MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 1794–1805 doi: 10.1111/j.1365-294X.2012.05538.x

INVITED REVIEWS AND META-ANALYSES

Next-generation sequencing technologies for environmental DNA research

SHADI SHOKRALLA, JENNIFER L. SPALL, JOEL F. GIBSON and MEHRDAD HAJIBABAEI

Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

Abstract

Since 2005, advances in next-generation sequencing technologies have revolutionized biological science. The analysis of environmental DNA through the use of specific gene markers such as species-specific DNA barcodes has been a key application of next-generation sequencing technologies in ecological and environmental research. Access to parallel, massive amounts of sequencing data, as well as subsequent improvements in read length and throughput of different sequencing platforms, is leading to a better representation of sample diversity at a reasonable cost. New technologies are being developed rapidly and have the potential to dramatically accelerate ecological and environmental research. The fast pace of development and improvements in next-generation sequencing technologies can reflect on broader and more robust applications in environmental DNA research. Here, we review the advantages and limitations of current next-generation sequencing technologies in regard to their application for environmental DNA analysis.

Keywords: environmental DNA, environmental monitoring, metagenomics, next-generation sequencing.

Received 15 October 2011; revision received 11 February 2012; accepted 14 February 2012

Introduction

Genomic analysis of complex environmental samples is becoming an important tool for understanding evolutionary history and functional and ecological biodiversity. It bypasses the need for laboratory cultivation and/or isolation of individual specimens. Examples of typical environmental samples include soil, water, sediments, passively collected aquatic, terrestrial, and benthic specimens, gut contents and faeces.

Advances in conventional Sanger DNA-sequencing technology led to large-scale, broad-scope biosystematics projects with a wide range of applications (e.g. the Barcode of Life initiative; Hajibabaei *et al.* 2007). DNA barcoding employs standardized species-specific genomic regions (DNA barcodes) to generate vast DNA libraries for the primary purpose of identification of unknown specimens. The cytochrome *c* oxidase subunit I (COI) gene region, for example, is capable of discerning between closely related species across Animalia (Hebert *et al.* 2003; Meusnier *et al.* 2008). Similarly, 16S ribosomal RNA (16S) is commonly used for bacterial identification (Sogin *et al.* 2006; Flanagan *et al.* 2007). The internal transcribed spacer (ITS) region of the nuclear ribosomal DNA is employed in studies of fungi (Nilsson *et al.* 2008). Plant DNA barcoding has relied heavily upon regions of plastid DNA including maturase K (*matK*) and ribulose-bisphosphate carboxylase (*rbcl*) (CBOL Plant Working Group 2009; Burgess *et al.* 2011). A number of other marker genes have been employed for biodiversity analysis at different phylogenetic depths or in certain taxonomic groups.

The traditional DNA-sequencing method (Sanger *et al.* 1977) can only sequence specimens individually and, therefore, is inadequate for processing complex environmental samples, especially for large-scale studies. These samples often contain mixtures of DNA from hundreds or thousands of individuals. Although conventional sequencing has provided the most efficient method for the development of large DNA barcode reference libraries, the number of individuals in an environmental sample is beyond the scope of its ability (Hajibabaei *et al.* 2011). Recovering DNA sequences

Correspondence: Shadi Shokralla, Fax: 1519 824 5703;
E-mail: sshokral@uoguelph.ca

from the thousands of specimens present in an environmental bulk sample requires the ability to read DNA from multiple templates in parallel; something that next-generation sequencing technologies do effectively, and with ever-lowering costs.

Next-generation sequencing (NGS) platforms have made it possible to recover DNA sequence data directly from environmental samples (e.g. Sogin *et al.* 2006). These data have been used in a variety of applications, including comparing microbiota in healthy versus diseased individuals (e.g. Andersson *et al.* 2008; Zhang *et al.* 2009); inferring the health of an ecosystem by analysing its biodiversity (Hajibabaei *et al.* 2011); studying ancient DNA (Haile *et al.* 2009; Sønstebø *et al.* 2010; Boessenkool *et al.* 2011); and diet analysis from DNA fragments in faeces or gut contents (Deagle *et al.* 2009). By comparing obtained sequences to a growing standard reference library of known organisms, taxa present in an environmental sample can be identified with high confidence. Advanced computational methods have made it possible to infer biodiversity measures across time and space by annotating and clustering DNA sequences using a combination of assignment and phylogenetic techniques (Hajibabaei *et al.* 2011). A recent explosion in both number and breadth of studies employing NGS platforms demonstrates a paradigm shift in ecological research towards the use of high volumes of sequence data.

Articles published in this special issue are good examples of situations in which NGS analysis provides an effective means of addressing difficult questions in ecology by targeting DNA in environmental samples. The revolution in NGS technologies is also reflected in several competing sequencing systems and their rapid advancement (Fig. 1). It is often difficult for users to determine the appropriate NGS platform for their ecological research. Our review presents the existing NGS platforms along with a comparison of the benefits and limitations of each system with regard to environmental DNA research.

History and advances of next-generation sequencing technologies

The conventional DNA-sequencing approach was introduced by Sanger *et al.* (1977) and is capable of recovering up to 1 kb of sequence data from a single specimen at a time. The most advanced version of automated Sanger sequencers is capable of sequencing up to 1 kb for 96 individual specimens at a time. In the last few years, a series of high-throughput sequencing devices have been commercially introduced based on different chemistries and detection techniques. These NGS technologies can potentially generate several hundred thousand to tens of millions of sequencing reads in parallel. This massively parallel throughput sequencing capacity can generate sequence reads from fragmented libraries of a specific genome (i.e. genome sequencing); from a pool of cDNA library fragments generated through reverse transcription of RNA molecules (i.e. RNAseq or transcriptome sequencing); or from a pool of PCR-amplified molecules (i.e. amplicon sequencing). In all cases, sequences are generated without the need of a conventional, vector-based cloning procedure that is typically used to amplify and separate DNA templates. As such, some of the cloning bias issues that impact sequencing evenness in sequencing projects may be

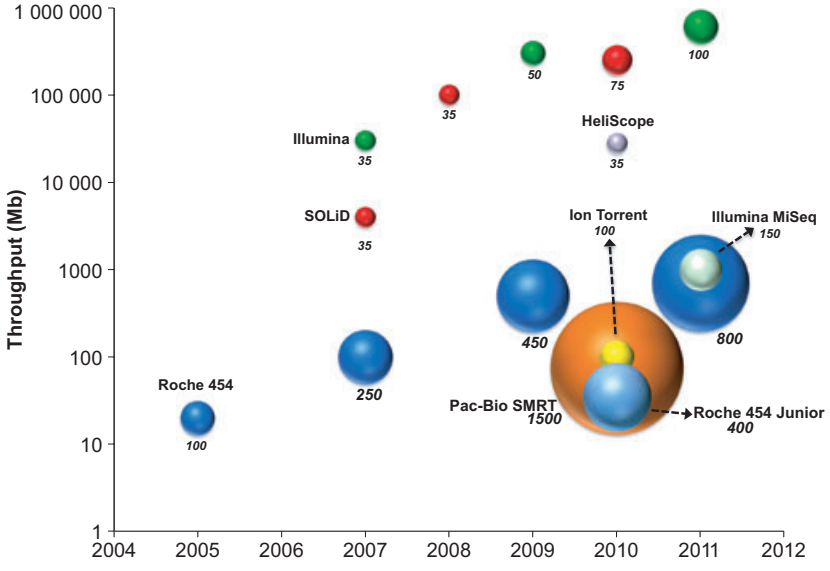

Fig. 1 Historical development of next-generation sequencing technologies. The diameter of each bubble represents the sequencing read length of the platform [in base pairs (bp)]. Colours correspond to individual platforms.

avoided, although each NGS platform may have its own associated limitations (Mardis 2008a).

Since their introduction in 2005, high-throughput NGS technologies have faced several general challenges. The first has been the improvement in sequencing output, in terms of read length and accuracy. The second challenge has been the total output of the sequencing experiment in relation to the cost and the labour expended. The third challenge is related to the amplification step prior to sequencing. This final challenge includes different sources of PCR bias, formation of chimeric sequences and secondary structure-related issues (Mardis 2008a; Shendure & Ji 2008). New technologies promise to fundamentally change the nature of genomics-based studies, especially when coupled with the computational algorithms necessary to analyse their vast sequencing output (Mardis 2008b). A timeline comparison of different high-throughput platforms in regard to read length and total output illuminates the rapid progress in sequencing capabilities of NGS machines (Fig. 1). In fact, sequencing technology's rapid progress has now out-paced computational processing as predicted by Moore's law (http://www.genome.gov/sequencingcosts/).

Although available NGS technologies utilize quite diverse chemistry and base incorporation/detection tools, they share two main steps: library fragmentation/amplicon library preparation and detection of the incorporated nucleotides (Glenn 2011; Zhang *et al.* 2011). NGS technologies can be classified into two main categories. The first group are PCR-based technologies, which currently include four commercially available platforms: Roche 454 Genome Sequencer (Roche Diagnostics Corp., Branford, CT, USA), HiSeq 2000 (Illumina Inc., San Diego, CA, USA), AB SOLiD™ System (Life Technologies Corp., Carlsbad, CA, USA) and Ion Personal Genome Machine (Life Technologies, South San Francisco, CA, USA). The other group, called 'single-molecule' sequencing (SMS) technologies, are non-PCR based and do not include an amplification step prior to sequencing. Two single-molecule sequencing systems have been announced recently: HeliScope (Helicos BioSciences Corp., Cambridge, MA, USA) and PacBio RS SMRT system (Pacific Biosciences, Menlo Park, CA, USA). Here, we briefly describe available NGS platforms in each category.

PCR-based next-generation DNA-sequencing technologies

Roche 454 genome sequencers

Introduced in 2005, the 454 Genome Sequencer was the first NGS technology to become commercially available. It uses real-time sequencing-by-synthesis pyrosequencing technology. In 454 pyrosequencing, each nucleotide incorporated by DNA polymerase results in the release of a pyrophosphate molecule. This release initiates a series of downstream reactions to produce light by the action of the enzyme luciferase. The amount of generated light is directly proportional to the number of nucleotides incorporated (Margulies *et al.* 2005). The 454 pyrosequencing workflow includes the immobilization of the library fragments on either sepharose or styrofoam beads whose surfaces carry oligonucleotides complementary to the 454-specific adapter sequences ligated or PCR-generated onto both ends of the fragmented library. The library fragments are amplified through emulsion PCR thermal cycling into individual water:oil micro-reactors that contain PCR ingredients. Each library fragment is amplified on the surface of one bead in a single micro-reactor, generating billions of copies of the same fragment covering the surface of the DNA bead. The amplified beads are recovered from emulsion oil followed by an enrichment step to keep only the amplified beads and discard the failed ones. The enriched beads are then prepared as single-stranded and annealed to a specific sequencing primer. These beads are then arrayed into a picotiter plate (PTP) that is designed to have more than one million wells per plate. Each of the wells can hold only one amplified DNA bead. Four layers of engineered beads are deposited into the PTP. From bottom to top, diluted pyrosequencing enzyme beads, DNA amplified beads, pyrosequencing enzyme beads and, finally, PPiase beads. All bead layers are deposited by centrifugation. The PTP is then sequenced en masse in the 454 GS pyrosequencing instrument. The sequencing steps include the flow of repetitive cycles of nucleotide solutions (T, C, A and G). The PTP is seated opposite a CCD camera that records the emitted light from each bead with the flow of different nucleotide solutions. The current GS FLX+ system provides 200 nucleotide flow cycles to generate up to 800-bp sequencing reads. The generated raw signals are processed by 454 pyrosequencing analysis software and then screened by various quality filters to remove poor-quality sequences (Mardis 2008a). Roche 454 genome sequencers are available in two versions: a GS FLX+ system (1 M sequence read capacity) and the recently introduced the GS junior system (100 k sequence read capacity) (Fig. 1; Table 1).

Illumina sequencers

Illumina (formerly known as Solexa) introduced the genome analyzer in 2007. Due to its high capacity, it soon became a workhorse for whole-genome resequencing applications, including human and model organ-

Table 1 Comparison of currently available next-generation sequencing technologies

Category	Platform	Read length (bp)	Max. number of reads/run	Sequencing output/run	Run time
PCR-based NGS technologies	Roche 454 GS FLX	400–500	1×10^6	≤500 Mb	10 h
	Roche 454 GS FLX+	600–800	1×10^6	≤700 Mb	23 h
	Roche 454 GS Junior	400–450	1×10^5	~35 Mb	10 h
	Illumina HiSeq 2000	100–200	6×10^9	≤540–600 Gb	11 d
	Illumina HiSeq 1000	100–200	3×10^9	≤270–300 Gb	8.5 d
	Illumina GAIIx	50–75	6.4×10^8	≤95 Gb	7.5–14.5 d
	Illumina MiSeq	100–150	7×10^6	≤1–2 Gb	19–27 h
	AB SOLiD 5500 system	35–75	2.4×10^9	~100 Gb	4 d
	AB SOLiD 5500 xl system	35–75	6×10^9	~250 Gb	7–8 d
	Ion Torrent -314 chip	100–200	1×10^6	≥10 Mb	3.5 h
	Ion Torrent -316 chip	100–200	6×10^6	≥100 Mb	4.7 h
	Ion Torrent -318 chip	100–200	11×10^6	≥1 Gb	5.5 h
SMS technologies	Helicos HeliScope	30–35	1×10^9	~20–28 Gb	≤1 d
	Pacific Biosciences system	≥1500	50×10^3	~60–75 Mb	0.5 h

ism genomic projects. The Illumina platform utilizes a sequencing-by-synthesis approach coupled with bridge amplification on the surface of a flow cell. Each flow cell is divided into eight separate lanes. The interior surfaces of the flow cells have covalently attached oligos complementary to specific adapters that are ligated onto the library fragments. DNA fragment-to-oligo hybridization on the flow cell occurs by active heating and cooling steps. This is followed by a subsequent incubation with the amplification reactants and an isothermal polymerase that generates millions of clusters of the library fragments. In the sequencing step, each cluster is supplied with polymerase and four differentially labelled fluorescent nucleotides that have their 3′-OH chemically inactivated. This blocking modification ensures that only a single base will be incorporated per flow cycle. After each nucleotide is incorporated, an excitation followed by an imaging step takes place to identify the incorporated nucleotide in each cluster. A chemical deblocking treatment removes the fluorescent group and allows the incorporation of the following nucleotide during the next flow cycle. The sequence of each cluster is computed and subjected to quality filtering to eliminate low-quality reads (Shendure & Ji 2008). Today, four versions of Illumina sequencers are commercially available. The HiSeq 2000, HiSeq 1000 and Genome Analyzer IIx have sequencing outputs of up to 600, 300 and 95 Gb, respectively. The recently introduced MiSeq platform can generate up to 150-bp sequencing reads with a total throughput of 1.5–2 Gb per run (Fig. 1; Table 1). In 2012, Illumina introduced HiSeq2500 platform as an upgrade of HiSeq2000. This new platform can generate up to 120 Gb of data in 27 h, enabling researchers to sequence an entire genome in 24 h (i.e. 'Genome in a Day'). Illumina has announced that this platform will be commercially available in the second half of 2012.

Applied Biosystems SOLiD sequencer (Life Technologies)

In 2007, Applied Biosystems (Life Technologies) introduced SOLiD technology as their NGS platform. Unlike the previous two platforms, SOLiD uses a sequencing-by-oligo ligation technology. This process couples oligo adaptor-linked DNA fragments with complementary oligos immobilized on the surface of 1-mm magnetic beads. The beads are individually amplified through emulsion PCR. After amplification, the beads are covalently attached to the surface of a specially treated glass slide that is placed into a fluidics cassette within the sequencer. The ligation-based sequencing process starts with the annealing of a universal sequencing primer that is complementary to the SOLiD-specific adapters ligated to the library fragments. Four semi-degenerate 8-mer fluorescent oligos are added along with DNA ligase in an automated manner within the instrument. When a matching 8-mer oligo hybridizes to the DNA fragment sequence adjacent to the universal primer at the 3′-end, DNA-ligase seals the phosphate backbone. After the ligation step, a fluorescent readout identifies the ligated 8-mer oligo, which corresponds to one of the four possible bases. A subsequent chemical cleavage of the linkage between the fifth and sixth bases of the 8mer oligo takes place, removing the fluorescent group and enabling a further ligation round. The second sequencing round is initiated with hybridization of an *n*-1 positioned universal primer, and subsequent rounds

of ligation-mediated sequencing. The same process is repeated with *n*-2, *n*-3 and *n*-4 positioned universal primers. The generated fluorescence from the five universal primers is decoded with a two-base calling processing software. In the SOLiD system, two slides can be processed in a single run; one slide receives sequencing reagents while the second slide is being imaged (Mardis 2008b). Today, Applied Biosystems SOLiD sequencers are available in two versions, the 5500 system and the 5500xl system, with up to 100- and 250-Gb sequencing capacity, respectively (Fig. 1; Table 1).

Life Technologies Ion Torrent

In 2010, Life Technologies introduced the Ion Personal Genome Machine (PGM) as a postlight sequencing technology. This system relies on the real-time detection of hydrogen ion concentration, released as a by-product when a nucleotide is incorporated into a strand of DNA by the polymerase action. Ion Torrent uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a single DNA template from the library. Beneath the wells are an ion-sensitive layer and a proprietary ion sensor to detect the change in hydrogen ion concentration because of nucleotides incorporation (Miller *et al.* 2011; Rothberg *et al.* 2011). The Ion Torrent platform can utilize one of the three available ion chips: 314; 316; or 318, which can generate up to 10 Mb, 100 Mb or 1 Gb, respectively, according to the required sequencing coverage (Fig. 1; Table 1). In 2012, Life Technologies introduced a new generation of Ion semiconductor sequencers; the Ion Proton bench top sequencer. Ion Proton chips will deliver the human genome or human exome in just a few hours. Ion Proton chips will be available in two versions: Ion Proton I chip with 165 million wells (about 100-fold more than the Ion 314 chip); and Ion Proton II chip with 660 million wells (about 1000-fold more than the Ion 314 chip). Both of these chips use CMOS semiconductor chip technology to capture chemistry changes instead of light and translate these changes into digital data. Life Technologies has announced that this platform will be commercially available soon.

Single-molecule DNA-sequencing technologies

Helicos biosciences HeliScope

HeliScope, introduced in 2008, was the first commercially available, single-molecule sequencing (SMS) system. It utilizes sequencing-by-synthesis on a single DNA molecule (Harris *et al.* 2007; Pushkarev *et al.* 2009). The library construction step depends on the preparation of single-stranded DNA fragments. No amplification step is required. During the sequencing cycles, repetitive cycles of the DNA polymerase and one of the four fluorescently labelled nucleotides are flowed in, resulting in template-dependent extension of DNA strands according to the flowed nucleotide. The fluorescent nucleotides are modified to stop the polymerase extension until the incorporated nucleotide's fluorescence is captured and the images are recorded with a highly sensitive CCD camera connected to a fluorescent microscope. A washing step then takes place to wash off the unincorporated nucleotides as well as the by-products of the previous cycle. After washing, fluorescent labels on the extended strands are chemically cleaved and removed. Another cycle of single-base extension, label-cleaving and imaging follows (Ewing *et al.* 1998; Harris *et al.* 2008; Zhang *et al.* 2011). HeliScope is capable of producing approximately 1 billion sequence reads (Fig. 1; Table 1).

Pacific Biosciences SMRT DNA sequencer

The single-molecule real-time (SMRT) DNA-sequencing platform, introduced in 2010 by Pacific Biosciences, is another example of a real-time, fluorescence-based, SMS platform (Korlach *et al.* 2010). It requires no amplification step for sample preparation as it involves a single-molecule sequencing-by-synthesis approach. This platform utilizes a nano-structure called a Zero Mode Waveguide (ZMW) for real-time observation of DNA polymerization (Levene *et al.* 2003; Korlach *et al.* 2008). It consists of tens of thousands of subwavelength, ten nanometre diameter holes, fabricated by perforating a thin metal film supported by a transparent substrate. During the sequencing workflow, the complimentary DNA strand is synthesized from the single-stranded template by the action of DNA polymerase planted at the bottom of each waveguide. Four different-coloured phosphor-linked nucleotides are utilized in this platform. Unlike other technologies, the fluorescence label is attached on the terminal phosphate group rather than the nucleotide base, leading to the release of the fluorescence moiety with the nucleotide incorporation (Pushpendra 2008; Flusberg *et al.* 2010). This approach does not need a washing step between each nucleotide flow, thus accelerating the speed of nucleotide incorporation as well as improving sequence quality. This technology uses the natural capacity of DNA polymerase to incorporate ten or more nucleotides per second in several thousand parallel ZMWs (Eid *et al.* 2009; Zhou *et al.* 2010).

The pros and cons of NGS platforms

The major advantages of the 454 pyrosequencing platform are its long read length and its relatively short run time. Also, unlike other PCR-based NGS technologies, 454 pyrosequencing does not need to carry out an extra chemical deblocking step to allow DNA extension by the action of the DNA polymerase. This reduces the chances of premature chain termination and nonsimultaneous extension, both of which are causes of dephasing (Metzker 2010; Zhou *et al.* 2010). Longer sequences generated through 454 provide higher flexibility in terms of accurate annotation of reads in ecological applications involving nonmodel organisms. This has made 454 the most commonly used NGS platform for the analysis of environmental DNA for ecological applications. Reading of homopolymer regions, however, is challenging for 454 pyrosequencing because of a lack of terminating moiety to stop the extension run. The most common error type on this platform is insertion–deletion rather than substitution. This issue has caused concerns for scientists employing this platform in the analysis of environmental DNA because sequence errors may be interpreted as unique haplotypes representing rare biota (Sogin *et al.* 2006). However, this problem has been largely alleviated using computational tools to distinguish and filter out erroneous sequences (Quince *et al.* 2009). Another drawback for 454 is its relatively high cost for reagents per megabase sequencing output (Claesson *et al.* 2010).

The major advantage of both Illumina and SOLiD systems is that nucleotide detection is performed one at a time. Therefore, homopolymer regions can be accurately sequenced. The chemical deblocking step is performed prior to the next nucleotide incorporation in the Illumina system and prior to further ligation in the SOLiD system. Another advantage of both Illumina and SOLiD systems is the high output per run compared to 454 pyrosequencing. The main drawback of these systems is the relative short-read length because of optical signal decay and dephasing. This limits the application of these technologies in situations where no reference sequence is available to align, assign and annotate the short sequences generated. Meanwhile, the error rate is accumulative with longer sequencing reads in both systems (Zhou *et al.* 2010).

All PCR-based NGS systems share the common disadvantage of bias introduced during amplification. The amplification bias can affect NGS results in two stages. The first incidence is bias introduced during amplicon library preparation. Even before the invention of NGS technologies, research was being directed towards exploring the potential causes and extent of bias in PCR amplification (Polz & Cavanaugh 1998). Annealing temperature is an important parameter for primer binding; to reduce PCR bias of any primer set, the effect of the annealing temperature should be investigated by denaturing gradient gel electrophoretic analysis. Previous studies have indicated that bias can be reduced at lower temperatures when specific amplification is achieved (Ishii & Fukui 2001). PCR bias, however, is also strongly dependent on the number of replication cycles. In this case, bias can be reduced by keeping the number of cycles low (Suzuki & Giovannoni 1996; Qiu *et al.* 2001). Another means to reduce amplification bias is by setting the fastest ramping rate from the denaturation step to the annealing step using PCR cyclers with a fast ramping rate. This may, however, increase heteroduplex formation when PCR reaches the plateau phase (Kurata *et al.* 2004). In general, many studies have demonstrated that PCR bias can be considerably reduced using high template concentrations, wise primer selection, low cycle number, low annealing temperature and mixed replicate reaction preparations (Polz & Cavanaugh 1998; Lim *et al.* 2010). The second stage at which the amplification bias can be introduced is during library amplification prior to sequencing through either emulsion PCR or bridge PCR. Although this late amplification step is carried out with universal probes and can be considered bias-free amplification, it can exaggerate biased amplification in the original amplicon library preparation (Schuster 2008). Single-molecule, non-PCR sequencing technologies overcome this challenge by eliminating the need for template amplification.

With several NGS platforms available and different workflows required in various applications, it is often difficult to select the optimal platform for a specific workflow and application. Table 2 provides an overview of categories of workflows possible for each commercially available NGS platform.

Enhancements of NGS technologies

As mentioned in the previous section, some limitations of NGS platforms can negatively influence their optimal applicability and uptake in various applications. For example, PCR and other workflow-associated biases can lead to highly skewed sequencing results. Additionally, most NGS workflows are time-consuming, tedious and require highly skilled personnel. Several approaches and technologies are being developed to reduce biases as well as enhance and simplify NGS protocols.

One important group of enhancements involves target selection for NGS. These enhancements, also known as Sequence Capture, promise to both eliminate initial PCR amplification and allow selective analysis of large numbers of target sequences (e.g. exome sequencing).

Table 2 Possible applications of NGS platforms related to required sequencing coverage

	Required sequencing coverage								
	≤1 Gb					>1 Gb			
Category	RF	RJ	IMS	IT	PB	IHS	IMS	ABS	HH
Complete genome shotgun sequencing	+	+	+	+	+	+	−	+	+
Transcriptome sequencing	+	+	+	+	+	+	−	+	+
Short amplicon sequencing (≤200 bp)	+	+	+	+	+	+	−	+	+
Long amplicon sequencing (≥200 bp)	+	+	−	−	+	−	−	−	−
Paired-end sequencing	+	+	−	+	−	+	+	+	−
Multiplexing	+	+	+	+	+	+	−	+	+

NGS platform abbreviations: RF-Roche 454 GSFLX; RJ-Roche 454 Junior; IHS-Illumina HiSeq 2000; IMS-Illumina MiSeq; ABS-AB SOLiD systems; IT-Ion Torrent chips; PB-Pacific Biosciences system; HH-Helicos HeliScope.

Sequence capture involves two hybridization-based methods using oligonucleotide probes, either immobilized to a solid array 'Capture arrays' or in solution 'Baits', to capture the sequencing targets (Tewhey *et al.* 2009; Lee *et al.* 2011). The hybridization probes (60–120 bp) are specifically designed to capture target regions across the genome. Nonspecific hybrids are removed by washing and targeted DNA is then eluted for sequencing. Previous studies have shown that the uniformity and specificity of sequences obtained from a solution capture experiment tend to be slightly higher than that of array capture. In addition, solid array hybridization requires expensive hardware, such as a hybridization station. Although target enrichment sequence capture eliminates the need for an initial PCR step, it is necessary to start library preparation with a relatively large amount of DNA (Mamanova *et al.* 2010). Recently, DNA capture via hybridization has allowed the efficient exploitation of NGS for population genetic analyses of ancient DNA samples (Horn 2012). Examples of available Sequence Capture systems include Roche's NimbleGen, Agilent's SureSelect, RainDance Technologies' RainStorm and Illumina's TruSeq Exome Enrichment system. These tools are currently in use in studies of human and other model organisms, but they show significant potential for applications in environmental DNA research (Adey *et al.* 2010; Barzon *et al.* 2011; Jones *et al.* 2011; Faircloth *et al.* 2012). Other enhancements mainly involve modified sample preparation protocols (e.g. library construction). This category of NGS enhancement includes the system developed by Nextera for rapid and automated library construction for Illumina platforms (Caruccio 2011).

Advances in NGS technologies have provided the ability to read millions of DNA sequences in parallel, making them ideally suited for large-scale biodiversity analyses of environmental samples. Constructing mixtures of tagged or bar-coded DNA templates for sequencing has incredible potential for many applications (Binladen *et al.* 2007). This approach can dramatically accelerate ecological studies by allowing multiplexing of different target gene markers of a single bulk sample or multiplexing of a single marker from multiple samples (Valentini *et al.* 2009; Harris *et al.* 2010; Xu *et al.* 2012). Tags should be designed considering the used sequencing chemistry to reduce the likelihood of ambiguities because of potential sequencing errors. For example, the tags should not start with the same nucleotide as the sequencing chemistry adaptor ends; or end with the same nucleotide as the amplification primer starts. Also, not more than two identical successive nucleotides are allowed within the unique tag (Binladen *et al.* 2007). Tag-encoded amplicon sequencing has been utilized in many studies, for example, analysis of the human gut microbiome (Wu *et al.* 2010) and analysis of the cattle tick bacteriome (Andreotti *et al.* 2011). Although this multiplexing approach opens new avenues for many applications with a reasonable price (Sun *et al.* 2011), it is also important to consider potential biases caused by the addition of multiplexing identifier tags to primer oligos (Berry *et al.* 2011).

Application of NGS for analysing environmental DNA

Mass sequencing of environmental samples has been at the forefront of ecology and biodiversity research in recent years. NGS technologies have facilitated analysis of environmentally derived samples from a variety of ecosystems including freshwater, marine, soil, terrestrial and gut microbiota. The majority of these studies seek to answer the question of what is present in a given environment. Through the use of the massive amounts

of sequence data produced by NGS platforms, researchers have been able to observe the slight changes in community structure that may occur following anthropogenic or natural environmental fluctuations (Leininger *et al.* 2006; Fierer *et al.* 2007). These small alterations, although highly informative of ecosystem health and stability, are not discernible with less-sensitive, traditional, molecular tools such as Sanger sequencing (Sogin *et al.* 2006; Huse *et al.* 2010; Xu *et al.* 2012). Regardless of the ecosystem studied or the specific ecological question asked, the vast majority of studies making use of NGS platforms and environmental samples have employed the 454 pyrosequencing platform mainly because of its longer sequence read lengths. A variety of sample sources and sequence-generation workflows are outlined below.

Several studies have analysed soil bacterial diversity by examining 16S rDNA amplicons (e.g. Roesch *et al.* 2007; Rousk *et al.* 2010; Nacke *et al.* 2011). Results suggest that agricultural management of soil may significantly influence the diversity of bacteria and archaea (Roesch *et al.* 2007). Other studies have focused on soil fungal diversity in both forest and agricultural settings by analysing ITS amplicons (Acosta-Martínez *et al.* 2008; Buée *et al.* 2009; Jumpponen *et al.* 2010; Rousk *et al.* 2010). An alternate approach has been to target all soil microbiota, from archaea to fungi, using either total RNA (Fierer *et al.* 2007) or selected functional gene amplicons (Leininger *et al.* 2006).

Marine environments have also been the subject of ecological research employing NGS technology. Analyses of marine bacterial communities have been conducted using 18S rDNA (Huber *et al.* 2007) and 16S rDNA (Sogin *et al.* 2006) amplicons. Frias-Lopez *et al.* (2008) studied microbial community gene expression in ocean surface waters through transcriptomic sequencing analysis of cDNA libraries. Mou *et al.* (2008) investigated functional assemblages within seawater through a NGS analysis of functional metabolic gene regions within bacterioplankton. Marine eukaryotic microbiota were investigated through NGS analysis of 18S rDNA amplicons (Stoeck *et al.* 2010). A shotgun sequencing approach was employed to investigate microbial and viral diversity in sea water (Williamson *et al.* 2008). Marine viromes were also investigated by Angly *et al.* (2006). Rare and extreme habitats such as acid mines (Edwards *et al.* 2006) and coral reefs (Wegley *et al.* 2007) are now also readily subjected to NGS-based analysis of biodiversity.

Four recent articles have outlined the application of NGS approaches to analysis of freshwater environmental samples. Ficetola *et al.* (2008) combined an NGS approach with conventional Sanger sequencing of cytochrome *b* amplicons to detect the presence of bullfrogs in freshwater samples. Also, freshwater microbialities were investigated with a whole-genome shotgun approach to provide further insight into fossil stromatolite communities (Breitbart *et al.* 2009). Amplicons of 18S rDNA were used to investigate protist diversity in freshwater samples (Medinger *et al.* 2010). Recently, short fragments of COI DNA barcodes were used to provide species-level identification of freshwater macroinvertebrates from benthic samples (Hajibabaei *et al.* 2011). This environmental barcoding study demonstrates the efficiency of 454 pyrosequencing in environmental biomonitoring projects by comparing benthic macro-invertebrate communities from both urban and conservation areas. The analysis of these benthic samples can provide a real-world test of NGS approaches for biomonitoring applications (see Baird & Hajibabaei 2012 in this issue).

Next-generation technology has also been employed in recent research into terrestrial environmental samples, both ancient and modern. Haile *et al.* (2009) utilized both 454 pyrosequencing and conventional Sanger sequencing methods in the analysis of ancient DNA recovered from Arctic permafrost cores. Sønstebø *et al.* (2010) analysed permafrost samples to identify ancient plant species. Both pathogens associated with colony collapse disorder in honey bees (Cox-Foster *et al.* 2007) and plant viruses from infected tomato plants have been identified with NGS technology (Adams *et al.* 2009). A variety of terrestrial microhabitats, specifically soil, leaf litter and canopy epiphytes in a Costa Rican rainforest, have been examined using NGS approaches (Creer *et al.* 2010; Porazinska *et al.* 2010). Amplicons of 16S rDNA have been utilized to explore the sensitivity of topsoil in determining vertebrate presence and diversity in regions with known species compositions (Andersen *et al.* 2011).

Many studies have used NGS technology in diet analysis and in the investigation of gut microbial ecology. Some of these studies have included analyses of herbivore diet from gut contents using the plastid *trnL* sequence (Pegard *et al.* 2009; Soininen *et al.* 2009; Valentini *et al.* 2009; Kowalczyk *et al.* 2011). Also, several studies have been conducted on the effect of diet on the gut microbiome of mice using 16S rDNA amplicons (Turnbaugh *et al.* 2008, 2009; Murphy *et al.* 2010; Ravussin *et al.* 2011; Serino *et al.* 2011). Recently, an investigation of the diet of bats was conducted using short COI amplicons. By enabling species-level identification of dietary components, NGS application to diet analysis allows a comprehensive relationship of the diet of sympatric cryptic species (Razgour *et al.* 2011).

Besides 454 pyrosequencing, the sequencing capacity of Illumina platforms has been successfully utilized for assessing microbial community diversity using short fragments of 16S rDNA (Lazarevic *et al.* 2009). Paired-

end sequencing can increase the read length for Illumina sequencing applications. However, potential sources of error, including sequencing artefacts and taxonomic misidentification, should be taken into consideration when using short-read NGS tools to discover the biodiversity of environmental samples (Degnan & Ochman 2011). A recent study (Miller *et al.* 2011) has shown the promise of the Ion Torrent semiconductor platform for the assessment of intraspecies genetic diversity in an endangered mammal species.

Concluding remarks

The introduction of and advancements in next-generation sequencing have revitalized research in environmental DNA. New methods of gathering sequence data, however, require optimization and benchmarking before being utilized on samples of unknown nature to avoid false negatives and biased results. The excitement of using new platforms has generated momentum among researchers to apply NGS tools in various applications involving environmental DNA. Rapid progress in the last 5 years has provided optimism for a bright future for the field of next-generation environmental DNA analysis.

Acknowledgements

This work was supported by the Government of Canada through funds from Genome Canada and the Ontario Genomics Institute (OGI-050) and by NSERC Canada.

References

Acosta-Martínez V, Dowd S, Sun Y, Allen V (2008) Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. *Soil Biology and Biochemistry*, **40**, 2762–2770.

Adams IP, Glover RH, Monger WA *et al.* (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology*, **10**, 537–545.

Adey A, Morrison HG, Asan *et al.* (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, **11**, R119.

Andersen K, Bird KL, Rasmussen M *et al.* (2011) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Andersson A, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE*, **3**, e2836.

Andreotti R, de León AAP, Dowd SE, Guerrero FD, Bendele KG, Scoles GA (2011) Assessment of bacterial diversity in the cattle tick *Rhipicephalus* (*Boophilus*) *microplus* through tag-encoded pyrosequencing. *BMC Microbiology*, **11**, 6.

Angly FE, Felts B, Breitbart M *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biology*, **4**, 2121–2131.

Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Barzon L, Militello V, Lavezzo E *et al.* (2011) Human papillomavirus genotyping by 454 next generation sequencing technology. *Journal of Clinical Virology*, **52**, 93–97.

Berry D, Mahfoudh KB , Wagner M, Loy A, (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied Environmental Microbiology*, **77**(21), 7846–7849.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

Boessenkool S, Epp L, Haile J *et al.* (2011) Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, **21**, 1806–1815.

Breitbart M, Hoare A, Nitti A *et al.* (2009) Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico. *Environmental Microbiology*, **11**, 16–34.

Buée M, Reich M, Murat C *et al.* (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, **18**, 449–456.

Burgess KS, Fazekas AJ, Kesanakurti PR *et al.* (2011) Discriminating plant species in a local temperate flora using the *rbcL+matK* DNA barcode. *Methods in Ecology and Evolution*, **2**, 333–340.

Caruccio N (2011) Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods in Molecular Biology*, **733**, 241–255.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA*, **106**, 12794–12797.

Claesson M, Wang Q, O'Sullivan O *et al.* (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, **38**, 1–13.

Cox-Foster DL, Conlan S, Holmes EC *et al.* (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, **318**, 283–287.

Creer S, Fonseca VG, Porazinska DL *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, **19**(Suppl. 1), 4–20.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Degnan PH, Ochman H (2011) Illumina-based analysis of microbial community diversity. *The ISME Journal*, **2011**, 1–12.

Edwards RA, Rodriguez-Brito B, Wegley L *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, doi: 10.1093/sysbio/sys004.

Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Fierer N, Breitbart M, Nulton J *et al.* (2007) Metagenomic and small-subunit rRNA analyses of the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*, **73**, 7059–7066.

Flanagan JL, Brodie EL, Weng L *et al.* (2007) Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *Journal of Clinical Microbiology*, **45**, 1954–1962.

Flusberg BA, Webster DR, Lee JH *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, **7**, 461–465.

Frias-Lopez J, Shi Y, Tyson GW *et al.* (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences, USA*, **105**, 3805–3810.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Haile J, Froese DG, MacPhee RDE *et al.* (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences, USA*, **106**, 22352–22357.

Hajibabaei M, Singer GAC, Clare EL, Hebert PDN (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, **5**, 24.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Harris T, Buzby P, Jarosz M *et al.* (2007) Optical train and method for TIRF single molecule detection and analysis. US patent application 20070070349.

Harris T, Buzby P, Babcock H *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.

Harris JK, Sahl JW, Castoe TA *et al.* (2010) Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. *Applied and Environmental Microbiology*, **76**, 3863–3868.

Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B*, **270**, S96–S99.

Horn S (2012) Target enrichment via DNA hybridization capture. *Methods in Molecular Biology*, **840**, 177–188.

Huber JA, Welch DBM, Morrison HG *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, **12**, 1889–1898.

Ishii K, Fukui M (2001) Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Applied and Environmental Microbiology*, **67**, 3753–3755.

Jones MA, Bhide S, Chin E *et al.* (2011) Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. *Genetics in Medicine*, **13**, 921–932.

Jumpponen A, Jones KL, Blair J (2010) Vertical distribution of fungal communities in tall grass prairie soil. *Mycologia*, **102**, 1027–1041.

Korlach J, Marks PJ, Cicero RL *et al.* (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Science*, **105**, 1176–1181.

Korlach J, Bjornson KB, Chaudhuri BP *et al.* (2010) Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology*, **472**, 431–455.

Kowalczyk R, Taberlet P, Coissac E *et al.* (2011) Influence of management practices on large herbivore diet—Case of European bison in Białowieza Primeval Forest (Poland). *Forest Ecology and Management*, **261**, 821–828.

Kurata S, Kanagawa T, Magariyama Y *et al.* (2004) Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Applied and Environmental Microbiology*, **70**, 7545–7549.

Lazarevic V, Whiteson K, Huse S *et al.* (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods*, **79**, 266–271.

Lee E-J, Pei L, Srivastava G *et al.* (2011) Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic Acids Research*, **39**, e127.

Leininger S, Urich T, Schloter M *et al.* (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**, 806–809.

Levene MJ, Korlach J, Turner SW *et al.* (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.

Lim YW, Kim BK, Kim C *et al.* (2010) Assessment of soil fungal communities using pyrosequencing. *Journal of Microbiology*, **48**, 284–289.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Mardis ER (2008a) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Mardis ER (2008b) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Margulies M, Egholm M, Altman W *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Medinger R, Nolte V, Pandey RM *et al.* (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, **19**(Suppl. 1), 32–40.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Meusnier I, Singer GAC, Landry J-F *et al.* (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 214.

Miller W, Hayes V, Ratan A *et al.* (2011) Genetic diversity and population structure of the endangered marsupial *Sarcophilus*

harrisii (Tasmanian Devil). *Proceedings of the National Academy of Sciences, USA*, **108**, 12348–12353.

Mou X, Sun S, Edwards RA, Hodson RE, Moran MA (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature*, **451**, 708–713.

Murphy EF, Cotter PD, Healy S *et al.* (2010) Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. *Gut*, **59**, 1635–1642.

Nacke H, Thürmer A, Wollherr A *et al.* (2011) Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PLoS ONE*, **6**, e17000.

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H (2008) Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics*, **4**, 193–201.

Pegard A, Miquel C, Valentini A *et al.* (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, **57**, 5700–5706.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.

Porazinska DL, Giblin-Davis RM, Esquivel A, Powers TO, Sung W, Thomas WK (2010) Ecometagenetics confirms high tropical nematode diversity. *Molecular Ecology*, **19**, 5521–5530.

Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, **9**, 847–850.

Pushpendra KG (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology*, **26**, 602–611.

Qiu X, Wu L, Huang H *et al.* (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology*, **67**, 880–887.

Quince C, Lanzén A, Curtis TP *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, **6**, 639–641.

Ravussin Y, Koren O, Spor A *et al.* (2011) Responses of gut microbiota to diet composition and weight loss in lean and obese mice. *Obesity*, doi: 10.1038/oby.2011.111.

Razgour O, Clare EL, Zeale MRK *et al.* (2011) High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecology and Evolution*, doi: 10.1002/ece3.49

Roesch LFW, Fulthorpe RR, Riva A *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**, 283–290.

Rothberg J, Hinz W, Rearick T *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Rousk J, Bååth E, Brookes PC *et al.* (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. *The ISME Journal*, **4**, 1340–1351.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, USA*, **74**, 5463–5467.

Schuster S. C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**(1) 16–18.

Serino M, Luche E, Gres S *et al.* (2011) Metabolic adaptation to a high-fat diet is associated with a change in the gut microbiota. *Gut*, **61**, 543–553.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

Soininen EM, Valentini A, Coissac E *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, **6**, 16.

Sønstebø JH, Gielly L, Brysting AK *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, **10**, 1009–1018.

Stoeck T, Bass D, Nebel M *et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21–31.

Sun Y, Wolcott RD, Dowd SE (2011) Tag-encoded FLX amplicon pyrosequencing for the elucidation of microbial and functional gene diversity in any environment. *Methods in Molecular Biology*, **733**, 129–141.

Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.

Tewhey R, Nakano M, Wang X *et al.* (2009) Enrichment of sequencing targets form the human genome by solution hybridization. *Genome Biology*, **10**, R116.

Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI (2008) Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe*, **3**, 213–223.

Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*, **1**, 6ra14.

Valentini A, Miquel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, **9**, 51–60.

Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environmental Microbiology*, **9**, 2707–2719.

Williamson SJ, Rusch DB, Yooseph S *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, **3**, e1456.

Wu GD, Lewis JD, Hoffmann C *et al.* (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology*, **10**, 206.

Xu M, Chen X, Qiu M *et al.* (2012) Bar-coded pyrosequencing reveals the responses of PDBE-degrading microbial com-

munities to electron donor amendments. *PLoS ONE*, **7**, e30439.

Zhang H, DiBaise J, Zuccolo A *et al.* (2009) Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences, USA*, **106**, 2365–2370.

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, **38**, 95–109.

Zhou XG, Ren LF, Li YT *et al.* (2010) The next-generation sequencing technology: a technology review and future perspective. Science China. *Life Sciences*, **53**, 44–57.

S.S. is a molecular microbiologist broadly interested in genomics technology development including laboratory protocols and data analysis tools especially for metagenomics and environmental applications. J.L.S. is a graduate student interested in evaluating the potential of NGS in providing biodiversity measurements at the community level for bioindicator species commonly used in biomonitoring programs and ecological assessments. J.F.G. is a molecular systematist with research interests including biodiversity of terrestrial arthropods, especially Diptera. He is currently developing NGS-based applications for monitoring arthropod diversity. M.H. is a molecular evolutionary biologist interested in studying biodiversity in different ecological settings through comparative analysis of genomics information. He currently leads the Biomonitoring 2.0 project (www.biomonitoring2.org), a large-scale effort that utilizes NGS technologies for monitoring environmental change.
