2015

Steps in Metagenomics: Let’s Avoid Garbage in and Garbage Out

Jacques Izard

Steps in Metagenomics: Let's Avoid Garbage in and Garbage Out

Jacques Izard

University of Nebraska–Lincoln

jizard@unl.edu

Why Metagenomics?

Is metagenomics a revolution or a new fad? Metagenomics is tightly associated with the availability of next-generation sequencing in all its implementations. The key feature of these new technologies, moving beyond the Sanger-based DNA sequencing approach, is the depth of nucleotide sequencing per sample.[1] Knowing much more about a sample changes the traditional paradigms of "What is the most abundant?" or "What is the most significant?" to "What is present and potentially significant that might influence the situation and outcome?" Let's take the case of identifying proper biomarkers of disease state in the context of chronic disease prevention. Prevention has been deemed as a viable option to avert human chronic diseases and to curb healthcare management costs.[2] The actual implementation of any effective preventive measures has proven to be rather difficult. In addition to the typically poor compliance of the general public, the vagueness of the successful validation of habit modification on the long-term risk, points to the need of

Published (as Chapter 1) in Izard, J., & Rivera, M.C., eds., *Metagenomics for Microbiology*
(Academic Press, 2015), pp 1–23. doi 10.1016/B978-0-12-410472-3.00001-4
Online @ https://doi.org/10.1016/B978-0-12-410472-3.00001-4

Published 14 November 2014.

defining new biomarkers of disease state. Scientists and the public are accepting the fact that humans are super-organisms, harboring both a human genome and a microbial genome, the latter being much bigger in size and diversity, and key for the health of individuals.[3,4] It is time to investigate the intricate relationship between humans and their associated microbiota and how this relationship modulates or affects both partners.[5] These remarks can be expanded to the animal and plant kingdoms, and holistically to the Earth's biome. By its nature, the evolution and function of all the Earth's biomes are influenced by a myriad of interactions between and among microbes (planktonic, in biofilms or host associated) and the surrounding physical environment.

The general definition of metagenomics is the cultivation-independent analysis of the genetic information of the collective genomes of the microbes within a given environment based on its sampling. It focuses on the collection of genetic information through sequencing that can target DNA, RNA, or both. The subsequent analyses can be solely focused on sequence conservation, phylogenetic, phylogenomic, function, or genetic diversity representation including yet-to-be annotated genes. The diversity of hypotheses, questions, and goals to be accomplished is endless. The primary design is based on the nature of the material to be analyzed and its primary function (Figure 1).

It All Starts with the Study Design

The goal is not to tell you how to do your science but to emphasize some aspects of study design that need careful attention because of the characteristics of the methodologies used in metagenomic studies. It begins by identifying the primary objective of the metagenomics project. What is the main scientific question you are trying to answer? More than one hypothesis can be tested depending on the scope of the experiment and the amount of associated data, or metadata, that you collect and use for your subsequent analyses.

The high-dimensionality characteristic of the metagenomics datasets is challenging and is revolutionizing microbiology analytical methodology. What is meant by high-dimensional dataset? Let's take as an example the Human Microbiome Project (HMP) 16S ribosomal RNA (rRNA)-based characterization of 10 sites from the digestive tract of 200 individuals. Such analysis required the collection of over 2000 samples, generating approximately 23 million high-quality sequence reads that were assigned to 674 taxonomic clades with their respective relative abundance per taxonomic level (e.g., from phylum to genus). For example, for the genus *Pyramidobacter*, the database stores the relative abundance at each taxonomic level, from the phylum (e.g., “Bacteria | Synergistetes”), the most inclusive taxonomic level, to the genus (e.g., “Bacteria | Synergistetes | Synergistia | Synergistales | Synergistaceae | *Pyramidobacter*”), the least inclusive taxonomic level, and all the taxonomic levels between the two.[6] From the same study, four body sites were further analyzed using whole metagenome shotgun (WMS) sequencing from approximately 100 individuals, generating a trillion nucleotides.[6] Another example can be extracted from the work of Giannoukos et al.[7] while developing rRNA depletion methodology for fecal samples. They obtained over 100,000 reads per sample.[7] In each example, each sample has a tremendous amount of genotypic and phenotypic information in addition to the metadata (e.g., age, sex, race, and others). In addition to the nucleotide data, information about other molecules (e.g., lipids, proteins, and metabolites) can be collected; increasing the complexity and multidimensionality of the dataset. The type of data collected will determine the type of analyses performed. These analyses can help answer questions such as: “What are the organisms present?”, “What can these organisms potentially do?”,’ “What is their metabolic capability?”, and “How do they influence the host?” (**Figure 1**). Planning the structure of samples and metadata acquisition as well as the analysis pipeline to be used, prior to the start of the experiment, will avoid bottlenecks and optimize utilization of funds.

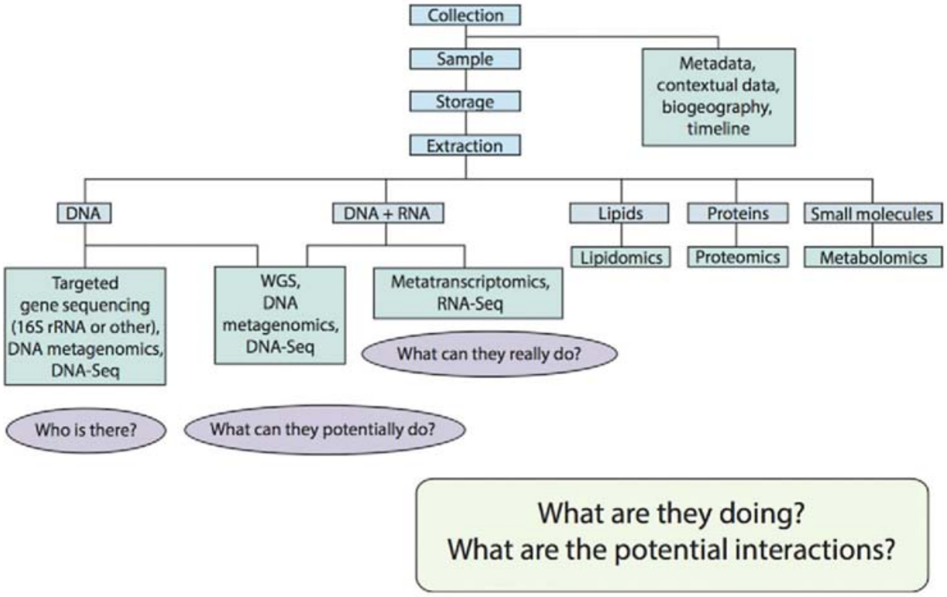
Collection
Sample
Storage
Extraction
Metadata, contextual data, biogeography, timeline
DNA
DNA + RNA
Lipids
Proteins
Small molecules
Lipidomics
Proteomics
Metabolomics
Targeted gene sequencing (16S rRNA or other), DNA metagenomics, DNA-Seq
WGS, DNA metagenomics, DNA-Seq
Metatranscriptomics, RNA-Seq
What can they really do?
Who is there?
What can they potentially do?
What are they doing?
What are the potential interactions?

Fig. 1. Metagenomic analysis process and some of the overarching questions that can be answered by the different methodologies.

During the study design phase, investigators need to take into consideration the ethical and legal issues related to metagenomics data collection and analysis. Some of the constrains of metagenomics studies utilizing human subjects include Institutional Review Boards, informed consent, and other issues related to the protection of the identifiable health information of the human subjects (e.g., HIPAA Privacy Rule in the United States). For examples of consent documentation and standard operating procedures, the National Institutes of Health HMP has made those document public and available online (http://www.hmpdacc.org).[8] It is essential for the consent procedures to accurately state what data will be gathered, how it will be used, and how it will be stored. All efforts should be made to secure information and confidentiality of the genetic material and associated data over time. This includes both the physical storage of the information, data deposition and data sharing, even when the samples are de-identified. For environmental samples, having the right of access and sampling permits is critical as geolocation is now required with the sample data submission to repository. It is important to point out that any samples collected from a host will contain a significant amount of the host genetic material. The potential contamination of samples with the host genetic material adds to the complexity of the metagenomics studies, and sophisticated computational pipelines for the removal of the contaminating reads are essential to generate meaningful conclusions and, in the case of human subjects, to protect the privacy and confidentiality of the sample donor. **Figure 2** shows the impact of human "contamination" on the amount and quality of the data collected using shotgun sequencing of human samples from 16 different body sites.[8] When working with different models, it should be noted that the genome of a brown rat is not that much smaller than that of a human (over 3 billion base pairs), and that the corn genome is over 2 billion base pairs. Although protists and fungi are much smaller, their genomes are still composed of few million base pairs. The knowledge of your biological system of interest will be critical to optimize the study design.

Have a Statistical Analysis Plan in Place Before Starting

Planning for statistical analysis should be an integral part of the study design. Although many experimental designs can be performed in metagenomics project, there is no single path to a successful strategy. While using metagenomic or metatranscriptomic approaches, it is essential to refer to the specific needs of each experiment.

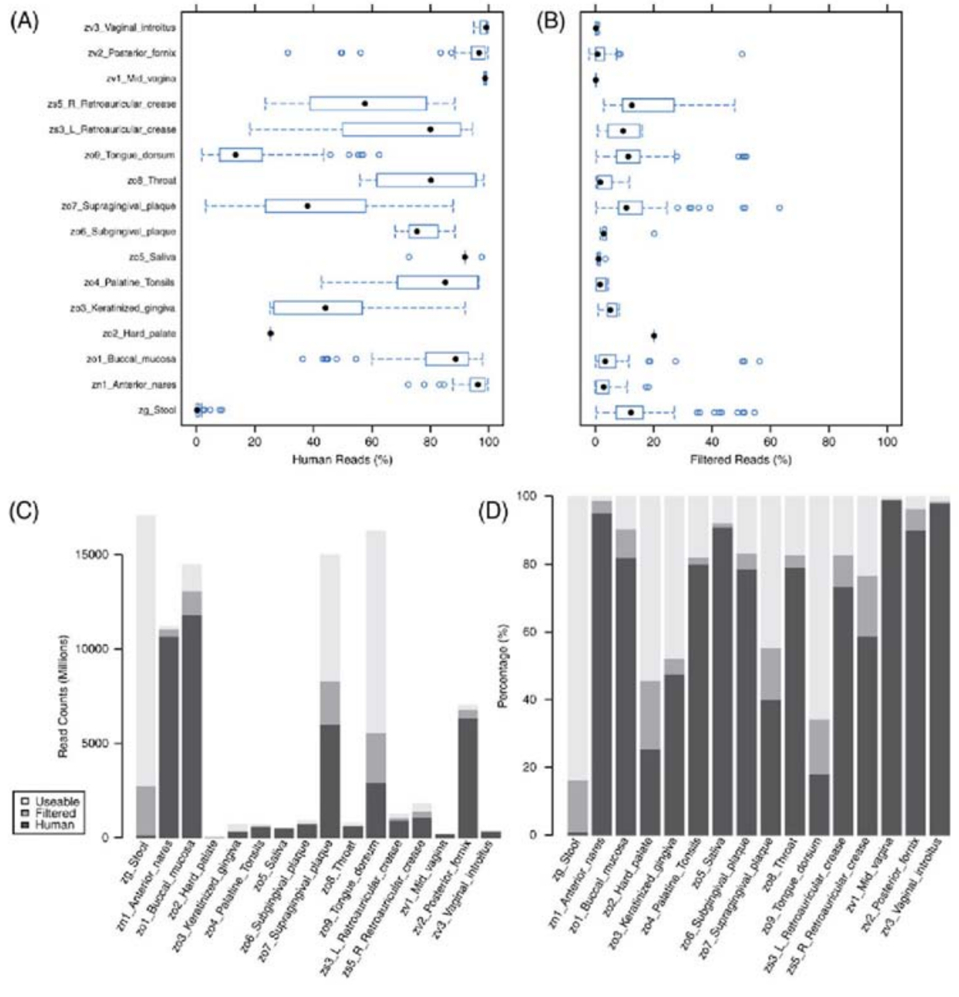

Fig. 2. Impact of quality and human filtering on shotgun metagenomic dataset. Thorough quality filtering and removal of reads resulting from human DNA contamination was performed on all shotgun metagenomic data of the Human Microbiome Project (average of 13 Gb/sample). The variation in fraction of reads per sample removed across the 18 body sites is shown by **(A)** boxplots for % of human and of **(B)** quality filtered reads. **(C)** Total amount of usable data (white) per site significantly varied because of (i) the different number of samples per site, (ii) the differential impact of human contamination (dark gray), and (iii) the differential impact of quality filtering (light gray). **(D)** Summary view of the usable fractions versus human and quality filtered data, per body site. (Reprinted by permission from Macmillan Publishers Ltd.[8])

The statistical analysis plan should take into account the characteristics of the experiment (in human studies, this would be the inclusion and exclusions criteria), the rate of sample acquisition (this would include the rate of human subject recruitment that will determine if you are working with one or more batch of datasets), the descriptive objectives, testable hypotheses, the statistical methods that might be stand alone or imbedded in bioinformatics tools or pipelines, etc. One of the direct advantages of planning ahead is that when you have the data in hand, you'll have a strategy in place to start the analysis. This is critical as next-generation sequencing provides a tremendous amount of data and you want to remain focused on your primary objective(s). After the accomplishment of your primary objective(s), exploratory analyses and additional hypotheses investigation or formulation is always a possibility.

The most basic question about the research plan should be "Are enough samples being collected from each site or from enough subjects to make meaningful conclusions?" To properly assess the degree of similarity or dissimilarity between bacterial communities, a measurable difference, or effect size, is necessary. In general, the smaller the effect size and the greater the variability within a group of samples, the larger the number of samples is required to achieve adequate statistical power.

For determining sample size for experiments using metagenomic taxonomic data, the work derived from the HMP provided the first available calculation and software package[9] (see chapter 6 by La Rosa and colleagues). For metagenomics and metatranscriptomics, standardized methods to assess the number of subjects (or independent samples) and reads are yet to be developed. If you are planning to use both a 16S rRNA gene-targeted approach and whole-metagenome shotgun sequencing, a two-stage experimental design is an option to focus on a subset of samples.[10]

The complexity of your sample will greatly influence the depth of sequence coverage in WMS and metatranscriptomics sequencing projects. As mentioned above, host genomic information can represent a significant amount of genetic data obtained through next-generation sequencing approaches, and this information should be part of an optimized study design.

If the complexity of the sample is low (as determined by more traditional methods), you may be able to estimate the depth of sequencing coverage needed, in order to sample the whole metagenome. Although each next-generation sequencing platform has its unique biases and associated errors (an issue not restricted to next-generation sequencing), metagenomic analyses assume that the reads are sampled randomly, independently, and evenly distributed across all the genomes in the metagenome.[11,12] To calculate the coverage, you need to know the amount of material (nucleotide amount) you are using and the size of the genomes or an average size for that environment. **Figure 3** provides

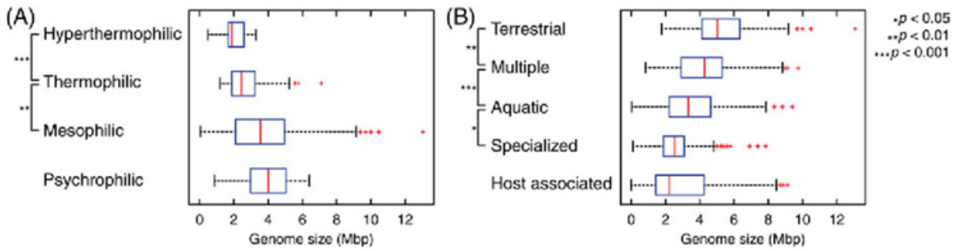

Fig. 3. Distribution of genome size based on temperature and habitat. **(A)** Distribution of genome sizes among prokaryotes with different growth temperature ranges. The differences in genome size between mesophiles, thermophiles, and hyperthermophiles are significant (Wilcoxon rank-sum test, $P < 1.9 \times 10^{-5}$ and $P < 7.9 \times 10^{-3}$ for mesophiles–thermophiles and thermophiles–hyperthermophiles, respectively), but not between psychrophiles and mesophiles (Wilcoxon rank-sum test, $P = 0.082$). **(B)** Distribution of genome sizes among different habitats. Habitats are ordered according to environmental variability from unvarying (host associated) to the most variable environment (terrestrial). The distributions of genome sizes differ between habitats (Wilcoxon rank-sum test, $P < 0.018$, $P < 0.0005$, $P < 0.0028$, for specialized-aquatic, aquatic-multiple, and multiple-terrestrial, respectively), with the exception of host-associated habitats (Wilcoxon rank-sum test, $P = 0.67$, for comparison between host-associated and specialized). The red vertical marks are the medians, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers (99% of all data if the data are normally distributed), and outliers are individually plotted as red crosses. Reprinted by permission from Oxford University Press.[16]

an overview of expected genome size in prokaryotes that can be complemented by other resources providing the exact information on specific genomes.[13–16] The correlation between G+C content and chromosome size can be positive, negative, or not significant depending on the clade from kingdom to species.[15] To our advantage, most chromosomes within a species have a similar pattern of correlation between G+C content and chromosome size; however, outliers are common.[15]

Longitudinal studies present their own challenges and can be independently analyzed at each time point, along the timeline as well as across body sites[17,18] (see chapter 7 in this book). When feasible, the collection of the metadata in between the time points is also critical in understanding the dynamic signatures of microbial population modification. Pooling the samples might seem to be a good strategy to reduce cost and reduce sample variation. However, this approach loses all of the low genetic representation and the ability to make inferences about the microbial population.

You might not find a metagenomic dataset to help or guide you in the experimental design phase. Instead, previous results using other molecular techniques or culture-based methods might be an alternative source of help in the design. If you were looking at the same question with a more traditional method, you should have enough samples to detect differences if they are present.

Metadata Is Needed to Provide Context to the Analysis

Critical to any metagenomic study is the quality and extent of the contextual metadata. Metadata is what will enhance your analysis beyond the most obvious evidence. It provides context to the experiments and allows for meaningful comparisons between studies, while deepening our understanding of the dataset. With a greater depth of information, a broader knowledge of the “environmental factors” is needed. Although not the focus of an experiment, seemingly extraneous data may become important. For example, information on the source of carbon for microbial metabolism might be later identified as a confounding variable in an experiment. It can be as simple as the source of sugar intake for a subject or the nature of the pollutant for a soil sample.

The information about the sample location or its relative position to other samples can be included in the analyses. The concept of biogeography goes beyond the description of environmental features that influence the spatial distribution of the microorganisms. It aims to understand the metabolic processes within the microbes’ own niche and their relationships with other biological niches. The niche might be the different sites in the oral cavity, along the digestive tube, or in the skin.[19–21] Large-scale data visualization and analysis tools have been created to help us better understand these positional aspects[22].

As we are discovering the microbiome as an interdependent organ of any biological system,[5] we may need to redefine what are the best associated data to collect along with the genomic sample. Although blood analyses might reflect the systemic inflammation of a human subject, the levels of air particles less than 2.5 mm in diameter ($PM^{2.5}$) that the subject is exposed to might contribute to the severity of their asthma, modifying the microbiome, which, in turn, can modify the responsiveness to medication.[23,24] In longitudinal datasets, seasons and length of the day have been shown to influence the ocean microbiome.[25]

Defining or re-defining the phenotype of interest might have a crucial importance. Because the phenotype is the results of the interaction between the genotype and the organism's environment in all its complexity, including the microbiome, we are required to renew our attention to the granularity of the defined phenotype. From the macro to the molecular scale, new considerations that were previously neglected because of the lack of significance might be at play when scrutinized with a different sliver or window of observation. Guidelines for data organization and naming standardization are already in place and are being improved upon, as described below.

Sampling: The Basis of Good Results

Although the technology of the sequencing platforms has evolved, they all focus on sequencing the nucleic acids, either DNA or RNA. The source of the microbiome sample greatly varies, from the environment, plants, insects, and animals to humans. The published data on environmental samples have been as diverse as soil, hot springs, seawater, air, as well as home and hospital surfaces. For plants, the associated microbiome above and below the ground has been studied. In insects, animals, and humans, multiple body sites have been investigated. In many of the subsequent steps, the hypothesis involved, the goals of the project, the available facilities and personnel, and the available funds play a role in the decision matrix.

Contamination will be detected as an integral component of the sample because of the depth of the data being acquired. Only a few years ago, understanding microbial diversity often led the investigator to do a series of cloning experiments resulting in the identification of approximately 100 randomly selected organisms per sample. Later, the availability of microarrays allowed the identification of few hundreds of organisms per sample. More recently, by using targeted 16S rRNA gene next-generation sequencing, tens of thousands of organisms can be identified per sample.[1,26] It is recommended to examine each step in the context of potential inadvertent contamination by nucleic material or potential inhibitor for downstream applications. This is particularly applicable to tools that are reused, where proper cleaning and sterilization procedures are essential. The following guidelines are simple ways to increase the quality of sample preparation. Not talking over a biological sample or wearing a facemask would eliminate contamination by the breath. While protecting the sample using gloves, we should not forget that a simple touch of the skin or a surface would contaminate the glove that, in turn, might contaminate the sample itself. Natural DNAses and RNAses may potentially damage the sample. It is most often about applying common sense in the context of the depth of the data to be gathered. In other words, if you want to know the microbiome of the banana peel on the plant but you drop the banana in the field, you are going to also learn about the microbiome of that square of earth as well as that of the fruit.

The proper sampling protocol is essential to a successful metagenomics study, since the accurate identification of many organisms depends on the collection and handling of the sample. Defining the geographical location or the specific body site, surface, depth volume, or quantity to be collected are necessary for sampling standardization. When possible, keep the samples concentrated and process them for immediate storage. Consistency in all aspects will both preserve the quality of the sample and limit the batch effect during the analysis, enhancing the signal of interest. Protecting the samples against the element (wind, sun, etc.) sounds to be a good advice, but keep in mind that sample desiccation is a common problem when working with small samples.

Analyzing true and technical replicates of a sample and assessing whether observed differences are statistically significant are a good practice. True replicates, when the same site is sampled more than once, are rarely done in metagenomics study as the sensitivity of the technique may easily show differences when sampling a site multiple times because of the biological organization of the site.[27] Technical replicates, when the sample is split for processing, are easy to perform for reassurance.[28,29]

Sample Storage

Storage and sampling are tightly linked issues. It is not always possible to have a freezer or an expert on location when the sample is collected. Solutions for these problems affecting the downstream steps need to be identified before starting the study. The nature of the type of sample is too diverse to enter in all the details, but one key question will drive the process: "How much sample do I really need?" The associated questions would be: "Do I need DNA, RNA, proteins, lipids, small molecules, etc., from the same sample?", "Will the sample be used for more than one application, preparation, or extraction?", as well as any other questions related to the present or the future study applications that might be of interest later on.

Many options are available, from immediate extraction to long-term storage in liquid nitrogen. The nature of the sample often dictates what is the best protocol to avoid sample desiccation, denaturation, lysis, degradation, etc. As immediate extraction on site or access to an −80°C freezer is not always an option, alternatives must be developed to preserve the sample, its integrity, and its value for the question(s) at hand. Similarly, for a vaccine, the quality of storage and its consistency might influence the sample quality. Multiple companies are offering sampling kits with fixative but those are rarely validated by comparative analysis. A metagenomic and metatranscriptomic comparison of human stools flash frozen, preserved in ethanol, or in RNA later show that those fixatives are compatible with large-scale self-collection by human subjects in a geographically disseminated cohort.[30]

Sampling cost is often neglected. You might have multiple steps in your process to reach the final storage space, and there is no issue with that. Optimize your process to be the most consistent for each sample or per batch of samples. It should be stressed that whether you work on a large human subject cohort or a large field collection, the cost of personnel, sampling equipment (single use when possible), and transient as well as permanent storage adds up quickly. With the sample collected and in storage, nucleotide extraction will be the next step.

Sample Extraction

The sample input into a metagenomics pipeline can be extremely diverse. The DNA and/or the RNA need to be extracted from the sample prior to any analysis. The type and source of the sample determines the most appropriate extraction protocol. This step, simplified by the availability of nucleic acid extraction kits, is crucial to the success of the analysis, as the quality of extracted DNA and/or RNA influences all subsequent steps. Before selecting the most appropriate extraction protocol, a careful review of the literature and validation of the protocol for your specific sample is recommended. The choice of protocol depends on the DNA or the RNA yield, shearing, removal of contaminants (which could be inhibitory to subsequent steps), and representation of diversity. A compiled list of extraction protocols for different sample sources has been recently published.[31] Some other criteria have to be taken in consideration as described below.

As mentioned above, the source of the sample is very important in the selection of the extraction protocol. A classic example of this is demonstrated by the inhibitory effect of humic acids in enzymatic reactions, such as polymerase chain reaction (PCR), performed using nucleic acids extracted from manure or soil.[32,33] Thus, elimination of humic acids needs to be part of the process, which might be already optimized by a compatible specific kit.

How the sample was preserved also matters. An example is the DNA recovery from formalin-fixed paraffin-embedded tissue, as the tissue is not readily available to traditional protocols.[34,35]

Differences in the structures of bacterial cell walls cause bacterial cell lysis to be more or less efficient.[36,37] The differential efficiency of the lysis can distort the apparent composition of the microbial communities and introduce bias in estimates of relative abundance.[36–39]

Consistency in sample handling and processing is key to avoid batch effect. Training, standard operating procedure, and good quality controls greatly help in minimizing the possibility of batch effect. Nucleic acids extraction automation is a good alternative when sufficient samples are available and the method of extraction has been validated.[40]

Extracting more than one macromolecule at the time is an option. Kits and protocols allow the purification of both DNA and RNA from the same sample, while others go further by recovering proteins as well.[31] An ongoing challenge is to purify other macromolecules from the same sample, which might require a different set of strategies.

Removing the host DNA might improve the quality of your analysis and decrease the cost of the sequencing by requiring magnitude(s) less of reads for the same amount of information. Differential lysis of eukaryotic cells (personal communication, Dr Eva Haenssler) and separation of methylated DNA based on CpG site methylation density between the host and the microbes[41] are the two strategies used by commercial kits. The attempt to decrease host DNA is not only limited to vertebrate hosts but successful contaminant DNA removal have also been performed in plants.[42,43]

Choosing the Right Platform

The cost of sequencing has drastically decreased (**Figure 4**), opening the door to many new investigations that were previously too costly. Although the cost per base of sequencing has decreased, the total cost of a run is still significant because the number of megabases sequenced per run is steadily increasing (**Table 1**). The initial entry cost might be still too high for some pilot projects. Based on those same parameters, traditional techniques such as PCR-DGGE (PCR followed by a denaturing gradient gel electrophoresis), cloning experiments followed by Sanger-based DNA sequencing, and microarrays approaches are here to stay.[44,45]

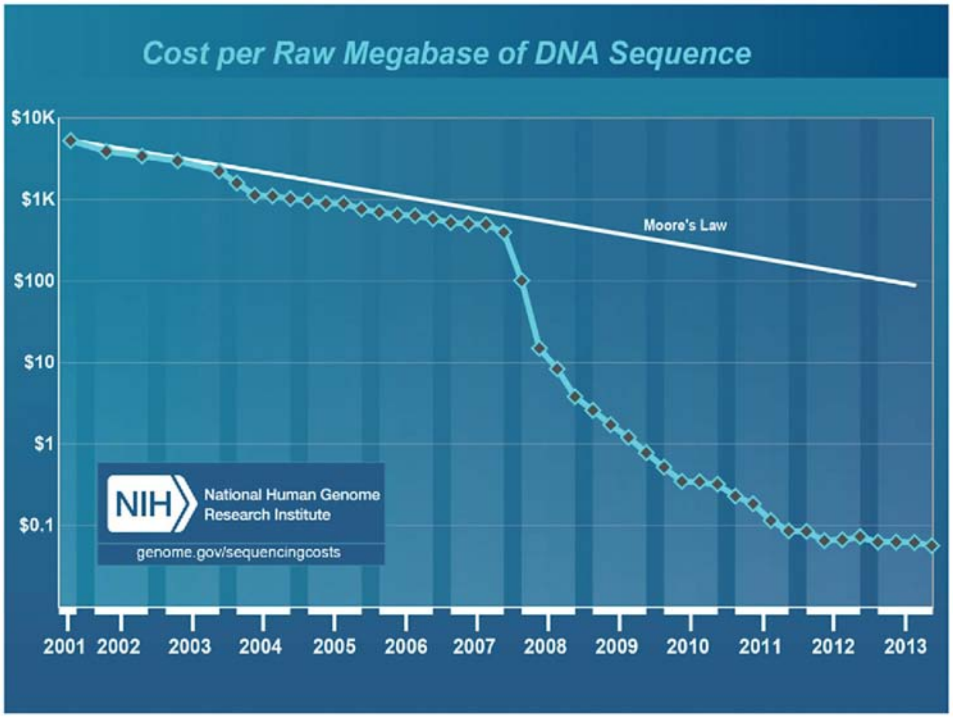

Fig. 4. Reduction of the cost of DNA sequencing over time. The white line reflects the Moore's Law pace. The Y axis shows, in logarithmic scale, the cost of sequencing per raw megabase of raw unassembled DNA sequence. The out-pacing of Moore's Law pace matches the availability of the first next-generation sequencing platforms, in 2008, competing with Sanger-based DNA sequencing technology. (Courtesy of the National Human Genome Research Institute.[45])

Which sequencing platform to use? Because of the varied nature of scientific studies, there is no single approach that is recommended. Detailed review of the literature, discussions with colleagues and sequencing facilities, cost, availability, turnaround time, and scope of the project will be part of the decision-making process. Let's not forget that the hypothesis and the goal should be the true drivers. Table 1.1 shows the characteristics of the different high-throughput sequencing technologies. Each sequencing platform is characterized by their strength and weaknesses regarding read length, bias in AT- or GC-rich regions and their ability to sequence homopolymers.[46,47]

How much sequencing depth is needed? Determining the number of reads required is a tradeoff between the minimal numbers of reads needed to allow an informative and statistical significant analysis and the available budget. This choice is driven by both the platforms and your experimental needs such as the previous knowledge of the relative

Table 1. Sequencing Platforms and Characteristics Based on Online Manufacturer Technical Specifications

Sequencer	*Read Length (b)*[a]	*Run Time (h) (d)*[b]	*Reads Per Run*	*Yield (b)*[a]	*Mate Pair Information*	*Use in Metagenomics*
ABI 3730xl[c]	500–900	6–10 h	–	0.05–0.08 Mb	Yes	Not anymore
Roche 454 GS Junior[d]	~400	10 h	~100,000	35 Mb	No	Yes
Roche 454 GS FLX+[d]	~700	23 h	1 million	700 Mb	No	Yes
Illumina MiSeq[d]	~300	5–65 h	25 million	0.3–15 Gb	Yes	Yes
Illumina NextSeq 500	~300	12-30 h	130–400 million	20–120 Gb	Yes	Yes
Illumina HiSeq 2500[d]	~125–150	7 h to 6 d	300 million to 2 billions	10–180 Gb	Yes	Yes
Illumina HiSeq Xd	~150	<3 d	3 billions	1.6–1.8 Tb	Yes	Not yet
5500 SOLiD[d]	~60	7 d	–	90 Gb	Yes	Yes
5500xl SOLiD[d]	~60	7 d	–	300 Gb	Yes	Yes
Ion PGM system[e]	~200 or ~400	2–4 h or 4–7 h	0.4–0.5 million on 314 chip 2–3 million on 316 chip 4–5.5 million on 318 chip	30–100 Mb on 314 chip 300 Mb to 1 Gb on 316 chip 600 Mb to 2 Gb on 318 chip	No	Yes
PacBio RS II SMRT[e]	4200–8500	0.5–3 h	50,000 per cell	275–375 Mb per cell	No	Yes

a. b stands for base and its multiple

b. h: hours; d: days

c. First-generation DNA sequencing or Sanger-based DNA sequencing technology. ABI 3730xl: Applied Biosystems 3730xl DNA Analyzer (Life Technologies Corporation, Carlsbad, CA).

d. Second-generation DNA sequencing. Roche 454 GS Junior and Roche 454 GS FLX+ systems from Roche Diagnostics Corporation (454 Life Sciences, Branford, CT). Illumina MiSeq, HiSeq 2500 and HiSeq X from Illumina, Inc. (San Diego, VA). 5500 and 5500xl SOLiD sequencer from Life Technologies Corporation (Carlsbad, CA).

e. Third-generation DNA sequencing. Ion PGM system from Life Technologies Corporation (Carlsbad, CA). PacBio RS II SMRT system is based on single-molecule, real-time (SMRT) DNA sequencing technology from Pacific Biosciences (Menlo Park, CA).

abundance of your organism(s) or metabolic pathway(s) of interest. If your metagenome or metatranscriptome is of a relatively low complexity, you can use available genome sequences to evaluate the coverage needed.[48] For a metatranscriptome, you'll have to adapt the sequencing coverage if your focus is the most abundant transcripts or the rare transcripts. It has been shown that millions of 16S rRNA reads do not appreciably increase the extracted information and that a cost-efficient read number is sufficient to discriminate adjacent sites.[1,9] In contrast, during the analysis of the stool microbiome of 100 individuals, increasing the depth of sequencing from 4.5 to 11.7 Gb on average per sample, the human fecal gene catalog increased from 3.3[49] to 5.1 million nonredundant microbial genes,[8] respectively.

Multiplexing of samples has both decreased the cost and allowed to control the number of reads for batch of samples. This approach tags each sample with a unique barcode that is also sequenced. The post-sequencing computing pipeline allows the reads to be binned based on the sample of origin, allowing many samples to be simultaneously sequenced.[50] Additional hidden costs that should be kept in mind are library construction required for preparing the DNA to be sequenced, kits, consumables, labor, instrument initial costs and maintenance, personnel support, indirect cost rate, etc. Further additional costs might be associated with the bioinformatics required for filtering low-quality reads, sequence assembly for pair-ended reads, removing human origin contaminating reads, providing raw or processed reads to your laboratory, and data submission. It's a discussion that you may want to have upfront with your collaborator and/or your sequencing facility of choice.

Read quality is always a parameter to take into account. One of the most common metrics for assessing sequence quality data is the Q score. Low Q scores (below 20) can lead to increase false-positive variants. Q_{20}, which represents an error probability of 1%, is an accepted community standard for a high-quality base, similar to the expectation of Sanger-based DNA sequencing. As the technologies improve, we can expect quality standard of Q_{30} (error probability of 1–1000) and above to be the norm.

Data Storage and Data Analysis

Next-generation sequencing moved us from the kilo- and megabytes size files to the mega- and terabytes size file world. Although this might not be of great importance when you are performing a single metagenomics experiment, it can quickly become an issue in large-scale studies. To put this in context, the HMP 16S rRNA-targeted approach generated about 250 megabases, while the shotgun sequencing approach produced over 3 terabases.[8] While the former can be handled on a traditional computer, the latter requires a lot of computing time (or CPU hours) on a computer or computer cluster with another class of technical specifications. An alternative is the use of remote or cloud computing power through virtual machine approaches.[51] Be sure that the data and related information is secured during transfer and in the cloud.

When focusing on 16S rRNA-targeted approach, the availability of packaged analysis pipelines greatly facilitates the process. Mothur and QIIME are not the only available options, although both have shown consistency of improvement and regular updates over the last few years.[52,53] These pipelines include statistical tools that allow a complete analysis of your dataset including your metadata. As we have been focusing on the quality of the input and output of metagenomic analysis, it is important to note that the denoising step is a crucial step that can increase microbial diversity (up to a meaningless amount if read quality filtering and chimera removal are not performed) or restrict the observed diversity based on the settings.[54] There is a balance that must be attained; however, this can be a bit more difficult to achieve when conducting the investigation of an understudied microbiome.

Whole genome shotgun sequencing leads to the information about the DNA and/or the RNA in the sample. The applications can and have been numerous. The focus might be on metabolism, discovering new metabolic or antibiotic pathways, phylogeny, site comparison, the distribution of single nucleotide polymorphism in the microbiome(s), the influence of cancer or antibiotic treatment, the behavioral effect, etc. From the same dataset, phylogenetic placement of the microbiota present in the sample can be obtained from the gene pool instead of the 16S rRNA gene as their relative abundance in the dataset is low.[8,55,56] Packaged analysis pipelines including statistical tools are available to download or as an Internet resource. An incomplete list of those resources includes CAMERA,[57] EBI metagenomics,[58] IMG/M,[59] MEGAN,[60] METAREP,[61] and MG-RAST.[62] For all metagenomics applications, commercial software replace or complement freely available tools.

All bioinformatics tools rely on databases to add layers of information, from phylogeny to function. While some are based on only one technology (such as the gene catalogs from METAhit and the HMP), others have evolved through generations of approaches and technological advancements such as COG,[63] KEGG,[64] GenBank, and all the other international depositories.[65] The lack of standardization, inconsistent annotation, and the different technologies leading to specific errors unknown to the investigator create some challenges. Curated databases are attempts to limit those issues and often decrease the dataset size by removing information (e.g., sequences) not relevant to the focus in question. Some of these databases include CAZy,[66] Greengenes,[67] HOMD,[14] and MetaCyc.[68] The power of additional layers of information is in their enrichment of the content that we can derive from a dataset. However, we should keep in mind that part of the information from the dataset is unavailable as it did not perfectly match to a previously obtained dataset. With the diversity of microbial strains yet to be sequenced, the answer to your scientific question might reside in the conserved proteins without associated function, or gene(s) or gene set that have never been deposited before.

Data and Publication

Any metagenomic project should include a plan for sharing the data collected to the scientific community, including sequence data and metadata. The International Nucleotide Sequence Database Collaboration (INSDC, http://insdc.org) hosts some of the repositories for the collection and dissemination of nucleic acid datasets. INSDC is a joint effort hosting the following computerized databases: DNA Data Bank of Japan (Japan), GenBank (USA), and the European Nucleotide Archive (based in the United Kingdom).[69]

The need to archive well-defined contextual metadata has been recognized by the community, leading to the creation of the Genomic Standards Consortium. Their mission is to work toward: 1) the implementation of new genomic standards, 2) methods of capturing and exchanging the information captured in these standards, and 3) harmonization of information collection and analysis efforts across the wider genomics community.[70] From this effort arose the creation of minimum information requirement for both genomes and metadata to be submitted to the journal and sequence repositories. The MIGS (minimal information about a genome sequence), MIMS (minimal information about metagenome sequence), MIMARKS (minimal information about marker gene sequence), and MIxS (minimum Information about any (X) sequence) specifications are checklists that both standardize and enhance our ability to further analyze datasets for either training or complementary analysis.[71,72] The adoption of such standards elevates the quality, accessibility, and utility of the information collected by the data repository.

As of yet, there is no standard format to present how the data was analyzed. In the best interests of all, the format should include the methods, tools, and parameters used in the analysis. One option is to make the information available as an online appendix to the published article. There is no such thing as pressing a button and getting the completed analysis. Professional scientists, students, and citizen scientists encounter the same issues. Similar standards of high quality should be put into service for the benefits of the biosphere.

Let’s Talk About the Status Quo

In science, the *status quo*, the existing state of affairs, and the dogma, the established opinion and doctrine, often go hand in hand. Every time a new technology challenges, the *status quo* resistance occurs, not always in the most constructive of ways. It is not our place to choose for you where you stand in the debate regarding the progresses supported by metagenomic approaches. One clear progress is the flow of data. It creates more statistical power to discriminate the aspect(s) of your hypothesis validation, and offers opportunities for validating previously published hypothesis and for hypothesis generation.

What about the ”old data,” the ones published using more restricted or better focused analyses? There is no current methodology that can yet replace quantitative PCR for detecting the relative abundance of host versus microbial genetic abundance. The previous approaches for cultivation-independent analyses are here to guide us by facilitating the analysis and providing the trampoline needed for the next discovery. The high dimensionality of the datasets is potentially a challenge, but it also brings new opportunities to create a validated system biology approach to better understand biological function.

The conceptual and practical details are project specific and all partners should be part of the discussion and project building (primary investigator, co-investigators, statistician, bioinformatician, core facilities, providers, suppliers, IT department, etc.). This is a call to students, professional scientists, and citizen scientists alike, to create new datasets and tools that are needed. Please research, share, and disseminate.

Acknowledgments — We would like to thank the colleagues who shared their frustrations and successes, and Rebecca S. Misra for drawing Figure 1. We apologize to those colleagues whose work was not included for lack of space. The laboratory is supported by grants CA166150 and DK097153 from the National Institutes of Health.

References

1. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 2011;108(Suppl 1):4516–22.
2. Centers for Disease Control and Prevention. CDC Health Disparities and Inequalities Report – United States, 2013. MMWR 2013;62(3).
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291(5507):1304–51.
4. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–14.
5. McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Loso T, Douglas AE, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci USA 2013;110(9):3229–36.
6. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, et al. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 2012;13(6):R42.
7. Giannoukos G, Ciulla D, Huang K, Haas B, Izard J, Levin J, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol 2012;13(3):R23.
8. Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;486(7402):215–21.
9. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PloS One 2012;7(12):e52078.
10. Tickle TL, Segata N, Waldron L, Weingart U, Huttenhower C. Two-stage microbial community experimental design. ISME J 2013;7(12):2330–9.
11. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 1988;2(3):231–9.
12. Wendl MC, Kota K, Weinstock GM, Mitreva M. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J Math Biol 2012;67(5):1141–1161.
13. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2012;40(Database issue):D571–9.

14. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database: J Biol Databases Curation 2010:baq013.
15. Li X, Du D. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PloS One 2014;9(2):e88339.
16. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol Evol 2013;5(5):966–77.
17. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science 2009;326(5960): 1694–7.
18. Gerber GK, Onderdonk AB, Bry L. Inferring dynamic signatures of microbes in complex host ecosystems. PLoS Comput Biol 2012;8(8):e1002624.
19. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, et al. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 2012;13(6):R42.
20. Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW, et al. A diversity profile of the human skin microbiota. Genome Res 2008;18(7):1043–50.
21. Zhang Z, Geng J, Tang X, Fan H, Xu J, Wen X, et al. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. ISME J 2013;(4):881–893.
22. Gonzalez A, Stombaugh J, Lauber CL, Fierer N, Knight R. SitePainter: a tool for exploring biogeographical patterns. Bioinformatics 2011;28(3):436–438.
23. Slaughter JC, Lumley T, Sheppard L, Koenig JQ, Shapiro GG. Effects of ambient air pollution on symptom severity and medication use in children with asthma. Ann Allergy, Asthma Immunol 2003;91(4):346–53.
24. Goleva E, Jackson LP, Harris JK, Robertson CE, Sutherland ER, Hall CF, et al. The effects of airway microbiome on corticosteroid responsiveness in asthma. Am J Res Crit Care Med 2013;188(10):1193–201.
25. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, 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.
26. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, et al. The human oral microbiome. J Bacteriol 2010;192(19):5002–17.
27. Zeeuwen PL, Boekhorst J, van den Bogaard EH, de Koning HD, van de Kerkhof PM, Saulnier DM, et al. Microbiome dynamics of human epidermis following skin barrier disruption. Genome Biol 2012;13(11):R101.
28. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS One 2012;7(2):e30087.

29. Aird L, Anderson S, Jumpstart Consortium Human Microbiome Project Data Generation Working Group. et al. Evaluation of 16S rDNA-based community profiling for human microbiome research. PloS One 2012;7(6):e39315.
30. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci USA 2014; 111(22):E2329–2338.
31. Dhaliwal A. DNA extraction and purification. Mater Methods 2013;3:191.
32. Watson RJ, Blackwell B. Purification and characterization of a common soil component which inhibits the polymerase chain reaction. Can J Microbiol 2000;46(7):633–42.
33. Yoshikawa H, Dogruman-Al F, Turk S, Kustimur S, Balaban N, Sultan N. Evaluation of DNA extraction kits for molecular diagnosis of human blastocystis subtypes from fecal samples. Parasitol Res 2011;109(4):1045–50.
34. John BA, Okello JZ, Devault Alison M, Kuch Melanie, Okwi Andrew L, Nelson K, et al. Comparison of methods in the recovery of nucleic acids from archival formalin-fixed paraffin-embedded autopsy tissues. Anal Biochem 2010;400:110–7.
35. Su JM, Perlaky L, Li XN, Leung HC, Antalffy B, Armstrong D, et al. Comparison of ethanol versus formalin fixation on preservation of histology and RNA in laser capture microdissected brain tissues. Brain Pathol 2004;14(2):175–82.
36. Chassy BM, Giuffrida A. Method for the lysis of Gram-positive, asporogenous bacteria with lysozyme. Appl Environ Microbiol 1980;39(1):153–8.
37. Frostegard A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, et al. Quantification of bias related to the extraction of DNA directly from soils. Appl Environ Microbiol 1999;65(12):5409–20.
38. Salonen A, Nikkila J, Jalanka-Tuovinen J, Immonen O, Rajilic-Stojanovic M, Kekkonen RA, et al. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods 2010;81(2):127–34.
39. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. PloS One 2010;5(4):e10209.
40. Nylund L, Heilig HG, Salminen S, de Vos WM, Satokari R. Semi-automated extraction of microbial DNA from feces for qPCR and phylogenetic microarray analysis. J Microbiol Methods 2010;83(2):231–5.
41. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PloS One 2013;8(10):e76096.
42. Wang HX, Geng ZL, Zeng Y, Shen YM. Enriching plant microbiota for a metagenomic library construction. Environ Microbiol 2008;10(10):2684–91.
43. Okubara P, Li C, Schroeder K, Schumacher R, Lawrence N. Improved extraction of *Rhizoctonia* and *Pythium* DNA from wheat roots and soil samples using pressure cycling technology. Can J Plant Pathol 2007;29(3):304–10.
44. Zimmerman N, Izard J, Klatt C, Zhou J, Aronson E. The unseen world: environmental microbial sequencing and identification methods for ecologists. Front Eco Environ 2014;12(4):224–31.

45. Institute NHGR. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program 2014 (Accessed February 12, 2014). http://www.genome.gov/sequencingcosts/

46. Dark MJ. Whole-genome sequencing in bacteriology: state of the art. Infect Drug Resist. 2013;6:115–23.

47. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13:341.

48. Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 2012;13:734.

49. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464(7285):59–65.

50. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PloS One 2008;3(7):e2836.

51. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinform 2011;12:356.

52. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75(23):7537–41.

53. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7(5):335–6.

54. Gaspar JM, Thomas WK. Assessing the consequences of denoising marker-based metagenomic data. PloS One 2013;8(3):e60458.

55. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9(8):811–4.

56. Tu Q, He Z, Zhou J. Strain/species identification in metagenomes using genome-specific markers. Nucleic Acids Res 2014;42(8):e67.

57. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res 2011;39(Database issue):D546–51.

58. Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, et al. EBI metagenomics: a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 2014;42(Database issue):D600–6.

59. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 2008;36(Database issue):D534–8.

60. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome Res 2011;21(9):1552–60.

61. Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methe BA, et al. METAREP: JCVI metagenomics reports – an open source tool for high-performance comparative metagenomics. Bioinformatics 2010;26(20):2631–2.
62. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 2008;9:386.
63. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28(1):33–6.
64. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38 (Database issue):D355–60.
65. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2008;36(Database issue):D25–30.
66. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 2014;42(1):D490–5.
67. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72(7):5069–72.
68. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2010;38(Database issue):D473–9.
69. Cochrane G, Karsch-Mizrachi I, Nakamura Y. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2011;39(Database issue):D15–8.
70. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic Standards Consortium. PLoS Biol 2011;9(6):e1001088.
71. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 2011;29(5):415–20.
72. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 2008;26(5):541–7.