

**Title: Making the most out of metagenomics – a guide from sampling to data analysis.**

Torsten Thomas<sup>1\*</sup>, Jack Gilbert<sup>2,3</sup>, Folker Meyer<sup>2,4</sup>

<sup>1</sup>School of Biotechnology and Biomolecular Sciences & Centre for Marine Bio-Innovation, The University of New South Wales, Sydney, NSW 2052, Australia

<sup>2</sup>Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A.

<sup>3</sup>Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.,

<sup>4</sup>Computation Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.,

\* corresponding author: email: [t.thomas@unsw.edu.au](mailto:t.thomas@unsw.edu.au) tel: +61-2-93853467

## **Abstract**

Metagenomics applies a suite of genomic technologies and bioinformatics tools to directly access the genetic content of entire communities of organisms. The field of metagenomics has been responsible for substantial advances in microbial ecology, evolution and diversity over the last 5-10 years and many research laboratories are actively engaged in it now. With the growing numbers of activities comes also a plethora of methodological knowledge and expertise that should guide the future developments in the field. This review aims to summarize the current metagenomic opinion, provide practical guidance and advice on sample processing, sequencing technology, assembly, annotation, experimental design, statistical analysis, data storage and sharing. As more metagenome data sets are generated the availability of standardized procedures as well as shared data storage and analysis become increasingly important to ensure output of individual projects can be assessed and compared.

## **Introduction**

Arguably, one of the most remarkable events in the field of microbial ecology in the last decade has been the advent and development of metagenomics. Metagenomics is defined as the direct genetic analysis of all genomes contained within an environmental sample. The field initially started with the cloning of environmental DNA followed by functional expression screening [1] and was then quickly complemented by direct random shotgun sequencing of environmental DNA [2, 3]. These initial projects did not only show proof-of-principle of the metagenomic approach, but also uncovered an enormous functional gene diversity in the microbial world around us [4].

Metagenomics provide access to the total functional gene composition of microbial communities and thus gives a much broader description than phylogenetic surveys that are often based only on the diversity of one gene, for instance, the 16S rRNA gene. On its own, metagenomics gives genetic information on potentially novel biocatalysts or enzymes, genomic linkages between function and phylogeny as well as evolutionary profiles of community function and structure. It can also be complemented with metatranscriptomic or metaproteomic approaches to describe expressed activities [5, 6]. Metagenomics is also a powerful tool for generating novel hypotheses of microbial function and the remarkable discoveries of proteorhodopsin-based photosynthesis or ammonia-oxidizing Archaea speak to this fact [7, 8].

The rapid and substantial cost reduction in next-generation sequencing has dramatically accelerated the development of sequenced-based metagenomics and in fact, the number of metagenome shotgun sequence datasets has exploded in the last few years. Now and in the future metagenomics will be used in the same manner as 16S

rRNA gene fingerprinting methods were used to describe microbial community profiles. It will therefore become a “standard” tool for many laboratories and scientists working in the field of microbial ecology.

This review aims to give an overview for the field of metagenomics, with particular emphasize in all steps involved of a “typical” sequence-based, metagenome project. We will describe and discuss sample processing, sequencing technology, assembly and annotation, experimental design and statistical analysis and finally data storage and sharing. Clearly, any kind of metagenomic dataset will benefit from the rich information available from other metagenome projects, and it is the hope that common, yet flexible, standards and vivid interactions between scientists in the field will facilitate this. This review article aims to summarize the current thinking in the field and will introduce scientists new to the field to current practices and key issues one needs to consider for a successful metagenome project.

### **Sampling and processing**

Sample processing is the first and most crucial step in any metagenomics project. It needs to ensure that the DNA extracted is representative of all cells present in the sample and that sufficient amounts of high-quality nucleic acids for subsequent library production and sequencing are obtained. This will obviously require specific protocols for each sample types and variety of robust methods for DNA extraction are available (e.g.[3, 9, 10]). Initiatives are also underway to explore the microbial biodiversity from tens of thousands of ecosystems using a single DNA extraction technology to ensure comparability [11].

If the target community is associated with a host (e.g. an invertebrate or plant), then fractionation and/ or selective lysis might be suitable to ensure that minimal host DNA is obtained (e.g.[9, 12]). This is particularly important when the host genome is large and hence might “overwhelm” the sequences of the microbial community in the subsequent sequencing effort. Physical fractionation is also applicable, when only a certain part of the community is the target of analysis, for example, viruses, bacteria or picoeukaryotes in seawater samples. Here a range of selective filtration or centrifugation steps, or even flow cytometry, can be used to enrich the target fraction [3, 13, 14]. Importantly, fractionation steps should be checked, to ensure that sufficient enrichment of the target is achieved and minimal “contamination” of non-target material occurs.

Physical separation and isolation of cells from the samples might also be important to maximize DNA yield or avoid co-extraction of enzymatic inhibitors (like humic acids) that might interfere with subsequent processing. This situation is particularly relevant for soil metagenome projects and substantial work has been done in this field to address the issue ([10] and reference therein). Direct lysis of cells in the soil matrix versus indirect lysis (i.e. after separation of cells from the soil) has a quantifiable bias in terms of microbial diversity, DNA yield and fragment length [10]. The extensive work on soil highlight the needs to ensure that extraction procedures are well benchmarked and multiple methods should be compared to ensure representative extraction of DNA.

Certain type of samples (such as biopsies or ground-water) often only yield very small amounts of DNA [15]. Library production for most sequencing technologies require high nanograms or micrograms amounts of DNA (see below) and hence amplification of starting material might be required. Multiple displacement amplification (MDA)

with the random hexamers and the phage phi29 polymerase is one option to increase DNA yields. This method can amplify femtograms of DNA to produce micrograms of product and thus has been widely used in single cell genomics and to a certain extent in metagenomics [16, 17]. As with any amplification method, there is sequence bias in the amplification and its impact will depend on the amount and type of starting material and the required number of amplification rounds to produce sufficient amounts of nucleic acids. This amplification bias can have significant impact on subsequent metagenomic community analysis [15] and this might make amplification permissible when the data from amplified and unamplified samples are compared.

### **Sequencing technology**

Over the last 10 years metagenomic shotgun sequencing has gradually shifted from classical Sanger sequencing technology to next generation sequencing (NGS). Sanger sequencing, however, is still considered the gold standard for sequencing due to its low error rate, long read length (>700bp) and large insert sizes (e.g. >30Kb for fosmids or bacterial artificial chromosomes (BACs)). All these aspects will improve assembly outcomes for shotgun data and hence Sanger sequencing might still be applicable, if generating close-to-complete genomes in low diversity environment are the work's objective [18]. A drawback of Sanger sequencing is the labor-intensive cloning process in its associated bias against genes toxic for the cloning host [19] and the overall cost per Gbp (appr. USD 400,000).

Of the NGS technologies both the 454/ Roche and the Illumina/ Solexa systems have now been extensively applied to metagenomic samples. Excellent reviews of these

technologies are available [20, 21], but a brief summary is given here with particular attention to metagenomic applications.

The 454/ Roche system applies emulsion polymerase chain reaction (ePCR) to clonally amplify random DNA fragments, which are attached to microscopic beads. Beads are deposited into the wells of a picotitre plate and then individually and in parallel pyrosequenced. The pyrosequencing process involves the sequential addition of all four deoxynucleoside triphosphates, which, if complementary to the template strand, are incorporated by a DNA polymerase. This polymerization reaction releases pyrophosphate, which is converted via two enzymatic reactions to the production of light. Light production of  $\sim 1.2$  million reactions is detected in parallel via a charge-coupled device (CCD) camera and converted to the actual sequence of the template. Two aspects are important in this process with respect to metagenomic applications. Firstly, the ePCR has been shown to produce artificial replicate sequences, which will impact any estimates of gene abundance. Understanding the amount of replicate sequences is a crucial part of understanding the data quality of sequencing runs and those replicates can be identified and filter out with bioinformatics tools [22, 23]. Secondly, the intensity of light produced when the polymerase runs through a homopolymer is often difficult to correlate to the actual number of nucleotide positions. Typically, this results in insertion or deletion errors of homopolymer greater than 6 position and can hence can cause frame-shifts, if open reading frames (ORF) are called on a single read. This type of error is, however, somewhat predictable and can be incorporated into models of ORF prediction [24]. Despite these disadvantages, the much cheaper cost of  $\sim$  USD 20,000 per Gbp has made 454/ Roche pyrosequencing a popular choice for shotgun-sequencing metagenomics. In addition,

the 454/ Roche technology produces average read length between 300-400 bp, which is long enough to cause only minor loss in the amount of reads that can be annotated [25]. Samples preparation has also been optimized so that tens of nanograms of DNA are sufficient for sequencing single-end libraries [26, 27], although pair-end sequencing might still require micrograms quantities. Finally, the 454/Roche sequencing platform offers multiplexing allowing for several samples to be analyzed in a single run of ~500 Mbp.

The Illumina/ Solexa technology immobilizes random DNA fragments on a surface and then performs solid-surface PCR amplification resulting in cluster of identical DNA fragments. Those are then sequenced with reversible terminators in a sequencing-by-synthesis process [28]. The cluster density is enormous with hundreds of millions of reads per surface channel and 16 channels per run on the HiSeq2000 instrument. Read length is now approaching 150 bp and clustered fragments can be sequenced from both ends. Using overlapping read pairs on a single template, merged reads can reach 200bps (from 100bp overlapping) or 300bp from (150bp overlapping). Yields of ~60 Gbp can therefore be typically expected in a single channel. Illumina/ Solexa has limited systematic errors, however, some data sets have shown high error rates at the tail ends of reads [29]. In general, clipping reads has proven a good strategy to eliminate the error in “bad” datasets, however data sets quality should also be used to detect “bad” sequences. The lower costs of this technology (~ USD 50 per Gbp) and recent success in its application to metagenomics, and even the generation of draft genome from complex dataset [30, 31], is currently making the Illumina technology an increasingly popular choice. As with 454/ Roche sequencing, starting material can be as low as a few nanograms, but larger amounts are required when

jump-libraries for longer insert libraries are made. The limited read length of the Illumina/ Solexa technology means that a larger proportion of unassembled reads than with 454/ Roche technology might be too short for functional annotation [25]. While some researchers assume that assembly of data is therefore advisable, we very much suggest to not use assembly as this has the potential to introduce biases by e.g. suppressing low abundance species since they can not be assembled and the fact that current software (e.g. MG-RAST) is capable of analyzing unassembled Illumina reads of 100bp and longer. Multiplexing of samples is also available for individual sequencing channel, with more than 500 samples multiplexed per lane. Another important factor to consider is run time, with a 2x100bp paired-end sequencing analysis taking approx. 10 days instrument time, in contrast to one day for the 454/ Roche technology. However, faster runtime (albeit at higher cost per Gbp of approx. USD 1000) can be achieved with the new Illumina MiSeq instrument. This smaller version of Illumina/ Solexa technology can also be used to “test-run” sequencing libraries, before analysis on HiSeq instrument for deeper sequencing.

There are a few additional sequencing technologies available that might prove themselves useful for metagenomic application, now or in the near future. Applied Biosystems SOLiD sequencer is one of them. The system has been extensively used, for example, in genome re-sequencing [32]. SOLiD arguably provides the lowest error rate of any current NGS sequencing technology, however does not achieve reliable read length much beyond 50 nucleotides. This will limit its applicability for direct gene annotation of unassembled reads or assembly to large contigs. However for the purpose of assembly or mapping of metagenomic data against a reference genome recent work showed encouraging outcome [33]. Roche is also marketing a

smaller-scale sequencer based on pyrosequencing with about 100 Mbp output and low per-run costs. This system might be useful, as relatively low coverage of metagenomes can establish meaningful gene profile [34]. Ion Torrent is another emerging technology and is based on the principle that nucleotide incorporation can be detected by protons release during DNA polymerization. These system promises read length of >100 bp and throughput in the order of magnitude of the 454/ Roche sequencing systems. Pacific Biosciences (PacBio) has released a sequencing technology based on single-molecule, real time detection in zero-mode waveguide wells. Theoretically, this technology on its RS1 platform should provide much greater read length than the other technologies mentioned, which would facilitate annotation and assembly. In addition a process called “strobing” will mimic pair-end reads. However accuracy of single reads for PacBio is current only at 85% and random reads are “dropped” making the instrument unusable in its current form for metagenomic sequencing [35]. Complete Genomics is offering a technology based on sequencing DNA nanoballs with combinatorial probe-anchor ligation [36]. Its read length of 35 nucleotides is rather limited and so might be its utility for *de novo* assemblies. While none of the emerging sequencing technologies have been thoroughly applied and tested to metagenomics samples, they offer however promising alternatives and even further reduction of costs.

### **Assembly**

If the research aims at recovering the genome of uncultured organisms rather than a functional description of the community, then assembly of short read fragments will be performed to obtain longer genomic contigs. All current assembly programs we

design to assemble clonal “genomes” and their utility for complex pan-genomic mixtures should be approached with caution and critical evaluation.

Two strategies can be employed for metagenomics samples, reference-based assembly (co-assembly) and *de novo* assembly. Reference-based assembly can be done using the software packages like Newbler (Roche), AMOS (<http://sourceforge.net/projects/amos/>) or Mira [37]. These software packages include algorithms that are fast and memory-efficient and hence can often be performed on laptop-sized machines in a couple of hours. Reference-based assembly works well if the metagenomic dataset contains sequences where closely related reference genomes are available. However slight difference in the true genome of the sample to the reference, such as large insertion, deletion or polymorphisms, can introduce substantial biases in the analysis.

*De novo* assembly typically requires larger computational resources and thus a whole class of assembly tools based on the deBruijn graphs was specifically created to handle the very large amounts of data [38, 39]. Machine requirements for the deBruijn assembler Velvet [40] or SOAP [41] are still significantly higher than for co-assembly, often requiring hundreds of gigabytes of memory in a single machine and runtimes frequently being days.

The fact that most (if not all) microbial communities include significant variation on a strain and species level makes the use of any existing assembly algorithm risky. The “clonal” assumptions built into most assemblers might lead to suppression of contig formation for certain heterogeneous taxa at specific parameter settings. This problem can be however be identified during the analysis. For example, varying k-mer length

in Velvet assembler might lead to completely different contigs being formed and this would warning sign that the assembly process impacts on the representation of diversity in the assembled contigs.

There are several points that need be considered when exploring the reasons for assembling metagenomic data. However, these can be condensed down to the following. Firstly, what is the length of the sequencing reads used to generate your metagenomic dataset and are longer sequences required for annotation? Pipelines like MG-RAST [42] only require 75bp or longer for gene prediction or similarity analysis that provides taxonomic binning and functional classification. However, on the whole, the longer the sequence information, the better is the ability to obtain accurate information from it. One obvious impact is on annotation, i.e. the longer the sequence, the more information and hence the easier it compare it to known genetic data e.g. via homology searches [25]. Annotation issues will be discussed in the next section.

Binning and classification of DNA fragments for phylogenetic or taxonomic assignment also benefit from long contiguous sequences and certain tools (e.g. Phylopythia) work only reliably over specific cut-off (e.g. 1Kb)[43]. Secondly, is the dataset assembled to reduce data processing requirements? Here as an alternative to assembling reads into contigs, clustering near-identical reads with cd-hit [44] or uclust [45] will provide clear benefits in data reduction. The MG-RAST pipeline uses also clustering as a data reduction strategy.

Fundamentally, assembly is also driven by the specific problem that single reads have generally lower quality and hence lower confidence in the accuracy than multiple reads which cover the same segment of genetic information. Therefore, merging reads

increases the quality of information. Obviously in a complex community with low sequencing depth/coverage it is unlikely to actually get many reads which cover the same fragment of DNA. Hence assembly may be of limited value for metagenomics.

Unfortunately without assembly, longer and more complex genetic elements (e.g. CRISPRS) cannot be analyzed. Hence there is a need for metagenomic assembly to obtain high-confidence contigs that enable the study of, for example, major repeat classes. However none of the current assembly tools is bias-free. Several strategies have been proposed to increase assembly accuracy [38], however strategies such as removal of rare k-mers are now no longer considered adequate, as rare k-mers do not represent sequence errors (as initially assumed), but reads from less abundant pan-genomes in the metagenomic mix.

### **Annotation**

There are two distinct paths that the annotation of metagenomics can follow. Firstly, reconstructed genomes are the objective of the study and assembly has been successful in obtaining large contigs. In this case, it is preferable to use existing pipelines for genome annotation, such as RAST [46] or IMG [47]. For this approach to be successful, minimal contigs length of 30,000bp or longer are required. Secondly, annotation can be performed on the entire community and relies on unassembled reads or short contigs. Here the tools for genome annotation are significantly less useful than those specifically developed for metagenomic analyses.

Annotation is the single, biggest computational challenge for most metagenomic projects and therefore deserves much attention now and over the next years. Current

estimates are that only between 20-50% of a metagenomic sequences can be annotated [48], leaving the immediate question of importance and function of the remaining genes. It should be noted that annotation is not done *de novo*, but via mapping to gene or protein libraries with existing knowledge (i.e. a non-redundant database). Any sequences that cannot be mapped to the known sequence space are referred to as orphans. These “orphan” are responsible for the seemingly never-ending, genetic novelty in microbial metagenomics (e.g.[49]). Two hypotheses exist for the roles of this unknown fraction. Firstly, the vast number of orphan genes encode for unknown biochemical functions. Secondly, orphan genes have no sequence homology with known genes, but potentially structural homology with known protein thus representing known protein families or folds. Future work will probably reveals that the truth lies somewhere between these two hypotheses [50]. Improving annotation of orphan genes we will rely on the challenging and labor-intensive task of protein structure analysis (e.g. via NMR and X-ray crystallography) and biochemical characterization.

Currently, metagenomic annotation relies on classifying sequences to known functions or taxonomic units based on homology searches against available “annotated” data. The process of annotation is based on predicting features associated with genes, which either based on intrinsic (i.e. not similarity-based and based on algorithm described, for example, in [24, 51-53]) or extrinsic (i.e. based on similarity searches with, for example, BLAST[54]) characters. Conceptually, the annotation is relatively simple and for small datasets (<10 000 sequences) manual curation can be used increase the accuracy of any automated annotation. Metagenomic dataset are typically very large, so that the latter is not possible. Automated annotation therefore

has to become more accurate and computationally inexpensive. Running a BLASTX similarity search is computationally expensive and thus require finances as much as ten time the cost of sequencing [55]. Computationally less demanding methods involving detecting feature composition in genes [43] have however limited success for short reads. With growing dataset sizes faster algorithms are urgently needed and several programs for similarity searches have been developed to resolve this issue [45, 56-58].

Several large-scale databases are available that process and deposit metagenomic dataset and IMG/M and MG-RAST are two prominent systems [42, 59]. MG-RAST alone contains more than 7000 users and >28,000 uploaded and analyzed metagenomes (of which 4300 are publicly accessible). This statistics demonstrate a move by the scientific community to centralize resources and standardize annotation. Both IMG/M and MG-RAST provide the ability to use stored computational results for comparison, enabling comparison of novel metagenomes to a rich body of other datasets without requiring the end-user to provide the computational means for re-analysis of all data sets involved in their study.

Other systems, such as CAMERA[60], offer more flexible annotation schema, but require that individual researchers understand the annotation of data and analytical pipeline well enough to be confident on their interpretation. MEGAN is another tool used for visualizing annotation results derived form BLAST searches in an functional or taxonomic dendrogram [61]. The use of dendrograms to display metagenomic data provides a collapsible network of interpretation, which makes analysis of particular functional or taxonomic groups visually quite easy.

Many reference databases are available to give functional context to metagenomic datasets, such as KEGG [62], COG/KOG [63], PFAM [64], TIGRFAM [65] to name a few. However, as no reference database covers all biological functions, the ability to visualize and merge the interpretations of all database searches within a single framework as implemented in the most recent version of MG-RAST and IMG-M is important. Thus it is essential that metagenome analysis platforms are able to share data in way that map and visualize them in the framework of other platforms. These metagenomic exchange languages should also reduce the burden associated with processing large datasets, as redundancy of search will be minimized and annotations are shared and mapped to different ontologies and nomenclatures to allow multifaceted interpretations. The Genomic Standards Consortium (GSC) with the M5 project is providing a prototypical standard for exchange of computed metagenome analysis results, one cornerstone of these exchange languages.

### **Experimental design and statistical analysis**

Owing to the high costs, many of the early metagenomic shotgun sequencing projects were not replicated or were focused on targeted exploration of microbial diversity. Reduction of sequencing cost (see above) and a much wider appreciation of the utility of metagenomics to address fundamental questions in microbial ecology, now require proper experimental designs with appropriate replication and statistical analysis. This design and statistical aspects, while obvious, are often not properly implemented in the field of microbial ecology [66]. However, many suitable approaches and strategies are readily available from the many decades of research in quantitative ecology of higher organisms (e.g. animals, plants). In a simplistic way, the data from multiple

metagenomic shotgun-sequencing projects can be reduced to tables, where the columns represent samples and the rows indicate either a taxonomic group or a gene function (or groups thereof), and the fields containing abundance or presence/ absence data. This is analogous to species-sample matrices in ecology of higher organisms and hence many of the statistical tools available to identify correlations and statistically significant patterns are readily transferable. The Primer-E package [67] is such a well-established tool allowing for a range of multivariate statistical analysis, including the generation of multi-dimensional scaling (MDS) plots, analysis of similarities (ANOSIM) and identification of the species or functions that contribute to the difference between two samples (SIMPER). Recently, multivariate statistics was also incorporated in a web-based tool, called Metastats [68], which revealed with high confidence discriminatory function between the replicated metagenome dataset of the gut microbiota of lean and obese mice [69]. In addition, the ShotgunFunctionalizeR package provides several statistical procedures for assessing functional differences between samples, both for individual genes and for entire pathways using the popular R statistical package [70].

Ideally and in general, experimental design should be driven by the question asked (rather than technical or operational restriction). For example, if a project aims to identify unique taxa or functions in a particular habitat, then suitable reference samples for comparison should be taken and ideally processed in consistent manner. In addition, variation between sample types can be due to true biological variation (something biologist would be most interested in) and technical variation and this should be carefully considered when planning the experiment. One should also be aware that many microbial systems are highly dynamic, so temporal aspects of

sampling can have a substantial impact on data analysis and interpretation. While the question of the number of replicates is often difficult to predict prior to the final statistical analysis and often boils down to rule of thumb “the more, the better” [66]. Also the level at which replication takes place is something that should not lead to false interpretation of the data. For example, if one is interested in the level of functional variation of the microbial community in habitat A, then multiple samples from this habitat should be taken and processed completely separately, but in the same manner. Taking just one sample and splitting it up prior to processing, will only provide information about technical, but not biological variation in habitat A. Taking multiple samples and then pooling them, will lose all information on variability and hence will be of little use for statistical purposes. Ultimately, good experimental design of metagenomic projects will facilitate integration of data set into new or existing ecological theories [71].

As metagenomic is gradually moving through a range of explorative biodiversity surveys it will also prove itself extremely valuable for manipulative experiments. This will allow for observation of treatment impact on the functional and phylogenetic composition of microbial communities. Initial experiments already showed promising results [72], however careful experimental planning and interpretations should be paramount in this field.

One of the ultimate aims of metagenomics is to link functional and phylogenetic information to the chemical, physical and other biological parameters that characterize an environment. While measuring all these parameters can be time- and cost-intensive, it allows retrospective correlation analysis of metagenomic data that

were perhaps not part of the initial aims of the project or might be of interest for other research questions. The value of such metadata cannot be underestimated for future research and, in fact, has become mandatory or optional for deposition of metagenomic data into some databases [59, 60].

### **Sharing and Storage of Data**

Data sharing has a long tradition in the field of genome research, but for metagenomic data this will require a whole new level of organization and collaboration to provide metadata and centralized services (e.g. IMG/M, CAMERA and MG-RAST) as well as sharing of both data and computational results. To enable sharing of computed results, some aspects of the various analytical pipelines mentioned above will need to be coordinated -a process, currently under way under the auspices of the GSC. Once this has been achieved, researchers will be able to download intermediate and processed results from any one of the major repositories for local analysis and/or comparison. The question of centralized versus de-centralized storage is also one of “who pays for the storage”, a matter with no simple answer. The US National Center for Biotechnology Information (NCBI) is mandated to store all metagenomic data, however, the sheer volume of data being generated means there is an urgent need for appropriate ways of storing vast amounts of sequences. As cost of sequencing continues to drop, while the cost for analysis and storing remains more or less constant, selection of data storage in either biological (i.e. the sample that was sequenced) or digital form in (de-) centralized archives might be required. Ongoing work and successes in compression of (meta-)genomic data [73] however might mean that digital information can still be stored cost-efficiently in the near future.

## Conclusion

Metagenomics has benefited in the last few years from many visionary investments both in financial and intellectual terms. To ensure that those investments are utilized in the best possible way, the scientific community should aim to share, compare and critically evaluate the outcomes of metagenomic studies. As data sets will become increasingly more complex and comprehensive, novel tools for analysis, storage and visualization will be required. This will ensure the best use of the metagenomics as a tool to address fundamental question of microbial ecology, evolution and diversity and to derive and test new hypotheses. Metagenomics will also be a tool as commonly and frequently employed as any other laboratory method and “metagenomising” a sample might become as colloquial as “PCRing”. It is therefore also important that metagenomics will be taught to students and young scientists in the same way as other techniques and approaches have been in the past.

## Acknowledgements

This work was supported by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357

## Reference:

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.** *Chem Biol* 1998, **5**(10):R245-249.
2. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428**(6978):37-43.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**(5667):66-74.

4. Simon C, Daniel R: **Metagenomic analyses: past and future trends.** *Appl Environ Microbiol* 2011, **77**(4):1153-1161.
5. Wilmes P, Bond PL: **Metaproteomics: studying functional gene expression in microbial ecosystems.** *Trends Microbiol* 2006, **14**(2):92-97.
6. Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I: **Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities.** *PLoS One* 2008, **3**(8):e3042.
7. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL *et al*: **Bacterial rhodopsin: evidence for a new type of phototrophy in the sea.** *Science* 2000, **289**(5486):1902-1906.
8. Nicol GW, Schleper C: **Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle?** *Trends Microbiol* 2006, **14**(5):207-212.
9. Burke C, Kjelleberg S, Thomas T: **Selective extraction of bacterial DNA from the surfaces of macroalgae.** *Appl Environ Microbiol* 2009, **75**(1):252-256.
10. Delmont TO, Robe P, Clark I, Simonet P, Vogel TM: **Metagenomic comparison of direct and indirect soil DNA extraction approaches.** *J Microbiol Methods* 2011, **86**(3):397-400.
11. Knight R, Desai N, Field D, Fierer N, Fuhrman J, Gordon J, Hu B, Hugenholtz P, Jansson J, Meyer F *et al*: **Designing Better Metagenomic Surveys: The role of experimental design and metadata capture in making useful metagenomic datasets for ecology and biotechnology.** *Nature Biotechnology* in review.
12. Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, Heidelberg KB, Egan S, Steinberg PD, Kjelleberg S: **Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis.** *ISME J* 2010, **4**(12):1557-1567.
13. Palenik B, Ren Q, Tai V, Paulsen IT: **Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity.** *Environ Microbiol* 2009, **11**(2):349-359.
14. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H *et al*: **The marine viromes of four oceanic regions.** *PLoS Biol* 2006, **4**(11):e368.
15. Abbai NS, Govender A, Shaik R, Pillay B: **Pyrosequence Analysis of Unamplified and Whole Genome Amplified DNA from Hydrocarbon-Contaminated Groundwater.** *Mol Biotechnol* 2011.
16. Lasken RS: **Genomic DNA amplification by the multiple displacement amplification (MDA) method.** *Biochem Soc Trans* 2009, **37**(Pt 2):450-453.
17. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS: **Genomic sequencing of single microbial cells from environmental samples.** *Curr Opin Microbiol* 2008, **11**(3):198-204.
18. Goltsman DS, Denev VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A *et al*: **Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum**

- ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms.** *Appl Environ Microbiol* 2009, **75**(13):4599-4615.
19. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM: **Genome-wide experimental determination of barriers to horizontal gene transfer.** *Science* 2007, **318**(5855):1449-1452.
  20. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11**(1):31-46.
  21. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24**(3):133-141.
  22. Niu B, Fu L, Sun S, Li W: **Artificial and natural duplicates in pyrosequencing reads of metagenomic data.** *BMC Bioinformatics* 2010, **11**:187.
  23. Teal TK, Schmidt TM: **Identifying and removing artificial replicates from 454 pyrosequencing data.** *Cold Spring Harb Protoc* 2010, **2010**(4):pdb prot5409.
  24. Rho M, Tang H, Ye Y: **FragGeneScan: predicting genes in short and error-prone reads.** *Nucleic Acids Res* 2010, **38**(20):e191.
  25. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: read length matters.** *Appl Environ Microbiol* 2008, **74**(5):1453-1463.
  26. White RA, 3rd, Blainey PC, Fan HC, Quake SR: **Digital PCR provides sensitive and absolute calibration for high throughput sequencing.** *BMC Genomics* 2009, **10**:116.
  27. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X *et al*: **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.** *Genome Biol* 2010, **11**(12):R119.
  28. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**(7218):53-59.
  29. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H *et al*: **Sequence-specific error profile of Illumina sequencers.** *Nucleic Acids Res* 2011, **39**(13):e90.
  30. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T *et al*: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331**(6016):463-467.
  31. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**(7285):59-65.
  32. Gulig PA, de Crecy-Lagard V, Wright AC, Walts B, Telonis-Scott M, McIntyre LM: **SOLiD sequencing of four *Vibrio vulnificus* genomes enables comparative genomic analysis and identification of candidate clade-specific virulence genes.** *BMC Genomics* 2010, **11**:512.
  33. Tyler HL, Roesch LF, Gowda S, Dawson WO, Triplett EW: **Confirmation of the sequence of 'Candidatus Liberibacter asiaticus' and assessment of microbial diversity in Huanglongbing-infected citrus phloem**

- using a metagenomic approach.** *Mol Plant Microbe Interact* 2009, **22**(12):1624-1634.
34. Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, von Mering C, Bebout BM, Pace NR, Bork P *et al*: **Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat.** *Mol Syst Biol* 2008, **4**:198.
35. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D *et al*: **Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *N Engl J Med* 2011, **365**(8):709-717.
36. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G *et al*: **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.** *Science* 2010, **327**(5961):78-81.
37. Chevreux B, Wetter T, Suhai S: **Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology.** *Proceedings of the German Conference on Bioinformatics* 1999, **99**:45-56.
38. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**(6):315-327.
39. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci U S A* 2001, **98**(17):9748-9753.
40. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821-829.
41. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24**(5):713-714.
42. 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, **2010**(1):pdb prot5368.
43. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4**(1):63-72.
44. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(13):1658-1659.
45. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26**(19):2460-2461.
46. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M *et al*: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.
47. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC: **IMG ER: a system for microbial genome annotation expert review and curation.** *Bioinformatics* 2009, **25**(17):2271-2278.
48. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I *et al*: **The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation.** *PLoS One* 2010, **5**(11):e15545.

49. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W *et al*: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.** *PLoS Biol* 2007, **5**(3):e16.
50. Godzik A: **Metagenomics and the protein universe.** *Curr Opin Struct Biol* 2011, **21**(3):398-403.
51. Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.** *Nucleic Acids Res* 2006, **34**(19):5623-5630.
52. Yok NG, Rosen GL: **Combining gene prediction methods to improve metagenomic gene annotation.** *BMC Bioinformatics* 2011, **12**:20.
53. Zhu W, Lomsadze A, Borodovsky M: **Ab initio gene identification in metagenomic sequences.** *Nucleic Acids Res* 2010, **38**(12):e132.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
55. Wilkening J, Desai N, Meyer F, A. W: **Using clouds for metagenomics — case study.** *IEEE Cluster* 2009.
56. Ye Y, Choi JH, Tang H: **RAPSearch: a fast protein similarity search tool for short reads.** *BMC Bioinformatics* 2011, **12**:159.
57. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
58. Wang W, Zhang P, Liu X: **Short read DNA fragment anchoring algorithm.** *BMC Bioinformatics* 2009, **10 Suppl 1**:S17.
59. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I *et al*: **IMG/M: a data management and analysis system for metagenomes.** *Nucleic Acids Res* 2008, **36**(Database issue):D534-538.
60. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J *et al*: **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.** *Nucleic Acids Res* 2011, **39**(Database issue):D546-551.
61. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17**(3):377-386.
62. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32**(Database issue):D277-280.
63. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
64. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.
65. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O: **TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes.** *Nucleic Acids Res* 2007, **35**(Database issue):D260-264.
66. Prosser JI: **Replicate or lie.** *Environ Microbiol* 2010, **12**(7):1806-1810.

67. Clarke KR: **Non-parametric multivariate analyses of changes in community structure.** *Australian Journal of Ecology* 1993(18):117-143.
68. White JR, Nagarajan N, Pop M: **Statistical methods for detecting differentially abundant features in clinical metagenomic samples.** *PLoS Comput Biol* 2009, **5**(4):e1000352.
69. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP *et al*: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457**(7228):480-484.
70. Kristiansson E, Hugenholtz P, Dalevi D: **ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes.** *Bioinformatics* 2009, **25**(20):2737-2738.
71. Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T: **Bacterial community assembly based on functional genes rather than species.** *Proc Natl Acad Sci U S A* 2011, **108**(34):14288-14293.
72. Mou X, Sun S, Edwards RA, Hodson RE, Moran MA: **Bacterial carbon processing by generalist species in the coastal ocean.** *Nature* 2008, **451**(7179):708-711.
73. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression.** *Genome Res* 2011, **21**(5):734-740.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.