Briefings in Bioinformatics Advance Access published May 15, 2013

Environmental bio-monitoring with high-throughput sequencing

Jing Wang, Patricia A. McLenachan, Patrick J. Biggs, Linton H. Winder, Barbara I. K. Schoenfeld, Vinay V. Narayan, Bernard J. Phiri and Peter J. Lockhart

Submitted: 6th February 2013; Received (in revised form): 11th April 2013

Abstract

There is much interest in using high-throughput DNA sequencing methodology to monitor microorganisms, complex plant and animal communities. However, there are experimental and analytical issues to consider before applying a sequencing technology, which was originally developed for genome projects, to ecological projects. Many of these issues have been highlighted by recent microbial studies. Understanding how high-throughput sequencing is best implemented is important for the interpretation of recent results and the success of future applications. Addressing complex biological questions with metagenomics requires the interaction of researchers who bring different skill sets to problem solving. Educators can help by nurturing a collaborative interdisciplinary approach to genome science, which is essential for effective problem solving. Educators are in a position to help students, teachers, the public and policy makers interpret the new knowledge that metagenomics brings. To do this, they need to understand, not only the excitement of the science but also the pitfalls and shortcomings of methodology and research designs. We review these issues and some of the research directions that are helping to move the field forward.

Keywords: *metagenomics; environmental bio-monitoring; high-throughput sequencing*

INTRODUCTION

The application of high-throughput sequencing protocols in metagenomics [1] offers the hope of a cost-effective and comprehensive means of assessing biotic diversity and ecological relationships for many complex animal, plant and microbial ecosystems. These protocols have the potential to advance understanding of human health [2], ecosystem health [3], food safety and security [4]; identify novel energy sources and drugs [5, 6]; facilitate large-scale monitoring of mammalian and endangered biodiversity [7]; uncover the nature of symbiotic [8] and endosymbiotic [9] relationships; determine the specificity of insects acting as biological control agents [10, 11]; and investigate the temporal and spatial trophic interactions of invertebrates, microbes and plants in farmed [12, 13] and natural [14] ecosystems—including evaluation of adaptive responses to environmental change [15, 16]. These examples illustrate some of the many questions that can now be addressed with high-throughput sequencing and metagenomics. That said, application of the

Corresponding author: Peter J. Lockhart, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. Tel: +646-356-9099; E-mail: P.J.Lockhart@massey.ac.nz

Jing Wang works as a bioinformatician in the pathogen discovery research group at the Institute of Environmental Science and Research (ESR), Upper Hutt, New Zealand. She is also a bioinformatics PhD student at the Technical University of Clausthal in Germany, with research interests in data mining techniques and their applications in biological databases areas.

Patricia McLenachan is a molecular biologist who provides technical support for the Massey University Genome Service.

Patrick Biggs is a bioinformatician/computational biologist specializing in high-throughput sequencing applications.

Linton Winder is an ecologist and entomologist, investigating the application of high-throughput sequencing for constructing ecological networks.

Barbara Schoenfeld is a PhD candidate who is studying recent and ancient endosymbioses in plants.

Vinay Narayan is a PhD candidate who is investigating marine ecosystems of Fiji.

Bernard Phiri is a PhD candidate who is investigating waterborne disease ecology in New Zealand.

Peter Lockhart is a recent recipient of a New Zealand Royal Society James Cook Fellowship who has been investigating applications of high-throughput sequencing in evolutionary ecology and systematics. He is Adjunct Professor of Biology at UNITEC (Auckland), and at the University of the South Pacific.

technology is not without its challenges. Educators need to be aware, not only of the potential for advancing scientific discovery but also of the methodological issues confronting researchers and the practical solutions being pursued.

Experimental and sampling issues

Southwood's classic 1966 text 'Ecological Methods' discusses the crucial importance of sampling strategies in ecological studies. The need for preliminary sampling, consideration of the number of samples needed and the importance of spatial and temporal aspects when designing such studies are all included. The principle of 'garbage in, garbage out' is as true today as it was when the most sophisticated tool that a biologist had was a quadrat; yet, there is undoubtedly a risk that researchers using high-throughput sequencing get caught in the headlights of new technology and fail to follow these fundamental and established principles of good experimental design [17, 18]. Cost effective yet properly replicated robust sampling is essential for any ecological or environmental study [19]. Although early shotgun-sequencing projects were understandably unreplicated or at best poorly replicated, the reduction in costs in high-throughput sequencing provides a new opportunity for studies to be undertaken that are robustly designed—and will ultimately reveal far more about ecological communities than work conducted before the genomic revolution [5].

Both the detection of species and their abundance may be achieved by extraction and sequencing of DNA [20], but a key question before interpreting data is whether the information gained is representative of the environment from which the sample was taken. To address this fundamental issue, studies should consider the following:

- sufficient replication [19];
- temporal and spatial heterogeneity, with the acceptance that heterogeneity in time and space within a sampled environment can significantly affect detection of species and estimates of relative abundance [21–24];
- the use of traditional accumulation curve techniques and associated rarefaction methods, recognizing that the primary goal of many biodiversity-related studies is to estimate either richness or diversity [25–30];
- sampling strategies should result in community and sample representativeness when estimating richness or diversity patterns [31] at the appropriate level of α, β or γ-diversity [32] for a given study;
- use of increasingly sophisticated computational approaches to optimally design sampling strategies, particularly in relation to the desired precision of an estimate and its confidence interval [1, 33–37];
- sequencing methodology is not without error, and this can lead to a false perception of diversity [38].

Sources of bias and error in the application of high-throughput sequencing techniques in metagenomics are described later in the text. Most observations have to date been made in studies involving microbial profiling. The findings are relevant to studies of other biological systems.

Inefficient DNA extraction will mislead community analyses

DNA extraction protocols that work more efficiently on some organisms than others have been shown to introduce a bias into estimates of microbial community diversity and structure. This has resulted in some species being over-represented and others being under-represented or absent [21, 39–45]. Although extraction can appear to yield DNA of good quality and quantity, where a protocol fails to extract the DNA from all organisms that are present, lower than anticipated diversity can be expected [40, 46]. This can result from inefficient cell lysis (cell wall or membrane differences) and or because other physiological and chemical properties of the organisms or their natural environment cause the DNA to be lost before, or during the extraction procedure (e.g. osmotic shock in low salt extraction buffers can cause loss of DNA during extraction protocols with salt tolerant species [47], DNA can adhere to mucus, extracellular matrix and/or soil particles [21, 44], DNA with low GC content can also degrade [47]).

Recently, the efficiency of DNA extraction methods has been investigated using mock communities (*in vitro* studies). Cultures of different (known) species and relative concentrations are combined for library construction and sequencing. The recovery and abundance of sequences has been compared with the known composition of the sample. Notably, significant variation in community structure and in the abundance of species has been observed between different extraction methods and from what was expected given the starting material [47–49]. The study of mock communities has an important place in the future of metagenomics, as they provide a means of investigating the relationship between biomass and numbers of sequence reads.

The purpose for the study has determined how researchers have been addressing these pitfalls; studies that make a 'snap shot' through time or geographical space might use a single method of extraction and live with the limitations of that method. Studies that aim to sample the full diversity of an ecological community suggest using several different DNA extraction methods [21, 43].

A low yield of DNA has been a problem for high-throughput sequencing protocols—pyrosequencing and some Illumina protocols require microgram amounts for DNA library construction. Low amounts of starting material can be a problem for detecting low abundance sequences [50, 1 and see references within]. In contrast, the Illumina NEXTERA library preparation method requires only 1 ng of DNA template. This approach involves transposon-mediated enzymatic shearing and is potentially susceptible to non-random coverage and contaminants in poor quality DNA preparations, which can reduce its efficiency (unpublished observations). Whole-genome amplification is an alternative for obtaining larger amounts of DNA starting template. However, it introduces its own biases and sources of error [1 and references within 21].

Random shotgun or amplicon sequencing?

High-throughput sequencing protocols for microbial profiling have involved either (i) polymerase chain reaction (PCR) amplification and sequencing of targeted gene loci (amplicon sequencing) and/or (ii) the sequencing of random genome fragments. Amplicon sequencing has most often been conducted for 16S rRNA sequences, as phylogenetic coverage in databases is greater for this molecule than any other molecule. Although amplicon sequencing is cost effective, PCR and sequencing of amplicons has a number of technical challenges that introduce biases into estimates of biological diversity. These issues are highlighted later in the text.

Random sequencing protocols involve fragmenting genomic DNA and then sequencing these fragments. This approach has been used less for microbial profiling than has 16S rDNA amplicon profiling. However, this situation might change with the appearance of platforms such as Illumina's MiSEQ [51] and Life Technology's Ion Torrent [52], as greater sequencing depth can be obtained at relatively low cost. There is an issue with the numbers of sequences required (sequencing effort) in a metagenomic data set, to ensure the recovery of low abundance members [53, 54]. Observed diversity increases as the total number of sequences increases. Hence, it is important to normalize the number of sequences analysed across different samples [53] and/or use statistical analysis methods that account for data sets with vastly different numbers of reads [54]. Simulations have shown that sequencing coverage impacts significantly on estimates of diversity, and phylogenetic methods of assessment become more important at low levels of sequence coverage [53].

Methods for random genome sequencing and targeting specific gene loci are not necessarily independent. From random shotgun sequences, rDNA sequences, for example, can be retrieved from the library of fragments and analysed separately [54–56]. As discussed under the section on 'Analytical Issues', there are specific computational issues relevant to different data types. Here, we first describe the nature of experimental biases that are important to consider for subsequent analyses.

Amplicon sequencing biases

There are a number of recognised properties of PCR that can mislead biodiversity estimates. These include polymerase error (which is estimated at 1 substitution per 10^5–10^6 bases [48]), the formation of chimeric/heteroduplex molecules [48, 57–59] and differential amplification efficiency [48, 50, 60–62]. Recommendations for laboratory practices that reduce these biases have been made [60 and references therein]. These include lowering PCR cycle numbers [59], the pooling of multiple reactions, high (>4 ng) template concentrations and the use of a proof-reading polymerase ([60], but also see [63]).

Some studies have investigated the effect of primer mismatch on amplification efficiency using simulated communities, and most concur that this is a major factor leading to errors in detection of taxa and the distortion of taxon frequencies within a community [48, 60, 61]. In a complex mixture of templates, sequences that do not have 100% match with the primer sequences can amplify at low efficiency or not at all [48, 60, 61], whereas perfectly matched sequences will be preferentially amplified and over-represented. Some studies recommend using degenerate primers or a mixture of non-degenerate primers, to overcome the problems of primer mismatch and amplification efficiencies [48, 50, 60, 62].

Mao *et al*. [62] have investigated the coverage rates of eight commonly used universal 16S rRNA primers, against the Ribosomal Database Project and against seven metagenomic data sets. They found that although some primers were genuinely universal (e.g. 1390R, 1492R) and showed high coverage over a wide selection of taxa, some (e.g. 27F) could be improved by the addition of degenerate bases, and others (e.g. 519F) missed particular phyla altogether. Hong *et al*. [43] estimate that in some cases, as much as 50% of microbial diversity can be absent when a single set of primers is used to amplify template DNA.

The choice of amplicon(s) is important; the most variable region of the 16S rRNA molecule appears to be the V1–2 region [50], but there are nine variable regions in total in the gene. The observed diversity within the sample can vary depending on which regions are chosen [48–50, 62, 64], and the choice of region can determine what stringencies of quality-score refinement should be used for data analysis [37, 56]. Amplicon length has been shown to affect the assessment of the number and relative abundances of species from communities from the termite hind gut [50] and from hydrothermal vent fluids [65]; in both studies, it was found that libraries constructed from smaller amplicons (<400 and 100 bp, respectively) contained greater species diversity with species of low abundance and more divergence being represented. Longer amplicons were disproportionately lost in downstream bioinformatics owing to errors, and the libraries contained more artefacts such as chimeras, heteroduplexes and misprimed sequences.

A high annealing temperature in the PCR reaction can exacerbate biases caused by differences in primer homology, and the greatest diversity is seen when low (47–52°C) annealing temperatures are used [60, 61]. The effect of varying the number of cycles in the PCR reaction has been investigated; Sipos *et al*. [61] found little difference, but others [59, 60] recommend using the minimum number of cycles (between 12 and 30) necessary to provide sufficient template for the next step in the sequencing protocol. This number varies between samples and should be determined experimentally. Ahn *et al*. [63] found the proportion of chimeric sequences could be reduced significantly by reducing the number of PCR cycles from 30 to 15 (32–1%, respectively), and that the numbers of chimeric sequences were higher when a high fidelity Taq was used for PCR. There is risk of chimera formation when partially formed PCR products act as primers to amplify homologous and/or similar sequences. The rate of chimera formation has been suggested as being from 5 to 45% [48, 57], which underscores its importance for consideration.

We reiterate here, the point noted by Schloss *et al*. [48], that the various platforms of high-throughput sequencing have been developed primarily for genome sequencing. Accuracy is not so much of an issue, given the high coverage afforded by the assembly of multiple reads. Error rates for an assembled genome are low due to coverage, although error rates for an individual sequence might be high.

The error rate of 0.01–0.02 errors/per total base call for Roche-454 sequencing is considered to be high [57]. PCR and sequencing errors can create singleton Operational taxonomic Units (OTUs) and lead to an overestimation of species richness in a sample. This can be overcome by using stringent post-sequencing quality filtering—for example, by excluding singleton OTUs and sequences that cannot be taxonomically classified from the analysis [48, 49, 58, 59].

PCR primers for the 16S rRNA V3 region have been shown also to be non-specific, amplifying 18S rRNA sequences, which have been wrongly annotated in GenBank as 16S rRNA gene sequences [66].

Variation in 16S rRNA copy number is a further source of error, as estimates of relative abundance of 16S rRNA sequence types can be affected by both copy number variation and organism abundance. In the novel study by Kembel *et al*. [67], copy number was normalized by using phylogenetic assignment of 16S rRNA sequences, and ancestral state reconstruction was used to infer copy number in unknown environmental samples. In analyses of 16S rRNA data from two previous environmental studies, they showed that differences in 16S rDNA copy number could sometimes lead to underestimation of the most abundant taxa and an overestimation of rare taxa.

ANALYTICAL ISSUES

Improving quality of the data

The past 12–24 months has seen the publication of many articles reporting the nature and significance of Roche 454 sequencing errors for metagenomic studies. Others acknowledge the likelihood and impact of errors with other technologies including Illumina sequencing and also more recently for the Ion Torrent. However, few direct comparisons have been reported [68].

High-throughput sequencing, and pyrosequencing in particular, is well known to be susceptible to sequencing errors that can falsely elevate estimates of species diversity by an order of magnitude [56]. Approaches to reduce this impact of intrinsic error include quality score analyses and modifications to alignment and/or clustering methods [56]. Quality score analysis involves removal of low quality sequences from the data set, or parts of sequences, depending on the algorithm used, and the kinds of downstream analyses to be performed. It is also important to remember the different error profiles generated by Illumina sequencing by synthesis approaches, in comparison with those generated by pyrosequencing technologies. Consequently, there is an algorithmic difference in the way these quality scores are processed, in that quality trimming and/or processing tools are platform-specific.

High levels of technical replicate variation have been reported in 16S rRNA amplicon sequencing [57, 69]—intra and inter-sequencing centre variation can be significant—variation in thermocycler calibration, reagent concentration and sampling artefacts are reported explanations for technical variation [48, 57, 69]. Such errors could significantly alter estimates of β diversity [69].

A number of analytical approaches and pipelines have been developed to reduce sequencing error rates with 454 and Illumina sequencing [57]. These include (i) removal of reads with ambiguous base calls; (ii) trimming sequences with low quality scores; and (iii) for 454 data application of (a) computationally intensive denoising algorithms (such as PyroNoise and DeNoiser, which correct base calls by modelling the original flow diagram [57]) and (b) less computationally intensive algorithms (single linkage clustering and SeqNoise). There are also heuristics that fit observed number of OTUs to expected number of OTUs—these are used to reduce the number of spurious OTUs and phylotypes. The study of Schloss *et al*. [48] implemented a quality filtering pipeline to better understand the effect of different sources of error on microbiome interpretation. Their study provides insight into the sources of error that can confound environmental community analyses.

In general, pipelines for analyses in metagenomics require validation before comparative analyses are undertaken. Figure 1 illustrates this point for Illumina data collected for a terrestrial freshwater sample. The difference in estimates of relative abundance of different bacterial groups is the result of using different orders of operation (read overlapping and quality trimming) in pipeline processing. The cause of this difference is currently undetermined, but it provides a further point of caution emphasizing the importance of standardization of protocols that are to be applied in comparative studies of environmental monitoring [37]. In other words, the exact description of the bioinformatics processing is as important as the processing of the sample in the laboratory after its collection.

Taxonomic classification

Taxonomic classification is often the first step in a metagenome project. After receiving millions of short reads from a high-throughput sequencing platform, the first question to answer, after taking steps to ensure data quality is 'What are they?'. Accurate and robust prediction of the source organism for each short read is essential for identification and enumeration of the organisms in a given sample. Such knowledge provides a 'roadmap' and foundation for ecological and environmental studies.

-There are three major approaches being used for taxonomic assignment: phylogenetic-tree-based methods, similarity search methods and composition-based methods (Table 1). Development of software tools to implement these approaches has been an active research area. Here, we only list some of the most popular tools (Table 1). Discussion of these tools here is space restricted.

Phylogenetic methods are based on reconstruction of evolutionary models for targeted molecular markers. 16S rRNA genes have been most commonly used for microbial studies, and this is strongly reflected in experimental studies published to date. Although useful because of the extent of phylogenetic representation in databases, in some instances, this molecule has been found to exhibit insufficient phylogenetic resolution for species identifications (e.g. as with some species of *Vibrio* [92]). The mitochondrial cytochrome c oxidase subunit I (COI) gene is a popular candidate marker for animals [93], whereas the chloroplast genes for the large subunit of ribulose-bisphosphate carboxylase (*rbc*L) and and a group II intron splicing factor (*mat*K) [94, 95]

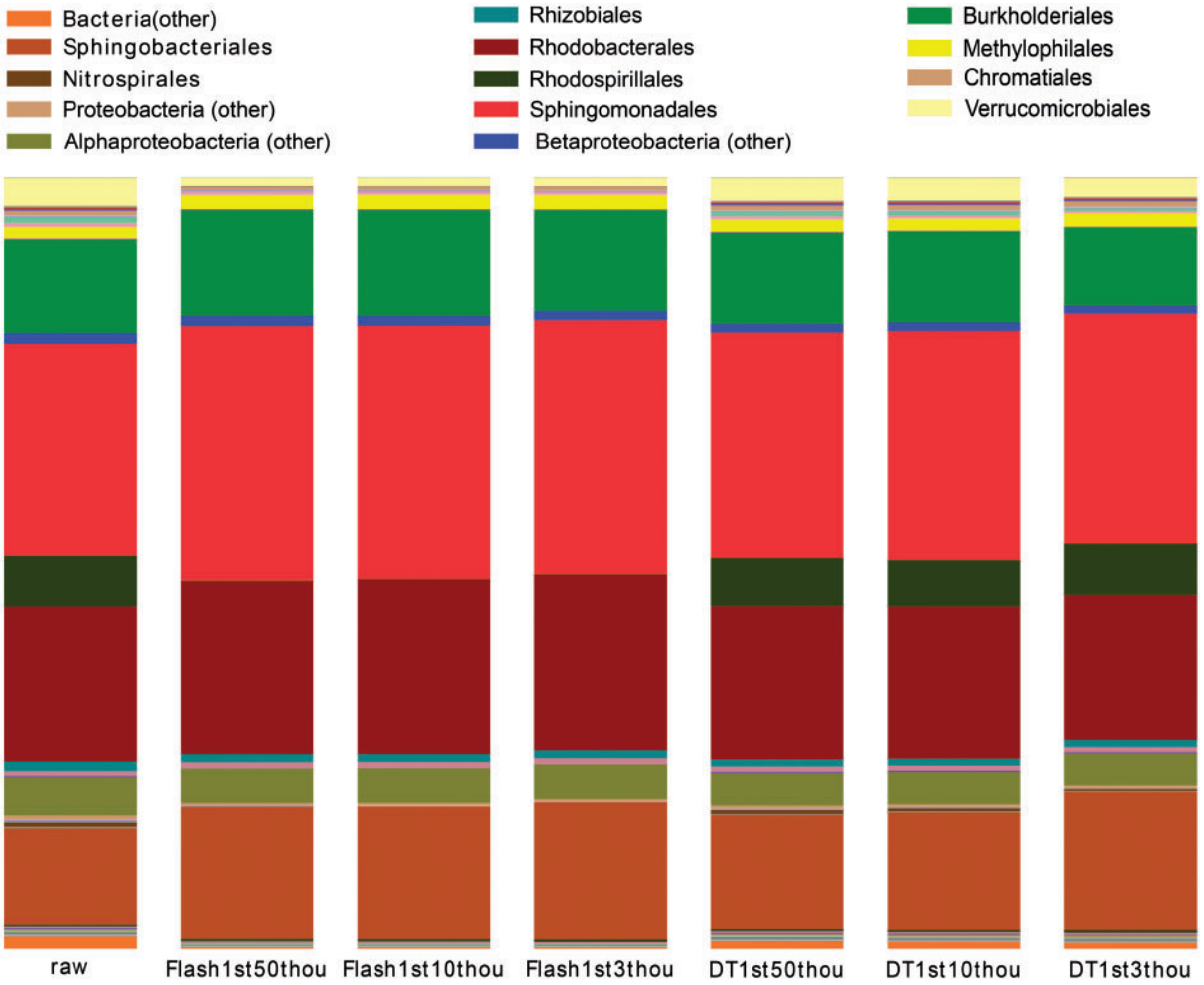

Figure 1: An example of the effect of the ordering of bioinformatics processing on 16S rRNA PCR products generated using an Illumina MiSeq with 150 bp reads. Approximately 700 000 reads were pre-processed using either the FLASH overlap aligner [70] or DynamicTrim (v. 2.0; part of the SolexaQA suite of Perl scripts [71]) first, and then processed with the other tool as a second round of processing. Quality trimming was performed with DynamicTrim at three quality levels (0.05, '50thou'; 0.01, '10thou'; 0.003, '3thou') to see whether this had an effect. In all, 200 000 reads of 253 bp were then taken and run through the QIIME pipeline (v. 1.5.0; default parameters [72]) in comparison with 200 000 unprocessed sequences ('raw'). The figure shows a distribution at the taxonomic level of order, with bacterial orders indicated where they were present at over 0.5%. It can be seen that the distributions obtained when FLASH was used first similar to the 'raw' data set and is different to that obtained when DynamicTrim was used first. The absence of Rhodospirillales in the FLASH first data sets is particularly noticeable in this example.

have been accepted for use in plants. Recognition of the potential significance of hybridization in animal and plant evolution [15] has also meant that nuclear markers have been considered [93, 96, 97]. There is no universal sequence found across viral genomes [98], and perhaps for this reason, phylogenetic methods have not been widely applied in viral metagenomic studies. In addition, sequencing design can limit application of phylogenetic methods. For example, although phylogenetic methods can be applied to shotgun metagenome approaches, if the genome coverage is insufficient, gene markers might constitute only a small percentage of a given sample, resulting in comparative data not being available.

Similarity search methods include comparison-based, homology or alignment-based methods. Basic Local Alignment Search Tools (BLAST) [99] and profile Hidden Markov Models (pHMM) [100] are two algorithms that have been successfully implemented for helping to identify homologies. They compare metagenome sequences to reference

Table 1: Commonly used approaches for taxonomic assignment

	Phylogenetic methods	Similarity search	Composition-based
Underlying Algorithms and Methods	Maximum likelihood, Bayesian Inference, Neighbour-Joining	BLAST pHMM LCA	Interpolated Markov models NBC k-means/k-nearest-neighbour
Pros	Marker gene databases and multiple alignments are well curated and maintained.	Some matured pipelines have been tested and applied. Makes use of all available reference data and is therefore the most comprehensive method for detecting taxa that have already been described.	Faster than the similarity-based approach once the model is built.
Cons	Not applicable for viruses, as there are no universal markers. Thus, design of primers for more specific loci is required.	Similarity search, i.e. BLAST search, is computational intensive. False positive assignments need to be examined manually. In most cases, this will not be practical. The assignment of reads is limited to the taxonomic range represented in the database.	Requires rather long sequences as analysis inputs. Needs to improve accuracy
Implemented Applications	EPA [73] FastTree [74] pplacer [75] Greengenes [76]	MEGAN [77] Sort-ITEMS [78] CARMA3 [79] MetaPhyler [80] QIIME [72]	INDUS [81] NBC [82] MetaBin [83] TACOA [84] MetaCluster [85, 86]
		PhymmBL [87]	
	AMPHORA [88], MLTreeMap [89] and SAP [90]		
		SPHINX [91]	

databases. Although BLAST is commonly used and effective at identifying homologies, there are computational issues to consider, and these have been outlined in the section later in the text. Once a BLAST search is completed, a key task is interpreting the BLAST output. Early programs like MG-RAST [55] assume that only the best hits with low e-values are to be trusted, and less significant hits are discarded. Such an approach lacks sensitivity, and the results need to be interpreted carefully. An alternative approach is the Lowest common ancestor (LCA) algorithm implemented in MEtaGenome ANalyzer (MEGAN) [78, 101, 102]. LCA allows sequences to be assigned to higher taxonomic levels if the minimum assignment of a taxonomic node on the NCBI tree of life does not meet a threshold of statistical significance or where there is ambiguity in the assignment of query sequences to database sequences. In the latest version of MEGAN [102], the 'min support' parameter (minimum number of reads that need to be assigned to a taxon to identify it) has been increased from 3 to 5. This makes the approach more conservative, but the parameter needs to be evaluated in the context of the increasing length of high-throughput sequencing reads and the trade-off between sensitivity and accuracy. The effectiveness of MEGAN is also impacted by database representation. This potential problem is illustrated in Figure 2.

If the sequence from which a query sequence originates is not in the database, then MEGAN will generally assign the read to the most closely related homolog in the database. For example, relatively few complete genome sequences are available for protists, thus query sequences matching orthologues in these organisms are often only distantly related. LCA is a powerful approach that makes maximum use of available reference sequences, but its susceptibility to uneven representation of taxa and missing data in the reference databases needs to be taken into account when interpreting results.

An integrated environment approach, provided by the Python-based software pipeline Quantitative Insights Into Microbial Ecology (QIIME) [72] has been devised to analyse 16S rRNA amplicon sequences from 454 (and now also Illumina) sequence data. This publically available software is available as a pre-packaged virtual machine, ready to perform analyses once installed. There are advantages in this approach to software distribution, as the process then

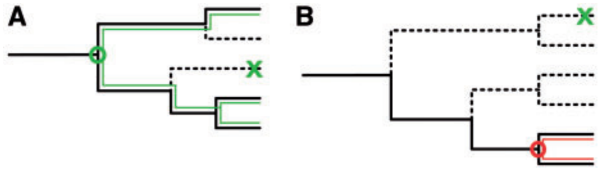

Figure 2: Impact of database representation on LCA assignment to ancestral nodes. Solid lines link homologues identified by BLAST to a query sequence. Dashed lines link query sequences to homologues that are not represented in the data base. Circles indicate the nodes to which LCA assigns the query sequence, whereas crosses identify the correct placement of the query sequence within the taxonomic hierarchy. (**A**) LCA assigns the query sequence to the node of the most recent common ancestor of all taxa that produce significant BLAST hits. Where the source organism or close relatives of the source organism are present, LCA provides a good indication of the taxonomic group to which the source organism belongs; (**B**) where the source organism and relatives of the source organism are poorly represented in the database, LCA can give a misleading assignment of the taxonomic group. Our observations are that this problem of miss-assignment is more significant for taxa whose genomes have not been fully sequenced.

becomes less dependent on the user being able to set up a bioinformatics pipeline for their analyses, and having all the necessary internal dependencies fulfilled. QIIME performs many analysis steps (read filtering, OTU picking by sequence similarity, taxonomic assignment, phylogenetic tree generation, diversity metric calculation and visualization by various methods including principal components analysis), and can be set up to run in a serial or parallel mode.

Composition-based methods have three steps. The first step involves computing a model or a profile on a set of known sequences. Interpolated Markov models [103] are typically developed for a set of reference sequences for which there is known taxonomic information. Features like GC content, codon usage or oligonucleotide frequencies are common characteristics used for computing such models. The second step involves characterizing an unknown set of metagenomic sequences for the same features that are used to describe the set of reference sequences. In the third step, a comparison is made of the reference and metagenome profiles so that taxonomic ranks can be assigned to the metagenomic sequences. The model building can be computationally expensive. However, once the models are built (step 1), the latter steps are generally faster than alignment-based methods such as BLAST.

To achieve the best results, sometimes instead of using only one of the strategies discussed earlier in the text, combined methods are implemented (Table 1). For example, SPHINX a hybrid binning approach combines similarity and composition methods. In SPHINX, an extended LCA method is implemented with a k-mer filtering step. Protein encoding sequences from microbial genomes are clustered based on their tetra-nucleotide frequencies with a k-means clustering approach. For each cluster, a centroid is computed, and the sequences are translated into protein sequences. The first step in the taxonomic classification consists of computing the distance of the metagenomic fragment to all cluster centroids. The fragment is then assigned to the cluster whose centroid has the smallest distance. After a BLASTx search of the metagenomic fragment against the translated sequences in this cluster, the SOrt-ITEMS algorithm (extended LCA algorithm) is used for the final classification. Applications such as the Automated Phylogenomic Inference Pipeline for bacterial sequences (AMPHORA), MLTreeMap and the Statistical Assignment Package (SAP) have been classified as similarity search methods in some reviews because they use alignment algorithms like BLAST or HMM. However, they also build evolutionary models for specific gene loci, and as such they might also be considered as phylogenetic analysis methods. Bazinet and Cummings [104] have carried out a comparative evaluation of sequence classification programs. In practice, scientists need to balance the trade-off between assignment accuracy and resource requirements when making a selection on which tool(s) to use.

Computational issues with BLAST

The large number of short sequences obtained from metagenomic samples has led to an exponential growth of data in public reference databases such as the NCBI nt or nr databases, and sequence similarity searches have become the bottleneck of metagenomic data analysis. To complete an analysis in a timely manner, scientists need to choose a similarity search tool. Although there are some alternative tools available, BLAST [99] (BLAST+) programs still remain the most widely used tools. BLAST is the most validated method of matching query and database sequences, and it remains the benchmark tool to evaluate the completeness and correctness of other alternatives. Further, although other alternatives offer greater speeds of execution, appropriate analysis pipelines still need to be integrated into routine environmental biomonitoring. For these reasons, it is important to outline the computational issues with BLAST. 'Which BLAST program do I use?' and 'which platform is going to be used to run BLAST?' are two important questions to be answered. BLASTP will match proteins searched against a protein database but will miss incorrectly translated query sequences and also sequences in poorly annotated reference genomes. BLASTX translates queries into all six reading frames, and in so doing increases sensitivity. However, this also increases computation time of the homology search. Similarly, TBLASTX, which translates both the query and reference sequences into all reading frames, is ideal in terms of sensitivity. However, in practice, it is not feasible to use in metagenomic studies because of its long run time and because of the need for high end computing resources. BLASTN is the least computationally intensive BLAST program because it compares nucleotide against nucleotide databases (such as the NCBI nt database). However, it is less sensitive than BLAST protein searches when database coverage is poor and less valuable than BLASTX where gene functional analysis is an important part of the metagenomic project.

Running parallel BLAST is not a trivial task, although BLAST for multiple queries is an 'embarrassingly parallel problem' [105], which requires no or little effort to separate the problem into a number of parallel tasks. The local NCBI BLAST+ [106] algorithm is a multithreading version of BLAST, which could take advantage of modern multi-core desktop computers. It has three steps: word matching, ungapped alignment and gapped alignment. Only the first step has been implemented using multithreading; therefore, in practice, NCBI BLAST+ does not significantly improve the speed of searching large databases and reduce runtime. In mpiBLAST [107], a reference database can be fragmented as well as the input queries. However, an out-of-dated version of BLAST (NCBI BLAST) is used in mpiBLAST, and there have been no updates in the development of mpiBLAST since April 2010 based on the information on the mpiBLAST website.

In addition to the effort on parallelizing BLAST software and reference databases, specific hardware is required to run parallel versions of BLAST. Accessing a high performance computing (HPC) facilities poses another barrier to scientists in terms of cost and usability. In most HPC facilities, the computing resources are managed by job scheduling tools, and researchers need to have some understanding of an HPC environment to run BLAST analyses.

Alternatively, graphic processing unit (GPU) BLAST has been implemented to take advantage of GPU parallelism using Compute Unified Device Architecture (CUDA) framework [108, 109]. However, this development is still in the stage of proof-of-concept and only BLASTP has been implemented.

Alternative similarity search algorithms

Far less sensitive than BLAST, but faster is the similarity search algorithm of BLAT [110]. The solution for attaining faster speed in BLAT is to index reference genomes using non-overlapping k-mers and to save the index in memory. However, the fast speed obtained using non-overlapping k-mers also means that sensitivity is sacrificed. In the context of metagenomic analyse, MG-RAST [55] and MetaBin [83] use BLAT as their homology search tool (Table 1).

Two new BLASTX alternatives: Reduced Alphabet based Protein similarity Search (RAPSearch) [111, 112] and Protein Alignment Using a DNA Aligner (PAUDA) [113] are also important to mention. A reduced protein alphabet idea is used in both approaches. In addition, PAUDA uses BOWTIE2 [114] as its mapping engine. RAPSearch2 and PAUDA reportedly run up to 100× and 10,000× faster than BLASTX, respectively. Within an environmental biomonitoring setting, e.g. in responding to an infectious disease outbreak, methods with fast execution times will be needed for obtaining timely results. A concern with PAUDA might be loss of sensitivity. However, the authors report identification of similar orthology groups at all taxonomic hierarchical levels in empirical analyses they have undertaken [113], suggesting that although more reads are likely to be unassigned with PAUDA, the method is nevertheless suitable for measuring changes in the relative abundance of species.

In parallel with the development of BLAST-like search tools, there are other similarity approaches that include application of the Smith–Waterman algorithm [115]. For this, research has leveraged multi-processors in a GPU. The Smith–Waterman algorithm involves a dynamic programming search strategy, which explores all possible alignments between two sequences and then produces an optimal local alignment. It is computationally unrealistic to obtain optimal local assignments with traditional implementations of the Smith–Waterman algorithm, given the exponential growth of protein and DNA databases and with the size of metagenomic data sets. For this reason, researchers have been making use of CUDA-enabled applications in metagenomic studies [109, 116–118]. However, GPU implementations require special hardware and software environments, and for this reason, they are not yet accessible to most scientists in the metagenomics field. The implementation, without or with few modifications, of accelerated Smith-Waterman search strategies into existing GPU/CUDA pipelines that currently use BLAST is a subject of considerable interest. Although the runtime of GPU Smith–Waterman has been reduced significantly, there are still no reported performance evaluations for rendering GPU Smith–Waterman output for taxonomic classification. In other words, the downstream analysis has not yet been evaluated.

In most cases, to run the data analysis process seamlessly, some programming skills are required to parse or alter the input or output files such as format, structure and layout. The amount of time and effort required to make the pipeline flexible are substantial and sometimes not achievable under some time and budget constraints. For this reason, some of the BLAST-like tools and GPU Smith–Waterman are still not a favoured option after considering other factors involved a metagenomic project. This scenario equally applies to other cases where a new bioinformatics tool is introduced into a metagenomic analysis pipeline.

DATABASE ISSUES

Schnoes *et al*. [119] has discussed reference database errors and their impact on the interpretation of BLAST results. As aforementioned, we (and others) have observed a strong presence of model organisms in MEGAN outputs, which reflects the problem of taxonomic bias in the database. Further, there is also evidence of misclassification of entries in the NCBI GenBank database [e.g. 66]. Such errors are propagated into BLAST outputs, which then results in incorrect taxonomic assignment. Another database issue concerns the GI-TID dump file from the NCBI Taxonomy database. This file (GI-TID) is used by some applications as a look up file for parsing sequence search results, i.e. assigning the BLAST output to the NCBI taxonomy. An assumption for the dump file is that one GI entry in the NCBI has one and only one associated taxonomy ID, and one taxonomy ID can be assigned to multiple GIs. In reality, for some entries, there are two associated TIDs. For example, the entry GI 29028372 has two taxon IDs. Taxon ID 10679 is for a virus and the other one, 562, is for its host's taxon ID. Errors such as this diminish database integrity.

Key points

- High-throughput sequencing overcomes (i) selection biases of traditional culture-based method in environmental profiling, (ii) identify minor but ecological significant microbes in the environment and (iii) make amenable for study other complex biotic ecosystems. However, there are many points to consider for its effective implementation.
- The length of high-throughput sequencing reads, the limitation of reference databases, sequencing errors and taxonomic assignment strategies are some of the possible factors that contribute to the high percentage of short reads from metagenomic projects that cannot be matched with their target organisms.
- In most current metagenomic projects, unassigned sequences are omitted from data analysis. So far, there are no easy answers to questions like 'do these sequences result from errors?' or 'have we found something new?' and 'how can we confirm our findings?'
- *In vitro*-simulated communities play an important role in the investigation of different sequencing and data analysis techniques. They provide a valuable experimental approach for testing limitations of methodology and for evaluating the potential of metagenomics in studies where the relative abundance of organisms and their biomass is assessed.
- During the past few years, the development of high-throughput and low-cost sequencing technologies has been faster than the speed of data analysis. Data storage, transfer and sharing are difficult problems owing to the large volume of sequencing data. Computational resources for processing the data are a bottleneck in most metagenomics projects. Sequence data generation has increased at a greater rate than Moore's Law. Alternatives to BLAST that speed up metagenomics analyses, which will facilitate rapid reporting of bio-monitoring results, are active area of much research interest.
- Multiple bioinformatics tools and applications are used for data analysis process. Integration of these tools with computational resources into pipelines is an important direction for research, requiring collaboration between biologists and computer specialists. Standardized pipelines will improve data analysis productivity and provide a solid foundation for future comparative metagenomic studies.

FUNDING

P.J.L. and L.H.W. thank UNITEC and Massey University for financial support. B.J.P. is the recipient of an Allan Wilson Centre PhD studentship. V.V.N. thanks the University of the South Pacific and the Oceania Development Network for financial support of his PhD studies. P.J.B. was partly funded by New Zealand Marsden Fund project 08-MAU-099 (Cows, starlings and Campylobacter in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution). J.W. thanks the Crown Research Institutes (CRIs) Capability Fund (CF), Ministry of Business, Innovation and Employment that supports her position at ESR.

References

1. Thomas T, Gilbert J, Meyer F. Metagenomics—a guide from sampling to data analysis. *Microb Inf Exp* 12012;**2**:3.
2. Gonzalez A, Jose C, Shade A., *et al*. Our microbial selves: what ecology can teach us. *EMBO Rep* 2011;**12**: 775–84.
3. Garcia GD, Gregoracci GB, de O, Santos E, *et al*. Metagenomic analysis of healthy and white plague-affected mussismilia braziliensis corals. *Microb Ecol* 2013;**65**:1076–86.
4. Carter MQ, Xue K, Brandl MT, *et al*. Functional metagenomics of Escherichia coli O157:H7 interactions with spinach indigenous microorganisms during biofilm formation. *PloS One* 2012;**7**: e44186.
5. Knight R, Jansson J, Field D, *et al*. Unlocking the potential of metagenomics through replicated experimental design. *Nat Biotechnol* 2012;**30**:513–20.
6. Kennedy J, Flemer B, Jackson SA, *et al*. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Mar Drugs* 2010;**8**: 608–28.
7. Calvignac-Spencer S, Merkel K, Kutzner N, *et al*. Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. *Mol Eol* 2013;**22**:915–24.
8. Woyke T, Teeling H, Ivanova NN, *et al*. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 2006;**443**:950–5.
9. Sloan DB, Moran N. Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biol Lett* 2012;**8**:986–9.
10. Vorsino AE, Wieczorek AM, Wright MG, *et al*. Using evolutionary tools to facilitate the prediction and prevention of host-based differentiation in biological control: a review and perspective. *Ann Appl Biol* 2012;**160**:204–16.
11. Jurado-Rivera JA, Vogler AP, Reid CAM, *et al*. DNA barcoding insect-host plant associations. *Proc. Biol Sci R Soc* 2009;**276**:639–48.
12. Holland JM, Smith BM, Birkett TC, *et al*. Farmland bird invertebrate food provision in arable crops. *Ann Appl Biol* 2012;**160**:66–75.
13. Lundgren JG, Ellsbury ME, Prischmann DA. Analysis of the predator community of a subterranean herbivorous insect based on polymerase chain reaction. *Ecol Appl* 2009;**19**: 2157–66.
14. Ellwood MDF, Foster WA. Doubling the estimate of invertebrate biomass in a rainforest canopy. *Nature* 2004;**429**: 549–51.
15. Hoffmann AA, Sgrò CM. Climate change and evolutionary adaptation. *Nature* 2011;**470**:479–85.
16. Marx CJ. Can you sequence ecology? Metagenomics of adaptive diversification. *PLoS Biol* 2013;**11**:e1001487.
17. Manly BFJ. *The Design and Analysis of Research Studies*. Cambridge: Cambridge University Press, 1992;353.
18. Southwood T, Henderson PA. *Ecological Methods*. 3rd edn. Oxford: Blackwell Science, 2000;75.
19. Prosser JI. Replicate or lie. *Environ Microb* 2010;**12**:1806–10.
20. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 2005;**6**:805–14.
21. Delmont TO, Robe P, Cecillon S, *et al*. Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* 2011;**77**:1315–24.
22. Lombard N, Prestat E, Van Elsas JD, *et al*. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol Ecol* 2011;**78**:31–49.
23. Zarraonaindia I, Smith DP, Gilbert JA. Beyond the genome: community-level analysis of the microbial world. *Biol Philos* 2013;**28**: 261–82.
24. Zambrano L, Contreras V, Mazari-Hiriart M., *et al*. Spatial heterogeneity of water quality in a highly degraded tropical freshwater ecosystem. *Environ Manage* 2009;**43**: 249–63.
25. Ugland KI, Gray JS, Ellingsen KE. The species—accumulation curve and estimation of species richness. *J Anim* 2003;888–97.
26. Thompson GG, Thompson SA, Withers PC, *et al*. Determining adequate trapping effort and species richness using species accumulation curves for environmental impact assessments. *Austral Ecol* 2007;**32**:570–80.
27. Unterseher M, Jumpponen A, Opik M, *et al*. Species abundance distributions and richness estimations in fungal metagenomics—lessons learned from community ecology. *Mol Ecol* 2011;**20**:275–85.
28. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;**6**:e1000667.
29. Dole-Olivier MJ, Castellarini F, Coineau N, *et al*. Towards an optimal sampling strategy to assess groundwater biodiversity: comparison across six European regions. *Freshwater Biol* 2009;**54**:777–96.
30. Hughes JB, Hellmann JJ, Ricketts TH, *et al*. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* 2001;**67**:4399–06.
31. Zinger L, Gobet A, Pommier T. Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* 2012;**21**:1878–96.
32. Whittaker RH. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol Monogr* 1960;**30**:279–338.
33. Wang JF, Jiang CS, Hu MG, *et al*. Design-based spatial sampling: theory and implementation. *Environ Model Softw* 2013;**40**:280–8.
34. Wang JF, Stein A, Gao BB, *et al*. A review of spatial sampling. *Spat Stat* 2012;**2**:1–14.
35. Zhang J, Zhang C. Sampling and sampling strategies for environmental analysis. *Int J Environ Anal Chem* 2012;**92**: 466–78.
36. Smith KL, Jones ML. Allocation of sampling effort to optimize efficiency of watershed-level ichthyofaunal inventories. *Trans Am Fish Soc* 2008;**137**:1500–6.

37. Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PloS One* 2013;**8**:e53608.

38. Dickie IA. Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytol* 2010; **188**:916–8.

39. Krsek M, Wellington EM. Comparison of different methods for the isolation and purification of total community DNA from soil. *J Microbiol Methods* 1999;**39**:1–16.

40. Stach JEM, Bathe S, Clapp JP, *et al*. PCR-SSCP comparison of 16S rDNA sequence diversity in soil DNA obtained using different isolation and purification methods. *FEMS Microbiol Ecol* 2001;**36**:139–51.

41. LaMontagne MG, Michel FC, Holden PA, *et al*. Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. *J Microbiol Methods* 2002;**49**:255–64.

42. Carrigg C, Rice O, Kavanagh S, *et al*. DNA extraction method affects microbial community profiles from soils and sediment. *Appl Microbiol Biotechnol* 2007;**77**:955–64.

43. Hong S, Bunge J, Leslin C, *et al*. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* 2009;**3**:1365–73.

44. Feinstein LM, Sul WJ, Blackwood CB. Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl Environ Microbiol* 2009;**75**:5428–33.

45. Temperton B., Field D, Oliver A, *et al*. Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J* 2009;**3**:792–6.

46. Zhao F, Xu K. Efficiency of DNA extraction methods on the evaluation of soil microeukaryotic diversity. *Acta Ecol Sinica* 2012;**32**:209–14.

47. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an *in vitro*-simulated microbial community. *PLoS One* 2010;**5**:1–10.

48. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 2011;**6**:e27310.

49. Diaz PI, Dupuy AK, Abusleme L, *et al*. Using high throughput sequencing to explore the biodiversity in oral bacterial communities. *Mol Oral Microbiol* 2012;**27**:182–201.

50. Engelbrektson A, Kunin V, Wrighton KC, *et al*. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* 2010;**4**:642–7.

51. Caporaso JG, Lauber CL, Walters WA, *et al*. Ultra-high-throughput microbial community analysis on the Illumina HiSEQ and MiSEQ platforms. *ISME J* 2012;**6**:1621–4.

52. Whiteley AS, Jenkins S, Waite I, *et al*. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *J Microbiol Methods* 2012;**91**: 80–8.

53. Lemos LN, Fulthorpe RR, Triplett EW, *et al*. Rethinking microbial diversity analysis in the high throughput sequencing era. *J Microbiol Mhethods* 2011;**86**:42–51.

54. Shah N, Tang H, Doak TG, *et al*. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac Symp Biocomput* 2011;**16**:165–76.

55. Glass EM, Wilkening J, Wilke A, *et al*. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* 2010. pdb.prot5368.

56. Bowen De León K, Ramsay BD, Fields MW. Quality-score refinement of SSU rRNA gene pyrosequencing differs across gene region for environmental samples. *Microb Ecol* 2012;**64**:499–508.

57. Haas BJ, Gevers DE, Ashlee M, *et al*. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 2011;**21**: 494–504.

58. Quince C, Lanzen A, Davenport RJ, *et al*. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011; **12**:38.

59. Patin NV, Kunin V, Lidström U, *et al*. Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microb Ecol* 2013;**65**:709–19.

60. Lee CK, Herbold CW, Polson SW, *et al*. Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PloS One* 2012;**7**:e44224.

61. Sipos R, Székely AJ, Palatinszky M, *et al*. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* 2007;**60**:341–50.

62. Mao DP, Zhou Q, Chen CY, *et al*. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol* 2012;**12**:66.

63. Ahn JH, Kim BY, Song J, *et al*. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J Microbiol* 2012;**50**:1071–4.

64. Vasileiadis S, Puglisi E, Arena M, *et al*. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLoS One* 2012;**7**:e42671.

65. Huber JA, Morrison HG, Huse SM, *et al*. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ Microbiol* 2009;**11**: 1292–302.

66. Prosdocimi EM, Novati S, Bruno R, *et al*. Errors in ribosomal sequence datasets generated using PCR-coupled 'panbacterial' pyrosequencing, and the establishment of an improved approach. *Mol Cell Probes* 2013;**27**: 65–7.

67. Kembel SW, Wu M, Eisen JA., *et al*. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* 2012;**8**: e1002743.

68. Luo C, Tsementzi D, Kyrpides N, *et al*. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 2012; **7**:e30087.

69. Zhou J, Wu L, Deng Y, *et al*. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* 2011;**5**:1303–13.

70. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011;**7**:1–8.

71. Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010;**11**:485.

72. Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2011;**7**:335–36.

73. Berger SA, Krompass D, Stamatakis A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol* 2011; **60**:291–302.
74. Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**:1641–50.
75. Matsen FA, Kodner RB, Armbrust EV. Pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 2010;**11**:538.
76. DeSantis TZ, Hugenholtz P, Larsen N, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;**72**: 5069–72.
77. Huson DH, Mitra S, Ruscheweyh H-J, *et al*. Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 2011;**21**: 1552–60.
78. Monzoorul HM, Ghosh TS, Komanduri D, *et al*. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 2009;**25**:1722–30.
79. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 2011;**39**:e91.
80. Liu B, Gibbons T, Ghodsi M, *et al*. MetaPhyler: taxonomic profiling for metagenomic sequences. IEEE International Conference. *Bioinform Biomed* 2010;95–100.
81. Mohammed MH, Ghosh TS, Reddy R, *et al*. INDUS: a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics* 2011;**12(Suppl. 3)**:S4.
82. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naïve Bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 2011;**27**: 127–9.
83. Sharma VK, Kumar N, Prakash T, *et al*. Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PLoS One* 2012;**7**:e34030.
84. Diaz NN, Krause L, Goesmann A, *et al*. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* 2009;**10**:56.
85. Yang B, Peng Y, Leung HCM, *et al*. MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. New York, NY: ACM, 2010; 170–9.
86. Wang Y, Leung HCM, Yiu SM, *et al*. MetaCluster 4.0: A novel binning algorithm for NGS reads and huge number of species. *J Comput Biol* 2012;**19**:241–9.
87. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 2009;**6**:673–6.
88. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 2008;**9**:R151.
89. Stark M, Berger SA, Stamatakis A, *et al*. MLTreeMap: accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 2010;**11**:461.
90. Munch K, Boomsma W, Huelsenbeck JP, *et al*. Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst Biol* 2008;**57**:750–7.
91. Mohammed MH, Ghosh TS, Singh NK, *et al*. SPHINX: an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* 2011;**27**:22–30.
92. Singh R, Narayan V, McLenachan PA, *et al*. Detection and diversity of pathogenic *Vibrio* from Fiji. *Environ Microbiol Rep* 2012;**4**:403–11.
93. Pickett SB, Bergey CM, Di Fiore AA. Metagenomic study of Primate insect diet diversity. *Am J Primatol* 2012;**10**: 622–31.
94. Xiang Q, Soltis D, Soltis P. Phylogenetic relationships of Cornaceae and close relatives inferred from matK and rbcL sequences. *Am J Bot* 1998;**85**:285.
95. Zhao X, Wang Q, Zhou J, Zhong Y. Phylogenetic relationships of taxaceae and its related groups inferred from nuclear 18S rRNA, chloroplast matK, rbcL, rps4, 16S rRNA and mitochondrial coxI sequences. *J Genet Mol Biol* 2006;**17**:81–93.
96. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. *PloS One* 2011;**6**:e19254.
97. Raupach MJ, Astrin JJ, Hannig K, *et al*. Molecular species identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Front Zool* 2010;**7**:e26.
98. Edwards RA, Rohwer F. Opinion: viral metagenomics. *Nat Rev Microbiol* 2005;**3**:504–10.
99. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
100. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
101. Huson DH, Auch AF, Qi J, *et al*. MEGAN analysis of metagenomic data. *Genome Res* 2007;**17**:377–86.
102. Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol* 2012;**856**:415–29.
103. Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* 2010;**11**:544.
104. Bazinet AL, Cummings MPA. comparative evaluation of sequence classification programs. *BMC Bioinformatics* 2012; **13**:92.
105. Lin H, Ma XMX, Chandramohan P, *et al*. Efficient data access for parallel BLAST. *19th IEEE International Parallel and Distributed Processing Symposium*. Washington, DC, USA: IEEE Computer Society, 2005;72b.
106. Camacho C., Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**: 421.
107. Darling AE, Carey L, Feng W. The design, implementation, and evaluation of mpiBLAST. *Proc ClusterWorld* 2003; 13–5.
108. Vouzis PD, Sahinidis NV. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics* 2011;**27**:182–8.
109. Ling C, Benkrid K. Design and implementation of a CUDA-compatible GPU-based core for gapped BLAST algorithm. *Proc Comput Sci* 2010;**1**:495–504.
110. Kent WJ. BLAT: the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.
111. Ye Y, Choi JH, Tang H. RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics* 2011; **12**:159.

112. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 2012;**28**:125–6.

113. Huson D, Chao X. A poor man's BLASTX—high throughput metagenomic protein database search using PAUDA. *Bioinformatics* 2013. doi: 10.1093/bioinformatics/btt254 (Advance Access publication 7 May 2013).

114. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–60.

115. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.

116. Ligowski L, Rudnicki W. An efficient implementation of Smith Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases. *Processing* 2009;1–8.

117. Liu Y, Schmidt B, Maskell DL. CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions. *BMC Res Notes* 2010;**3**:93.

118. Nguyen DH, Pham PH. Applying GPUs for Smith-Waterman sequence alignment acceleration. *GSTF J Comput* 2012. doi: 10.5176_2010–2283_1.2.56.

119. Schnoes AM, Brown SD, Dodevski I., *et al*. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 2009; **5**:13.