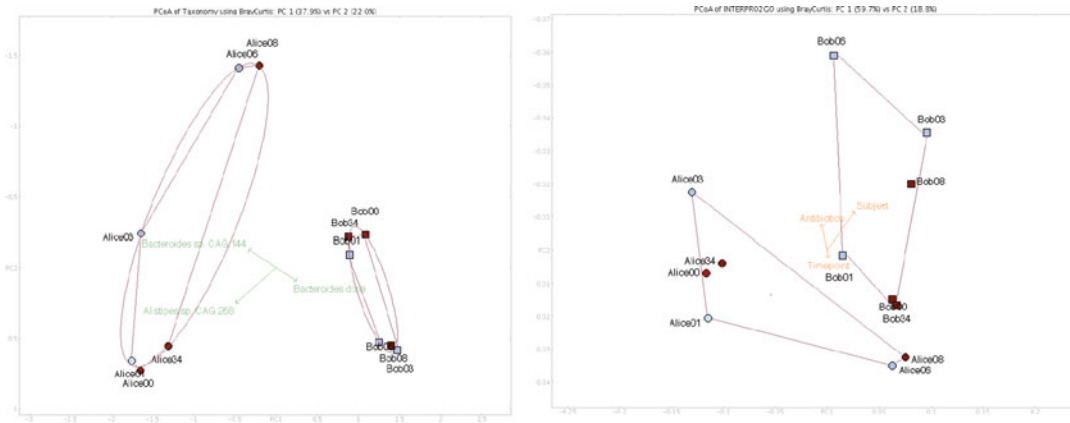
3 Comparison of Multiple Samples

Most modern metagenomics experiments include the collection and analysis of multiple samples to compare different groups with controls or study the dynamic changes of a microbial community over time. Hence, a very important feature of MEGAN is the ability to load multiple datasets into a single "comparison document" (megan file). This is a light-weight file that does not contain the original reads and alignments, but allows one to compare the taxonomic and functional diversity of multiple samples.

To be able to easily compare groups of samples and relate findings to features attached to samples, it is helpful to import metadata. Metadata should be provided in tabular format and connect the sample IDs to attributes whose values can be text, numeric, or boolean values. Using this information you can group samples in different visualizations. For example, this allows

easier interpretation of the principal component analysis (PCoA) plots in MEGAN. Principal components can be calculated using different distance measures including Bray–Curtis or simple Euclidean distances. MEGAN can include bi-plots and tri-plot vectors into the PCoA plot, which represent the top taxonomic or functional classes and metadata features, respectively, that correlate most with the differences between samples. Figure 9 shows multiple examples of PCoA plots including bi-plot and tri-plot vectors.

MEGAN can also calculate and visualize co-occurrence and correlation plots. For correlation there are two options. The first is useful for time series analysis, because it calculates correlations between different taxa. This can be used to determine how changes in abundance of one taxon influence changes in another, which makes it possible to detect potential interactions between taxa. To distinguish the effect of interactions between taxa from it being caused by an external influence, it is useful to check out the other attribute correlation plot, which calculates correlations between taxa and metadata. So, if, for example, two taxa are correlated to each other and correlated to the same external influence from the metadata, then they might be less likely to be influencing each other, but are perhaps both influenced by the same attribute of the metadata. An example of an attribute correlation plot is shown in Fig. 10.



(a) Taxonomic (b) Interpro2GO

Fig. 9 PCoA analysis of 12 samples associated with "Alice" (round shapes) and "Bob" (square shapes), from [14]. Time points of antibiotic intake are colored light blue, time points before and after antibiotic intake dark red. (**a**) A PCoA plot based on Bray–Curtis distances as calculated by MEGAN using the taxonomic abundances for the samples. The green vectors represent the bi-plot vectors. The samples are grouped by individual, showing the convex hulls of the groups as well as ellipses. (**b**) is based on the same data but using the abundances of GO terms in the InterPro2GO hierarchy and only showing the convex hulls of the group. Here the orange vectors are the tri-plot vectors, showing the relation of metadata values to the principal components

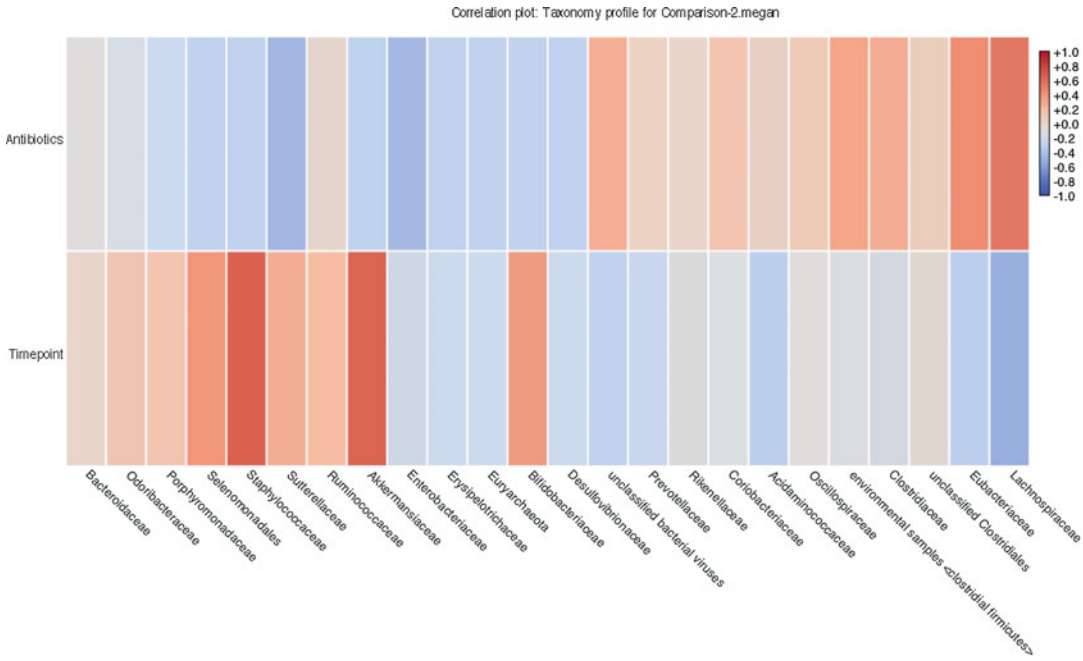


Fig. 10 Attribute correlation plot for the data from [14] for two healthy individuals taking antibiotics for 6 days (day 1–6). Correlation is shown as a heat map with red marking positive correlation between the attribute and the taxon and blue marking negative correlation. Correlations are shown for antibiotics intake (boolean) and time (day 0, 1, 3, 6, 8, and 34)

4 Outlook

It goes without saying that the quality and quantity of the input sequencing data limits the reliability of the output analysis. More directly, quality of the MEGAN hierarchy assignments is determined by the quality of the read alignment, which, in turn, depends on the chosen database and alignment tool. On the one hand, the database needs to be well annotated and comprehensive, as it is only possible to analyze the organisms or entities present in it. On the other hand, the alignment tool needs to be sensitive in order to identify the matching sequence. It is especially difficult to deal with sets of very similar sequences. Currently, for the human gut microbiome sequencing data analyzed with the basic short read pipeline, as much as 30% of reads are not assigned to any node in the course of the taxonomic analysis.

In order to avoid the bias introduced by the database one can also use one of the database-free strategies, e.g., k-mer counting. They are good for tracking the global changes in the data, but it is difficult to correct for possible contaminations. Although MEGAN does not support this type of analysis, it enables global comparisons with PCoA based on the profiles computed for each of the samples.

Another approach is assembly based analysis. In brief, the reads are assembled and then the scaffolds or contigs are annotated and investigated. This approach provides some information on gene co-localization at a cost of data loss in the form of unassembled reads and short contigs. Full metagenomic read assembly [21] is a very complex and computationally expensive task that MEGAN does not address.

Application of the long read sequencing technologies opens new perspective for metagenomics analysis. Long reads provide information on gene co-location on a single DNA molecule, and make assembly much easier. But, long reads also pose new algorithmic challenges in aspects of the protein alignment, hierarchy assignment, and abundance computation. As long read technologies continue to evolve, so, too, must the corresponding analysis algorithms.

MEGAN is a powerful visual analytics tool that provides a wide range of the algorithms for analysis of metagenomics sequencing data. MEGAN can run on hundreds of samples along with hundreds of metadata columns. It is the main workhorse of the Tubiom project where metagenomics profiles of 10,000 volunteers are collected and mined for correlations with the vast metadata (www.tuebiom.de).

4.1 MEGAN Resources

MEGAN Community software is freely available on the website: ab.inf.uni-tuebingen.de/data/software/megan6, together with the current mapping files for taxonomic and functional analysis.

Short read datasets presented in this chapter and used for visualizations are publicly accessible in MEGAN via MeganServer. The dataset used in the Long Read Pipeline section was downloaded from the supplementary material of Brown et al. [20]. Instructions for use of MEGAN and user support can be found on the MEGAN community website (megan.informatik.uni-tuebingen.de).

References

1. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68(4):669–685
2. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60
3. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21 (3):487–493
4. Sheetlin SL, Park Y, Frith MC, Spouge JL (2014) Frameshift alignment: statistics and post-genomic applications. Bioinformatics 30 (24):3575–3582
5. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R (2016) MEGAN community edition–interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol 12(6):e1004957. https://doi.org/10.1371/journal.pcbi.1004957
6. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH (2016) MALT: fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv 050559. https://doi.org/10.1101/050559
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

8. Geer LY, Marchler-bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH (2010) The NCBI biosystems database. Nucleic Acids Res 38(2009):492–496
9. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD (2014) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(D1):D213–D221
10. Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, Jones P, Leinonen R, McAnulla C, Maguire E, Maslen J, Mitchell A, Nuka G, Oisel A, Pesseat S, Radhakrishnan R, Rocca-Serra P, Scheremetjew M, Sterk P, Vaughan D, Cochrane G, Field D, Sansone SA. EBI metagenomics–a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 42(D1):D600–D606
11. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 40(D1):D284–D289
12. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
13. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R The seed and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 42 (D1):D206–D214
14. Willmann M, El-Hadidi M, Huson DH, *et al* (2015) Antibiotic selection pressure determination through sequence-based metagenomics. Antimicrob Agents Chemother 59 (12):7335–7345
15. Huson DH, Tappu R, Bazinet AL, Xie C, Cummings MP, Nieselt K, Williams R (2017) Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. Microbiome 5 (1):11
16. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma X, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138
17. Mikheyev AS, Tin MY (2014) A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour 14(6):1097–1102
18. Laver T, Harrison J, ONeill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quant 3:1–8
19. Yang C, Chu J, Warren RL, Birol I (2017) NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience 6(4):1–6
20. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB (2017) MinION nanopore sequencing of environmental metagenomes: a synthetic approach. GigaScience 6(3):1–10
21. Medvedev P, Georgiou K, Myers G, Brudno M (2007) Computability of models for sequence assembly. Gene 4645:289–301