METAGENOME ANALYSIS USING MEGAN

DANIEL H. HUSON and ALEXANDER F. AUCH
Center for Bioinformatics, Tübingen University, Sand 14, 72076 Tübingen, Germany

QI JI and STEPHAN C. SCHUSTER
310 Wartik Laboratories, PennState University, Center for Comparative Genomics, Center for Infectious Disease Dynamics, University Park, PA 1803, USA

In metagenomics, the goal is to analyze the genomic content of a sample of organisms collected from a common habitat. One approach is to apply large-scale random shotgun sequencing techniques to obtain a collection of DNA reads from the sample. This data is then compared against databases of known sequences such as NCBI-nr or NCBI-nt, in an attempt to identify the taxonomical content of the sample. We introduce a new software called MEGAN (Meta Genome ANalyzer) that generates species profiles from such sequencing data by assigning reads to taxa of the NCBI taxonomy using a straight-forward assignment algorithm. The approach is illustrated by application to a number of datasets obtained using both sequencing-by-synthesis and Sanger sequencing technology, including metagenomic data from a mammoth bone, a portion of the Sargasso sea data set, and several complete microbial test genomes used for validation proposes.

1. Introduction

Genomics is the study of the genome sequence of individual organisms. Most genome sequences available in databases today were obtained by "Sanger sequencing", using a shotgun approach that involves cloning small inserts of DNA and then determining their sequence using fluorescent dideoxynucleotides for termination and electrophoresis for measurement[7]. The NCBI website (`www.ncbi.nlm.nih.gov`) lists hundreds of bacterial, tens of archaeal and about one hundred eukaryotic genomes as being completely sequenced, or in the process of being sequenced.

Metagenomics has been defined as "the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms"[5], and its importance stems from the fact that 99% or more of all microbes are deemed unculturable. If we take a genome to be the entire genetic information of a single organism, then a *metagenome* can be defined as the entire genetic information of an ensemble of organisms, living in a common habitat. The aim of metagenomics is to understand the genetic diversity of a metagenome, ideally, by identifying the (relative abundances of) species present. Metagenomics promises to lead to the discovery of new genes that have useful applications in biotechnology and medicine[10].

One main technique in metagenomics is to apply large-scale random shotgun sequencing. A number of recent projects use Sanger sequencing to create datasets in this way, for

example, from an acid mine biofilm[12], sea-water samples [13], deep sea sediment[4], or soil and whale falls[11]. Recently, a new sequencing approach "sequencing-by-synthesis" was published that uses emulsion-based PCR application of a large number of DNA fragments and high-throughput parallel pyro-sequencing[6]. A single instrument is able to sequence 25 million bases within four hours, at a lower price, per base, than Sanger-based methods. Current drawbacks of the method are short read lengths of ≈ 100bp, in contrast to ≈ 800bp using Sanger sequencing, and a higher error rate. Moreover, sequencing of pair-ended reads is not yet possible.

Given present-day technology, obtaining the complete sequences of all genomes present in a metagenome is not feasible, even using Sanger sequencing and paired-end reads, as the amount of data required is too large and the assembly problem too difficult. More realistic goals are to determine the presence or absence of specific species of interest, or to obtain a rough overview of the taxa represented in a given metagenome.

In this paper we present a straight-forward approach to the latter problem. We describe a strategy for processing DNA reads collected within the frame-work of a metagenomics project and provide a new program called *MEGAN* (MEtaGenome ANalyser) that can be used to explore a metagenomics data set in a taxonomical context. The program employs a combinatorial algorithm, which we call "LCA-assignment", to estimate the taxonomical content of a metagenome, based on sequence comparisons.

We first illustrate this approach by application to a set of $302,692$ reads obtained from a sample of mammoth bone[8], using the sequencing-by-synthesis approach. We then address the question whether species can be identified with confidence from short reads of length 100. Finally, to demonstrate the applicability of the approach to data sets obtained using other sequencing approaches, we apply it to a subset of the Sargasso sea data[13].

Ease-of-use is a main design criterion of MEGAN. An analysis is initiated by simply opening a BlastX, BlastN or BlastZ file and is then performed interactively. For maximum portability, the program is written in Java and installers for Linux/Unix, MacOS and Windows are freely available for academic use from:
`http://www-ab.informatik.uni-tuebingen.de/software/megan`.

2. Processing metagenomic data

The following simple approach to metagenome analysis is a typical starting point for more sophisticated strategies (see Figure 1): First, randomly sequence a collection of DNA reads from the given sample. Second, perform Blast[1] comparisons of the reads against one or more reference databases, such as NCBI-nr, NCBI-nt, NCBI-env-nr, NCBI-env-nt[2], and additional genome specific databases, when appropriate. (Sequence comparison is the main computational bottle neck, which will grow more serve as the sizes of datasets and databases continue to grow.) Third, analyze the output of these comparisons and then assign individual reads to taxa, including higher-order taxa. Finally, for each taxon implicated, evaluate the provided evidence for its presence in the sample.

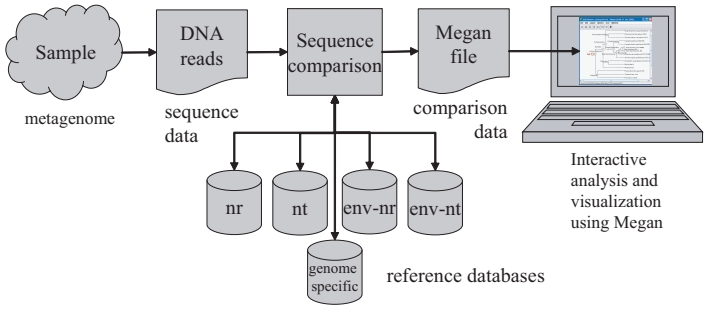


Figure 1. For a given sample of organisms, a randomly selected collection of DNA fragments is sequenced. The resulting reads are then compared with one or more reference databases using an appropriate sequence comparison program such as Blast. The resulting data is processed by MEGAN to produce an interactive analysis of the taxonomical content of the sample.

2.1. *Analysis of the Mammoth dataset*

As an example, we recently[8] used a metagenomics approach to analyze the DNA present in a sample of one gram of bone taken from a mammoth that was preserved in permafrost for 27,000 years. The project proceeded in the following steps. First, we used the Roche GS20 sequencing technology to randomly collect DNA from the sample, obtaining $302,692$ reads of mean length 95 base pairs (bp). We will refer to this as the *Mammoth dataset*.

To identify those reads that come from the mammoth genome, we performed BlastZ[9] comparisons of genome sequences for elephant, human and dog, downloaded from `www.genome.ucsc.edu`. As a result of this computation, in the mentioned paper[8] we estimate that at least 45.4% of the reads represent mammoth DNA.

We were interested in determining the possible sources of the remaining reads, as they probably represent micro-organisms that were present at, or immediately after, the time of the mammoth's death. To this end, we first used BlastX to compare all reads against the NCBI-nr ("non-redundant") protein database[2]. This resulted in a file of size 1.4GB containing $2,911,587$ local alignments of reads to sequences in the database. Of the $302,692$ reads, only $52,179$ have one or more alignments. We then loaded the results of the BlastX search into a preliminary version of MEGAN (then called GenomeTaxonomyBrowser[8]) and applied the LCA-assignment algorithm to compute an assignment of reads to taxa, thus obtaining an estimation of the taxonomical content of the sample.

Here we repeat this analysis, but are slightly more conservative and now employ a threshold of 30 for the bit score of matches, and will discard any isolated assignments, that is, any taxon that has only one read assigned to it. (We remove isolated assignments to avoid false positive identification of taxa due to sequencing errors or chance matches.) The LCA-assignment algorithm assigns $50,093$ reads to taxa and $2,086$ remain unassigned either because the bit score of their matches fall below the threshold or because they give rise to an isolated match.

A total of $19,841$ reads are assigned to Eukaryota, of which $7,969$ are assigned to Gnathostomata (jawed vertebrates) and thus presumably come from mammoth genes. Fur-

ther, a total of $16,972$ reads are assigned to Bacteria, 761 to Archea and 152 to viruses, respectively. These numbers are slightly lower than previously reported[8] due to our slightly more conservative settings. MEGAN can be used to summarize the results at different levels of the NCBI taxonomy, see Figures 2 and 3.

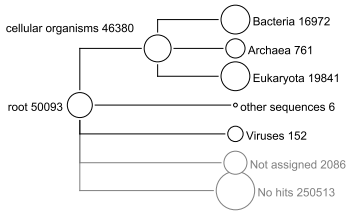


Figure 2. High-level summary of a MEGAN analysis of the mammoth dataset, based on a BlastX comparison of the $302,692$ reads against the NCBI-nr database. In all figures, each circle represents a taxon in the NCBI taxonomy and is labeled by its name and the number of reads that are assigned either directly to the taxon, or indirectly via one of its sub-taxa. The size of the circle is scaled logarithmically to represent the number of reads assigned directly to the taxon.

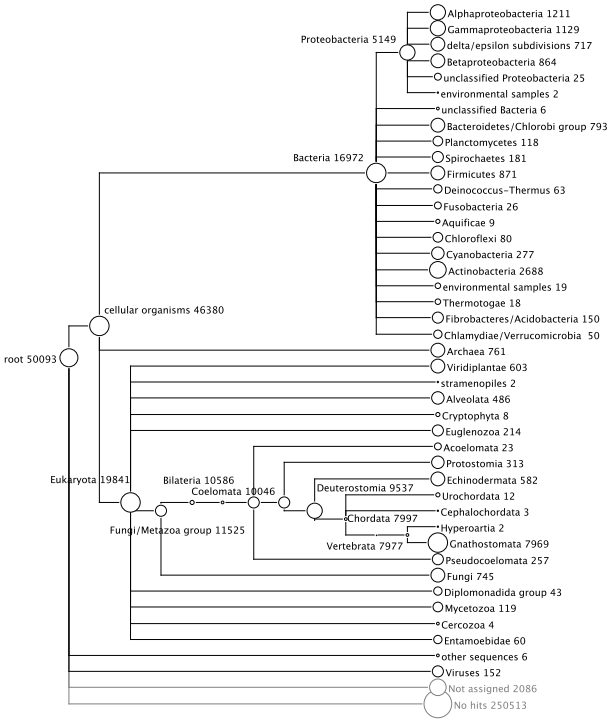


Figure 3. A more detailed MEGAN analysis of the mammoth dataset.

2.2. *Identifiability of species from short reads*

The average read length currently obtainable using Roche GS20 sequencing technology is $\approx$ 100bp. The question arises whether this sequence length is long enough to provide meaningful information on the taxonomical content of a metagenome. This can be addressed by collecting a set of reads from a known genome and then processing the data as a metagenome dataset.

In a first experiment, we considered 2000 reads sequenced from *E. coli K12*, using Roche GS20 sequencing technology. As *E. coli* is widely used in laboratories, this dataset may potentially give rise to many false positive identifications, as parts of its sequence occur by error in a number of different genome sequences.

In Figure 4 we show the resulting MEGAN analysis, based on a BlastX comparison of the reads against the NCBI-nr database, using a bit score threshold of 35 and discarding any isolated assignments. Of the 2000 reads, approximately 25% (448) have no hits and 116 reads are not assigned. Of the remaining 1436 reads, approximately 50% (699) are assigned to *Enterobacteriaceae*, thus making a correct assignment up to the family level. All other reads, except two, are assigned to super taxa, thus producing correct, if increasingly weak, predictions.

The two false positive assignments to *Haemophilus somnus* appear to be due to false entries in the NCBI-nr database: one of the assigned reads has a perfect BlastN match to 16S rRNA sequence in *E. coli* and the other has a perfect BlastN match to 23S rRNA sequence in *E. coli*. On the other hand, the matched sequences representing *Haemophilus somnus* in NCBI-nr are both labeled "hypothetical" proteins.

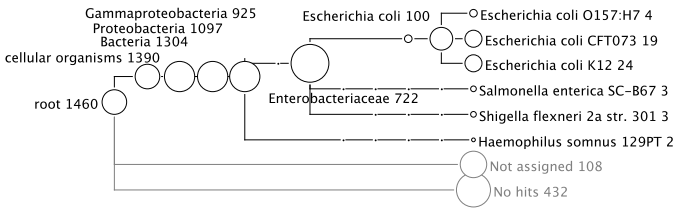


Figure 4. MEGAN analysis of 2000 reads collected from *E. coli K12* using Roche GS20 sequencing, based on a BlastX comparison with the NCBI-nr database.

In a second experiment, we considered 2000 reads sequenced from *Bdellovibrio bacteriovorus HD100* using Roche GS20 sequencing technology. In Figure 5(a) we show the resulting MEGAN analysis, based on a BlastX comparison of the reads against the NCBI-nr database, using a bit score threshold of 35 and discarding any isolated assignments. Of the 2000 reads, approximately 20% (397) have no hits and 5% (105) are not assigned. Of the remaining 1498 reads, approximately 70% (1062) are assigned to *Bdellovibrio bacteriovorus HD100*. All other reads are assigned to super taxa, thus producing correct, if increasingly weak, predictions. There are no false positive predictions.

In Figure 5(b) we show the MEGAN analysis obtained when using a copy of the NCBI-

nr database from which all sequences representing *Bdellovibrio bacteriovorus HD100* have been removed. This mimics the case in which reads are obtained from a genome that is not represented in the database. Of the 2000 reads, approximately 65% (1361) have no hits and approximately 15% (272) are not assigned. A small number of false positives occur up to the level of bacteria.

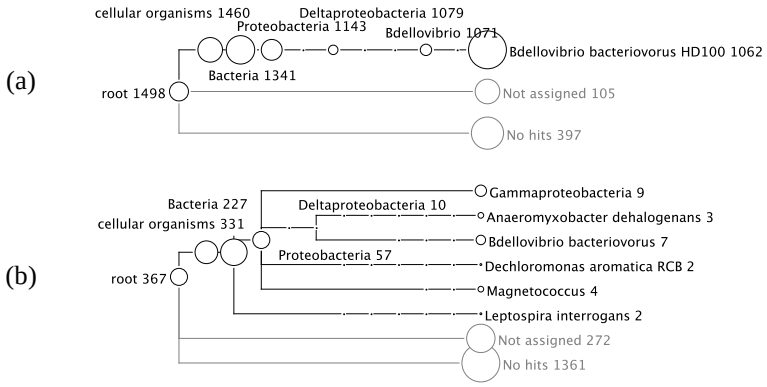


Figure 5. MEGAN analysis of 2000 reads collected from *Bdellovibrio bacteriovorus HD100* using Roche GS20 sequencing technology. (a) Analysis based on a BlastX comparison with NCBI-nr. (b) Similar analysis, but with all hits to database sequences representing *Bdellovibrio bacteriovorus HD100* removed, mimicking the situation in which the reads originate from a genome that is not represented in NCBI-nr.

These two experiments show that the LCA-assignment algorithm is quite conservative, avoiding false positive assignments at the price of producing quite large numbers of inspecific assignments. Further, they also indicate that the performance of this type of approach depends heavily on the set of sequences represented in the reference database. In particular, if close relatives are missing in the database, then reads from an unknown organism will give rise to many unassigned reads and possibly some false positive assignments, as well.

2.3. *Analysis of the Sargasso Sea data set*

In the Sargasso sea project[13], samples of sea water were collected and organisms of size $0.1 - 3.0$ μm were extracted to produce a metagenome dataset. In 4 separate experiments, approximately 1.9 million reads of average length 818 bp were collected using Sanger sequencing.

To explore the application of MEGAN to such data, we downloaded the first $10,000$ reads from `https://research.venterinstitute.org/sargasso/` and ran BlastX to compare the data against the NCBI-NR database. Only 1% (13) of the reads had no hits. A MEGAN analysis of the data using a bit score threshold of 100 and discarding all isolated assignments, assigned approximately 90% ($8,977$) to taxa, a majority of which (6811) were assigned to bacteria. The results are summarized in Figure 6. Interestingly, this analysis of a small portion of the Sargasso sea dataset is compatible with the analysis reported by Venter *et al.*[13], (although *Firmicutes* are missing, probably due to

the small size of the sub-sample), and also confirms findings that parts of the data set is contaminated with *Shewanella* and *Burkholderia*[3].

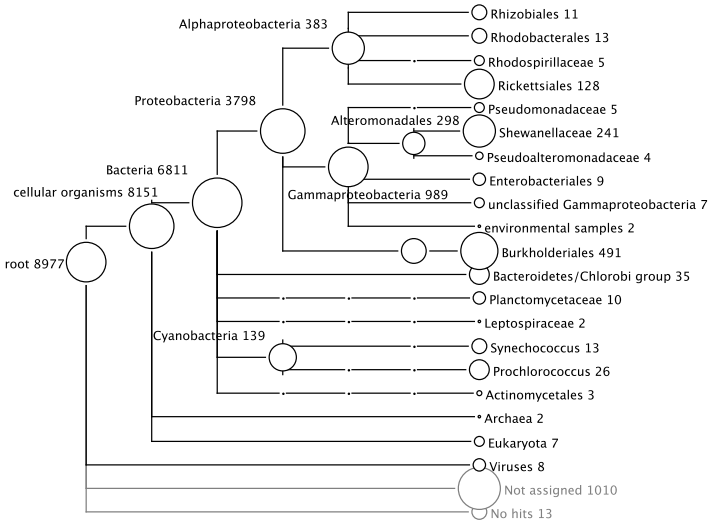


Figure 6. MEGAN analysis of $10,000$ reads of Sargasso sea data.

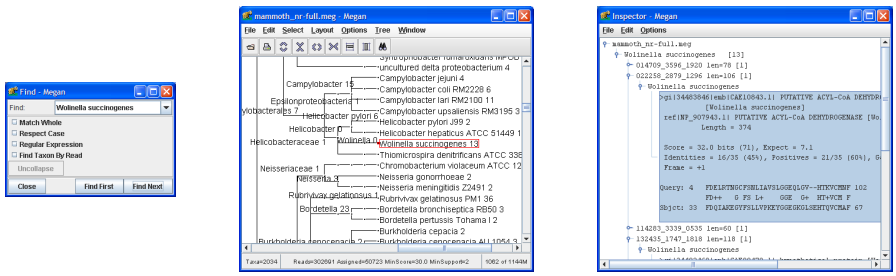
3. Analysis using MEGAN

At startup, MEGAN loads the complete NCBI taxonomy, currently containing over $280,000$ taxa, which can then be interactively explored, using customized tree-navigation features. However, the main application of MEGAN is to process results files generated by a comparison of sequencing reads with a database of annotated sequences. The program parse files generated by BlastX, BlastN or BlastZ, and saves the results as a series of read-taxon matches in a program-specific format. (Additional parsers may be added to process the results generated by other sequence comparison methods.)

The program assigns reads to taxa using the LCA-assignment algorithm (described in detail below) and then displays the induced taxonomy. Nodes in the taxonomy can be collapsed or expanded to produce summaries at different levels of the taxonomy. Additionally, the program provides a Find tool to search for specific taxa and an Inspector tool to view individual Blast matches (see Figure 7).

The approach uses a number of thresholds. First, a *min-score* threshold defines the minimum bit score that must be attained by a Blast alignment so that a read r is considered to *match* a given taxon t. Second, the *min-support* threshold specifies how many reads must be assigned to a specific taxon, or any taxon below it in the taxonomy, so that the taxon is identified as present.

The result of the LCA-assignment algorithm is presented to the user as the partial taxonomy T that is induced by the set of taxa that have been identified (see Figure 2). The



(a) (b) (c)

Figure 7. (a) MEGAN provides a Find tool to search for specific taxa of interest. (b) The result of a search is highlighted in a detailed summary of the analysis. (c) MEGAN provides an Inspector tool to view the individual sequence comparisons upon which the assignment of a particular read to a particular taxon is based.

program allows the user to explore the results both at a high level, and also a very detailed level, by providing methods for collapsing and expanding different parts of T. Each node in T represents a taxon t and can be queried to determine which reads have been assigned directly to t, and how many have been assigned to taxa below t. Additionally, the program allows the user to view the Blast alignments upon which a specific assignments is based (see Figure 7(3)).

4. Assignment of reads to taxa

MEGAN currently uses a simple combinatorial algorithm, which we call "LCA-assignment", in association with a number of different thresholds, to assign each read to a taxon at some level of the NCBI taxonomy.

The *LCA-assignment* algorithm operates as follows. Consider a read r and assume that the Blast computation has established matches to sequences representing a set of taxa $t(r) = \{t_1, t_2, \ldots, t_k\}$. We assign the read r to the *lowest common ancestor (LCA)* of $t(r)$ in the NCBI taxonomy. For example, if r matches *Campylobacter lari*, *Helicobacter hepaticus* and *Wolinella*, then r is assigned to the taxon *Campylobacterales*. If r does not match any sequence in the given reference database, that is, if $t(r) = \emptyset$, then r is assigned to the special taxon *no hits*. If r cannot be assigned to a taxon for other reasons, e.g. the read only matches sequences for which the taxon is unknown, then r is assigned to another special taxon *Not assigned*.

In this way, each read r in the dataset is assigned to one or more NCBI taxa, or to one of either special taxa. If the Blast matches computed for r involve only one or a few closely related species, then r will be assigned to a taxon near the tips of the taxonomy. If, on the other hand, r matches a wider range of taxa, then r will be assigned to a higher-level taxon. The read may even be assigned to the root of the taxonomy, if the sequence is completely unspecific.

To implement the LCA-assignment algorithm, we assign a binary *address* $a(t)$ to each every t in such away that if taxon s is an ancestor of taxon t, then the address $a(s)$ is a prefix of the address $a(t)$. Using this scheme, we can easily determine the lowest common

ancestor of a set of n taxa $\{t_1, \ldots, t_n\}$ by determining the longest common prefix of the corresponding set of addresses $\{a(t_1), \ldots, a(t_n)\}$, in $O(n \times \log K)$ steps, where K is the maximum depth of the taxonomy.

4.1. *Under- and over prediction*

We say that a read gives rise to an *under prediction*, if it is assigned to a taxon that lies above the true taxon in the taxonomy. Under prediction happens when a read comes from a gene that is widely conserved. We say that a read gives rise to a *false prediction*, if it is assigned to a taxon that is not the true taxon, nor one of its ancestors in the taxonomy. W say that a false prediction is an *over prediction*, if it is caused by the fact that the true sequence is not represented in the employed databases.

For example, all reads analyzed in Figure 5(a)-(b) come from the genome of *Bdellovibrio bacteriovorus HD100*. However, there is a substantial amount of under prediction both in (a) and (b), in particular of the taxon *Bacteria*, and a number of cases of over prediction in (b), ranging from *Anaeromyxobacter dehalogenans* and *Gammaproteobacteria* to *Leptospira interrogans*.

As a simple combinatorial method, the LCA-assignment algorithm is susceptible to both types of errors. We hope to develop a more sophisticated approach that will not only take the presence or absence of matches into account, but also will make use of the quality of the matches and the levels of similarly that are typical for given genes in given clades of sequences.

5. Summary

A metagenomics project aims at understanding the taxonomical content of an ensemble of organisms. The approach described in this paper is to use sequencing techniques to produce DNA reads, to perform similarity searches in databases of known sequences and then to analyze and explore the resulting comparison data using software such as MEGAN.

MEGAN is based on a robust algorithm for assigning reads to taxa and is designed as an easy-to-use exploration tool that quickly produces summaries of the data at different taxonomical levels. It offers tools to search for specific taxa in the data and to inspect the evidence supporting the presence of any given taxon. This software provides a solid starting point for producing a first analysis of a metagenomic dataset.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
2. D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. Genbank. *Nucleic Acids Res*, 1(33 (Database issue)):D34–38, 2005.
3. E.F. DeLong. Microbial community genomics in the ocean. *Nat Rev Microbiol.*, 3(6):459–69, 2005.
4. S. J. Hallam, N. Putnam, C.M. Preston, J.C. Detter, and D. Rokhsar. Reverse methanogenesis: Testing the hypothesis with environmental genomics. *Science*, 305:1457–62, 2004.

5. J. Handelsman. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685, 2004.
6. M. Margulies and *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.
7. D. Meldrum. Automation for genomics, part two: Sequencers, microarrays, and future trends. *Genome Research*, 10(9):1288–1303, 2000.
8. H. N. Poinar, C. Schwarz, Ji Qi, B. Shapiro, R. D. E. MacPhee, B. Buigues, A. Tikhonov, L. P. Tomsho D.H. Huson, A. Auch, M. Rampp, W. Miller, and S. C. Schuster. Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science*, 331:392–394, 2006.
9. S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res.*, 13:103 – 107, 2003.
10. H.L. Steele and W.R. Streit. Metagenomics: Advances in ecology and biotechnology. *FEMS Microbiology Letters*, 247(2):105–111, 2005.
11. S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308:554–557, 2005.
12. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram1, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–43, 2004.
13. J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers, and H. O. Smith1. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, 2004.