frontiers in
PLANT SCIENCE

REVIEW ARTICLE
published: 21 May 2014
doi: 10.3389/fpls.2014.00216

Analysis of plant microbe interactions in the era of next generation sequencing technologies

Claudia Knief*

Institute of Crop Science and Resource Conservation—Molecular Biology of the Rhizosphere, Faculty of Agriculture, University of Bonn, Bonn, Germany

Edited by:
Ann E. Stapleton, University of North Carolina Wilmington, USA

Reviewed by:
Xingyi Guo, Albert Einstein College of Medicine, USA
Ute Hentschel, University of Wuerzburg, Germany

***Correspondence:**
Claudia Knief, Institute of Crop Science and Resource Conservation–Molecular Biology of the Rhizosphere, Faculty of Agriculture, Nussallee 13, 53115 Bonn, Germany
e-mail: knief@uni-bonn.de

Next generation sequencing (NGS) technologies have impressively accelerated research in biological science during the last years by enabling the production of large volumes of sequence data to a drastically lower price per base, compared to traditional sequencing methods. The recent and ongoing developments in the field allow addressing research questions in plant-microbe biology that were not conceivable just a few years ago. The present review provides an overview of NGS technologies and their usefulness for the analysis of microorganisms that live in association with plants. Possible limitations of the different sequencing systems, in particular sources of errors and bias, are critically discussed and methods are disclosed that help to overcome these shortcomings. A focus will be on the application of NGS methods in metagenomic studies, including the analysis of microbial communities by amplicon sequencing, which can be considered as a targeted metagenomic approach. Different applications of NGS technologies are exemplified by selected research articles that address the biology of the plant associated microbiota to demonstrate the worth of the new methods.

Keywords: next generation sequencing, plant microbiota, phyllosphere, rhizosphere, metagenomics, amplicon sequencing

INTRODUCTION

Plants live in association with diverse microorganisms, which thrive below ground in the rhizosphere and above in the phyllosphere (Vorholt, 2012; Bulgarelli et al., 2013). They are found as endophytes within the plant, as epiphytes attached on plant surfaces and in the nearby soil around the roots. These microorganisms can have beneficial, neutral, or detrimental effects on plant health and development (Newton et al., 2010). The majority of the diverse plant colonizing microorganisms follows a commensal lifestyle; they do not cause obvious harm to the plant, nor do they exert a strong plant growth promoting effect as known for instance from symbiotic nitrogen-fixing bacteria or mycorrhizal fungi. The opening questions to better understand the association between plants and their associated microbiota are the "Who is there?" and "What are they doing?" These are extended by "How do they life under given conditions?" "How do they respond to environmental changes and perturbations?" "How do they interact with each other?" and "How do they affect plant health and development?" Finding answers to these questions will lead to a better understanding of the association between microorganisms and plants; a prerequisite to assess if and how associated microorganisms may be used in the future to support plant growth and improve crop yield.

DNA based studies of the plant associated microbiota are of high value to address the aforementioned questions. Genomic analyses of individual microbial strains or metagenomic studies of whole microbial communities provide insight into the composition and physiological potential of plant associated microorganisms. RNA based studies can extend such studies in order to elucidate the actual metabolic activities and regulatory mechanisms of the microbial cells under given conditions. NGS technologies have a tremendous impact on DNA and RNA based analysis methods; they allow finding answers to questions that could not be addressed before, largely due to technical and financial limitations. Thus, plant microbe associations can now be studied at a speed and depth as never before.

The present review summarizes the main features of the currently available NGS systems and gives a brief outlook about what may be expected in the future. It critically discusses limitations of NGS platforms and shows up ways to compensate these. Applications in the context of plant-microbe-interactions are highlighted that profit from these new technologies, focusing on metagenomic analyses.

NEXT GENERATION SEQUENCING TECHNOLOGIES

Different NGS systems have in common that they produce a massive amount of sequencing data (up to gigabases and soon even terabases) in parallel. Often, NGS instruments are classified as second and third generation sequencing technologies (e.g., Schadt et al., 2010; Niedringhaus et al., 2011; Pareek et al., 2011; Liu et al., 2012). There is no consistent definition for this terminology, and it is difficult to assign all different instruments unambiguously to one or the other category (Schadt et al., 2010; Thompson and Milos, 2011). In this review I refer to all those methods that depend on a PCR step for signal intensification prior to sequencing as second generation sequencing instruments, opposed to single molecule sequencing. Second generation sequencing technology includes the 454 instruments from Roche, the different Illumina platforms and the Life Technologies instruments, i.e., the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) and Ion Torrent sequencers. The only third generation sequencing instrument that

is currently commercially available is the PacBio RS by Pacific Biosciences.

COMMON AND DISTINCT FEATURES OF SECOND GENERATION SEQUENCING TECHNOLOGIES

The main characteristics of NGS sequencers are described here in a comparative way in order to point out similarities and differences. A detailed description of second generation sequencing platforms and principles can be found in dedicated reviews (e.g., Voelkerding et al., 2009; Metzker, 2010; Glenn, 2011; Pareek et al., 2011; Zhang et al., 2011; Liu et al., 2012; Shokralla et al., 2012; Mardis, 2013; Morey et al., 2013). Despite differences in terms of sequencing principle, all current second generation sequencing platforms have several shared features with regard to library preparation, library amplification and the sequencing process (**Figure 1**, **Table 1**).

LIBRARY PREPARATION

Library preparation can be done from DNA (genomic or PCR amplified fragments) or RNA as input material. The latter has to be converted into cDNA during the library preparation process, direct sequencing of RNA is not yet possible. Due to size limitations for library molecules, genomic DNA and often also mRNA is fragmented, which is usually done mechanically, e.g., by sonication or nebulization, or enzymatically. The fragment size of a library is critical and depends on the sequencing platform that is going to be used. The standard fragment size of Illumina libraries is between 300 and 550 bp including adapters. Longer fragments up to 800 bp can be sequenced if cluster density on the flow cell is reduced to prevent interference of library molecules during the sequencing process. The size of libraries prepared for 454 sequencing depends on the sequencing run conditions. To obtain long reads with a modal length of 700 bp, a size of approximately 1500 bp is recommended. Libraries prepared for sequencing on the small-scale 454 Junior instrument or for sequencing using the older FLX chemistry should be smaller (300–750 bp). Libraries that are sequenced on the Ion Torrent Personal Genome Machine (PGM) platform should never be longer than the requested read length.

Libraries are constructed by adding sequencing platform-specific DNA adapters to the DNA molecules. This enables binding of the library fragments to a surface, which is either a microbead (454, Ion PGM, SOLiD) or a glass slide (Illumina, SOLiD). Moreover, the adapters allow amplification of the library fragments by emulsion PCR (emPCR) or bridge PCR. When amplicons are sequenced, e.g., in microbial community analyses, adapters are often already added during PCR using fusion primer constructs.

Diverse library preparation kits are commercially available and even more protocols have been published that are adapted to the specific needs of research projects. During the last years, library preparation methods were streamlined to reduce costs and preparation time and to enable high throughput library preparation on automated systems (e.g., Adey et al., 2010; Caruccio, 2011; Neiman et al., 2012; Rohland and Reich, 2012; Langevin et al., 2013) Methods were also optimized to reduce potential bias, e.g., by excluding PCR amplification steps (Kozarewa et al., 2009; Adey et al., 2010; Mamanova and Turner, 2011; Oyola et al., 2012; Van Dijk et al., 2014). Another goal is the reduction of the amount of input material. This ranges from several micrograms down to hundreds of pictograms (e.g., Adey et al., 2010; Tariq et al., 2011; Parkinson et al., 2012; Bowman et al., 2013; Langevin et al., 2013). In microbial metagenomic studies, which often aim at in-depth analysis of gene diversity, it is advisable to prepare libraries from microgram amounts of input material to cover as much of the diversity as possible and obtain high sequencing depth. It also has to be considered that library preparation from just a few nanograms of input material will require additional PCR steps to amplify the material, which is a potential source of bias.

Library construction using standard methods can easily be outsourced. If library preparation is done by oneself, care has to be taken that the generated libraries are compatible with the sequencing platform that is used for sequencing, as adapters were in some cases modified since the release of the first instruments. For instance, the sequencing of libraries that are constructed according to an Illumina GAIIx protocol is not necessarily fully supported on HiSeq or MiSeq instruments. Details should be discussed prior to the preparation of libraries with the sequence provider.

BARCODING OF LIBRARIES

At least one of the library adapters usually carries a library specific DNA sequence, often a 6- to 12-mer, referred to as barcode, molecular identifier (MID) or tag. This barcode enables the pooling of different libraries, which can then be further processed and sequenced within the same region of a picotiterplate (454), a lane of a flow cell (Illumina, SOLiD) or on a chip (Ion PGM). Barcoding allows sequencing of a complex set of libraries at rather low depth, which is of particular interest in large-scale ecological or biodiversity studies comprising many samples. In amplicon sequencing projects, a sample specific barcode is often already added during PCR amplification of the target genes to enable parallel sample processing at an early step. It should be noted that bias may be introduced when using complex fusion primers with adapters and different barcodes. This can be compensated to certain extent by using a two-step PCR procedure (Berry et al., 2011).

Several different barcode sets have been developed by hand or using software tools. They vary in length and account more and more strictly for different types of sequencing errors and sequencing platform specific needs to maximize data output (Faircloth and Glenn, 2012 and references therein; Kircher et al., 2012; Buschmann and Bystrykh, 2013; Costea et al., 2013). In some articles the use of a dual barcoding strategy is proposed for paired end sequencing in order to decrease sample misidentification rate or to decrease the number of individually tagged PCR primers (Gloor et al., 2010; Carlsen et al., 2012; Degnan and Ochman, 2012; Kircher et al., 2012; Kozich et al., 2013).

LIBRARY AMPLIFICATION BY EMULSION PCR OR BRIDGE PCR

PCR amplification of the library molecules is required to increase signal intensity for the sequencing process. Amplification has to occur spatially separated for the individual library fragments on microbeads (454, PGM, SOLiD) via emPCR or on a glass surface

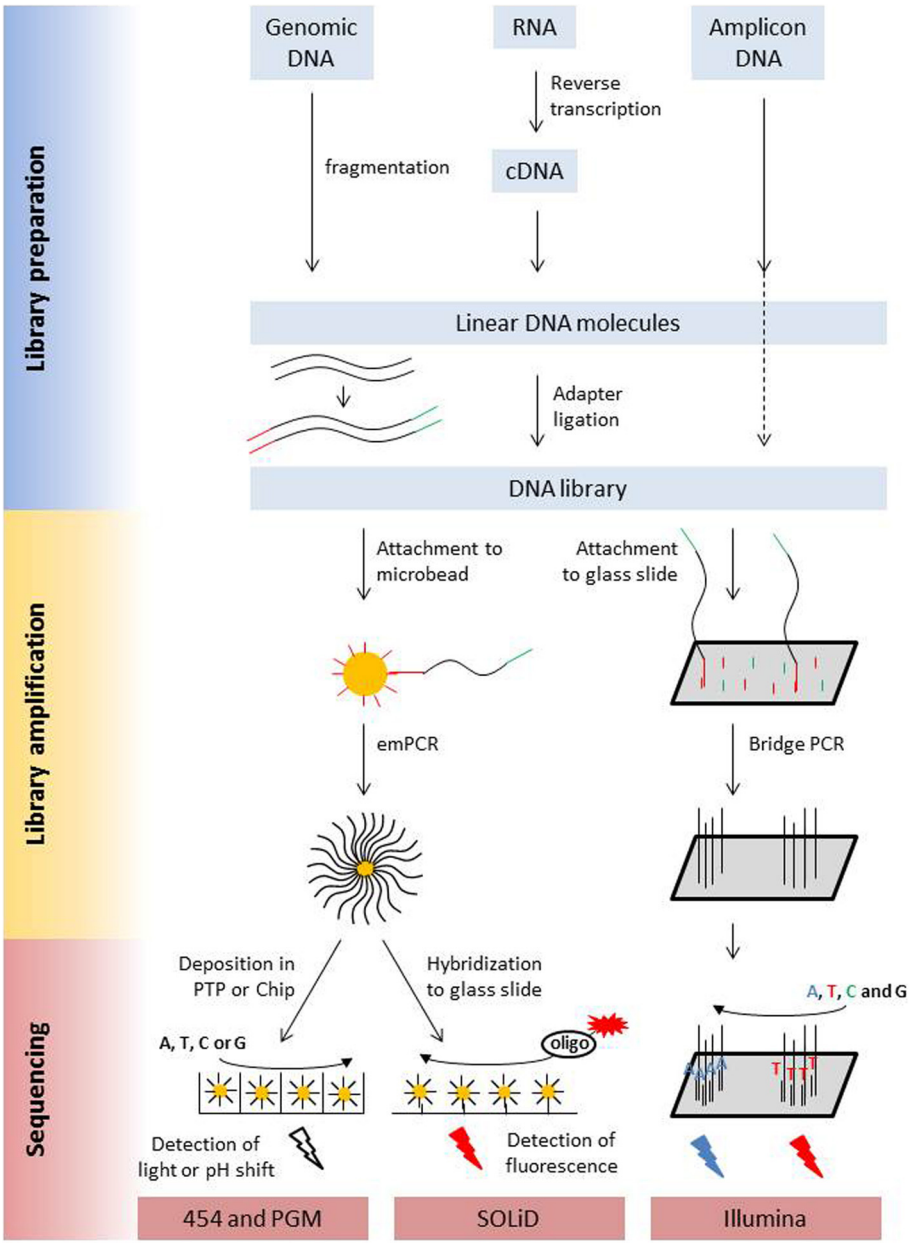

FIGURE 1 | Schematic presentation of the library preparation and sequencing process of the most commonly used next generation sequencing platforms. All different types of starting molecules are converted into doublestranded DNA molecules that are flanked by adapters. Adapters are sequencing platform specific and enable the binding of the library molecules to surfaces, either beads or a flow cell, where they are amplified prior to sequencing. Clonal amplicons are spatially separated on the glass slides, chips, or picotiterplate. Sequencing is either a sequencing by ligation process with fluorescently labeled oligonucleotides of known sequence (SOLiD) or a sequencing by synthesis process. During Illumina sequencing, four differently labeled nucleotides are flushed over the flow cell in multiple cycles, depending on the desired read length. During 454 and Ion PGM sequencing unlabeled nucleotides are flushed in a sequential order over the flow cell. Incorporation is detected via a coupled light reaction (454) or the detection of proton release during nucleotide incorporation.

(Illumina, SOLiD) via bridge PCR. Hybridization of the library fragments to the surfaces occurs via the adapters to surface-bound oligonucleotides. In the bead based method, each bead obtains only a single library molecule. The beads are spatially separated from each other during emPCR in individual water droplets in a water-oil emulsion. Beads with successfully amplified fragments are enriched and deposited in a picotiterplate (454), a semiconductor chip (Ion PGM) or hybridized to a glass surface (SOLiD) for sequencing. When library molecules are directly hybridized to a glass surface, their density on the surface has to be sufficiently low to prevent interference of library molecules, even after fragment amplification via bridge PCR (**Figure 1**).

Table 1 | Technological specifications of currently commercially available next generation sequencing platforms.

Company (former companies)	Platforms	Library amplification	Carrier of library molecules or beads during sequencing	Sequencing principle	Nucleotide modifications	Signal detection method	Dominant type of sequencing error
Roche (454 until 2006)	454 FLX Titanium 454 FLX+ 454 GS Junior Titanium	emPCR on microbeads	Picotiterplate	Pyrosequencing	None (except for dATP, which is added as thiol derivative dATPαS)	Optical detection of light, emitted in secondary reactions initiated by release of PP_i upon nucleotide incorporation	Indels in homopolymeric regions
Illumina (Solexa until 2007)	Illumina GAIIx Illumina HiSeq1000 Illumina HiSeq1500 Illumina HiSeq2000 Illumina HiSeq2500 Illumina MiSeq Illumina NextSeq 500 Illumina HiSeq X ten	Bridge-PCR on flow cell surface	Flow cell	Reversible terminator sequencing by synthesis	End-blocked fluorescent nucleotides	Optical detection of fluorescent emission from incorporated dye-labeled nucleotides	Substitutions, in particular at the end of the read
Life Technologies (Agencourt until 2006, Applied Biosystems until 2008)	SOLiD 4 SOLiD 5500 SOLiD 5500xl SOLiD 5500 W SOLiD 5500xl W	emPCR on microbeads; PCR on FlowChip surface for the 5500 W models	FlowChip	Sequencing by ligation	2-base encoded fluorescent oligonucleotides	Optical detection of fluorescent emission from ligated dye-labeled oligonucleotides	Substitutions, in particular at the end of the read
Life Technologies (Ion Torrent until 2010)	Ion PGM Ion Proton	emPCR on microbeads	Ion Chip, a semiconductor chip	Semiconductor-based sequencing by synthesis	None	Transistor-based detection of H^+ shift upon nucleotide incorporation	Indels
Pacific biosciences	PacBio RS	Not applied	SMRT cell	Single-molecule, real-time DNA sequencing by synthesis	Phosphor-linked fluorescent nucleotides	Real-time optical detection of fluorescent dye in polymerase active site during incorporation	Indels

Since the production and recovery of successfully templated beads from the water-oil emulsion during emPCR is time consuming, technically challenging and rather expensive, sequencing companies search for alternative methods to amplify library molecules. This has been realized in the recently released Wildfire technology for the SOLiD sequencer (SOLiD 5500 W) and is under development for Ion Torrent sequencers (Merriman et al., 2012).

THE SEQUENCING PROCESS

Sequencing is performed in a massively parallel manner for ten thousands to billions of library fragments. It occurs via repeated cycles of nucleotide addition by a DNA polymerase or ligase (SOLiD), detection of incorporated nucleotides and washing steps. Due to this iterative procedure including extensive washing steps, sequencing lasts several hours to days. In case of Illumina and SOLiD sequencing the four differently labeled nucleotides are flushed over the glass slide in parallel, while a sequential flooding of non-labeled native nucleotides occurs during 454 and Ion PGM sequencing. In the former case incorporation of nucleotides is detected based on specific fluorescent labels attached to the nucleotide, in the latter case products of the enzymatic nucleotide incorporation reaction are detected, i.e., proton or pyrophosphate release. While proton release can be directly measured as pH change by the semiconductor chip of the Ion Torrent instruments (Merriman et al., 2012), the pyrophosphate signal is further converted into a light signal via subsequent reactions including the enzyme luciferase (Ronaghi et al., 1998). The generation of a light signal has led to the term "pyrosequencing" for this technology.

The different strategies of adding nucleotides to the DNA template strand affect sequence read length. During Illumina and SOLiD sequencing, a blocking group at each of the (oligo-) nucleotides prevents the addition of more than one molecule, so that the sequence is increased by one (oligo-) nucleotide at each step and the full read length is determined by the number of sequencing cycles performed (Bentley et al., 2008). In contrast, 454 and Ion PGM sequencing result in sequence reads of variable length. Due to the fact that the four different nucleotides are applied in a specified sequential order, a variable number of nucleotides is incorporated after four cycles, depending on the sequence of the respective library molecules. Several nucleotides are incorporated within the same cycle if the DNA template strand shows a homopolymeric region. This comes along with a proportional increase in signal strength, so that signal intensity is used to calculate the number of incorporated nucleotides (Margulies et al., 2005).

SPECIFICATIONS OF THE DIFFERENT SEQUENCING PLATFORMS

Major progress has been made during the last years with regard to sequence read length and output (number of reads per run) by technically improving the instruments, the chemistry and base-calling algorithms. A compilation of current specifications as given in **Table 2** is useful to assess and compare the potential of the different instruments. The presented data were taken from the websites of the sequence providers. It should be kept in mind that those data were generated under optimum conditions. The specifications may not be met when more difficult sampling material is sequenced, e.g., libraries with more extreme GC content or of sub-optimal fragment length.

The SOLiD and Illumina HiSeq sequencers generate the largest amount of data per run at the lowest costs per base. Soon Illumina HiSeq instuments will produce up to 1000 Gb per run. At the same time, these platforms generate the shortest reads. In particular the very short SOLiD sequence reads are mostly used for resequencing and transcriptomics projects, in which reads can be mapped to known genomes, but not frequently in *de novo* sequencing projects. Between 8 and 11 days are needed to perform a run with maximum data output on these instruments. Illumina has developed strategies during the last years to reduce run time, resulting in the upgrade of the HiSeq 2000 instrument to HiSeq 2500. The upgrade allows sequencing in rapid run mode, which produces a smaller amount of data (approximately 25–30% of data compared to a so-called "high-output" run) within hours to 2 days, depending on the desired read length. The upgrade came along with an increase in maximum read length from 100 to 150 bp in rapid run mode.

The Illumina MiSeq platform was launched in 2011. This platform produces 22–25 million reads with a maximum length of 300 bp when using the new V3 chemistry. The costs per sequenced base are higher compared to the HiSeq instrument. However, the longer read length in combination with the lower read number can be of particular interest for amplicon sequencing projects. It is also very suitable for small scale metagenomics projects or initial sample evaluation prior to deep sequencing on a HiSeq. The newest releases from Illumina are the NextSeq 500 platform, which performs at intermediate scale in terms of output, read length, and costs per base compared to HiSeq and Miseq, and the HiSeqX ten, a package of 10 HiSeq sequencers, which allow even higher throughput than the HiSeq2500 in shorter time.

The 454 sequencer was the first commercially available NGS instrument (since 2005). In comparison to Illumina and SOLiD platforms, it generates longer reads (modal read length 750 bp, average read length 700 bp) in a shorter run time (1 day) using FLX+ chemistry. The total output per run of this platform is clearly lower in terms of reads (1 million) and bases (700 Mb). The higher costs per base are a major reason why its use is meanwhile often replaced by the aforementioned platforms, in particular in projects in which coverage is more important than read length, as it is for instance the case in transcriptomics projects, some metagenomic projects or amplicon sequencing projects. Also Roche has released a smaller-scale benchtop sequencing instrument, the 454 GS Junior (available since 2009). This sequencer produces approximately 100,000 reads per run with a modal read length of 450 bp, comparable to the read length obtained with the FLX+ platform when run with FLX chemistry instead of FLX+ chemistry.

The Ion Torrent PGM sequencer is available on the market since the end of 2010. Sequencing on this platform is done using semiconductor chips of different scale, which allow to sequence between 0.4 and 5.5 million reads. Read length on this platform increased successively from approximately 100 bp to meanwhile 400 bp. Sequencing on Ion instruments is very fast, taking only a couple of hours. The Ion Proton is a larger-scale instrument that produces 10-fold more bases per run using the Ion PI chip. A

Table 2 | Data output of currently commercially available next generation sequencing platforms.

Company platform	No of units on sequencing support	Sequencing run conditions and read length[a]	Sequencing run time[b]	Maximum data output per run[c]	Maximum output in mio reads[d]
ROCHE					
454 FLX+	1 PTP with gaskets to separate 2, 4, 8 or 16 regions	FLX (modal 450 bp, max. 600 bp)	10 h	450 Mb	1 per PTP (0.7 for amplicons)
		FLX+ (modal 700 bp, max. 1000 bp)	23 h	700 Mb	1 per PTP (0.7 for amplicons)
454 GS Junior Titanium	1 PTP	~450 bp	10 h	35 Mb	0.1 per PTP (0.07 for amplicons)
ILLUMINA					
HiSeq 2000/2500 (High output mode) V3 kits	8 lanes per flow cell, 1 or 2 flow cells per run	36 bp	2 days	95–105 Gb	165–185 per lane
		2 × 50 bp	5.5 days	270–300 Gb	
		100 bp	5 days	270–300 Gb	
		2 × 100 bp	11 days	540–600 Gb	
HiSeq 2000/2500 (High output mode) V4 kits	8 lanes per flow cell, 1 or 2 flow cells per run	36 bp	29 h	128–144 Gb	250 per lane
		2 × 50 bp	2.5 days	360–400 Gb	
		2 × 100 bp	5 days	720–800 Gb	
		2 × 100 bp	6 days	900–1000 Gb	
HiSeq 2500 (Rapid run mode) V3 kits	2 lanes per flow cell (not independent), 1 or 2 flow cells per run[e]	36 bp	7 h	18–22 Gb	125–150 per lane
		2 × 50 bp	16 h	50–60 Gb	
		2 × 100 bp	27 h	100–120 Gb	
		2 × 150 bp	40 h	150–180 Gb	
HiSeq X ten[f]	1 or 2 flow cells	2 × 150 bp	<3 days	1.6–1.8 Tb	3000 per flow cell
miSeq, V2 kits	1 lane, 1 flow cell	36 bp	4 h	540–610 Mb	12–15 per flow cell
		2 × 25 bp	5.5 h	750–850 Mb	
		2 × 150 bp	24 h	4.5–5.1 Gb	
		2 × 250 bp	39 h	7.5–8.5 Gb	
miSeq, V3 kits	1 lane, 1 flow cell	2 × 75 bp	24 h	3.3–3.8 Gb	22–25 per flow cell
		2 × 300 bp	55 h	13.2–15 Gb	
NextSeq 500 (High output mode)	4 lanes (not independent), 1 flow cell[e]	75 bp	11 h	25–30 Gb	400 per flow cell
		2 × 75 bp	18 h	50–60 Gb	
		2 × 150 bp	29 h	100–120 Gb	
NextSeq 500 (Mid output mode)	4 lanes (not independent), 1 flow cell[e]	2 × 75 bp	15 h	16–20 Gb	130 per flow cell
		2 × 150 bp	26 h	32–39 Gb	
LIFE TECHNOLOGIES					
SOLiD 5500xl	2 × 6 lanes	75 bp	5 days	160 Gb	160 per lane
		75 bp + 35 bp	8 days	220 Gb	
		60 bp + 60 bp	8 days	260 Gb	
SOLiD 5500xl W	2 × 6 lanes	50 bp	4 days	160 Gb	265 per lane
		75 bp	5 days	240 Gb	
		2 × 50 bp	8 days	320 Gb	
Ion PGM, 314 chip v2	1 Chip	200 bp mode	2.3 h	30–50 Mb	0.4–0.55 per chip
		400 bp mode	3.7 h	60–100 Mb	

(Continued)

Table 2 | Continued

Company platform	No of units on sequencing support	Sequencing run conditions and read length[a]	Sequencing run time[b]	Maximum data output per run[c]	Maximum output in mio reads[d]
Ion PGM, 316 chip v2	1 Chip	200 bp mode 400 bp mode	3.0 h 4.9 h	300–600 Mb 600 Mb–1 Gb	2–3 per chip
Ion PGM, 318 chip v2	1 Chip	200 bp mode 400 bp mode	4.4 h 7.3 h	600 Mb–1 Gb 1.2–2.0 Gb	4–5.5 per chip
Ion Proton, PI chip	1 Chip	200 bp mode	2–4 h	Up to 10 Gb	60–80 per chip
PACIFIC BIOSCIENCES					
PacBio RS II	Up to 16 SMRT cells	C2/P4 chemistry, mean read length ~8000 bp	2–3 h per cell	400 Mb per cell	0.05 per SMRT cell

[a] *"2 ×" refers to paired end runs; more run conditions in the given range are possible for Illumina instruments.*

[b] *Sequencing time does not include library amplification, except for the MiSeq and NextSeq platforms.*

[c] *Output for 2 flow cells per run in case of the Illumina HiSeq systems.*

[d] *The two reads of a paired end read are counted as one paired end read here.*

[e] *Lanes can only be independently loaded with different libraries if cluster amplification is done on the cBot.*

[f] *Not yet available, dedicated to human genome sequencing.*

larger scale chip (Ion PII) is announced for this platform. In terms of sequencing costs per base, the Ion PGM ranges in between 454 and Illumina/SOLiD technologies.

PAIRED END SEQUENCING AND MATE PAIR LIBRARIES

Most sequencers allow sequencing of library fragments from both ends. A corresponding reverse read can be assigned to each individual forward read in Illumina and SOLiD paired end sequencing mode. Since the average size of the library molecules is known, the distance between forward and reverse read is also known. This information is very helpful when performing assembly or read mapping. Paired end reads can also be used to improve sequence quality of short amplicons when overlapping reads are generated. Paired end sequencing is also possible on the Ion Torrent instruments and protocols are available, but this sequencing mode is not yet officially supported by the company.

Paired end sequencing can be done for library fragments of up to approximately 800 bp. However, in *de novo* sequencing projects read pairs spanning even larger distances are helpful to bridge longer repetitive regions (Mavromatis et al., 2012). Paired sequence reads spanning distances between 1.5 and 20 kb can be obtained from mate pair libraries. The construction principle of such libraries is shown in **Figure 2**. Mate pair libraries are sequenced in paired end run mode if available. On 454 instruments, mate pair libraries can also be sequenced; the reads will contain sequence information from both ends, separated by the linker sequence somewhere in the middle of the read.

The construction of mate-pair libraries is quite expensive not only monetarily, but also with regard to the amount of input material. Mate pair libraries spanning long distances need 15–20 µg of high molecular weight DNA of which most is lost during the enrichment step of the end-to-end ligated fragments. A certain percentage of library molecules will consist of molecules in which one of the two ends is only represented by a few nucleotides due to the random fragmentation process of the circularized molecules. Such short fragments cannot be assembled with certainty and are discarded. Moreover, the library construction procedure is not free of bias, which can negatively affect assembly, and the diversity of fragments can be rather low, in particular when the amount of input material is limited. When sequencing organisms with small genomes such as bacterial strains, a few hundred thousand reads are usually sufficient to cover the diversity of constructs present in a library. The use of sequencing platforms that produce long reads such as the PacBio instrument appears to become an interesting alternative to mate pair library sequencing.

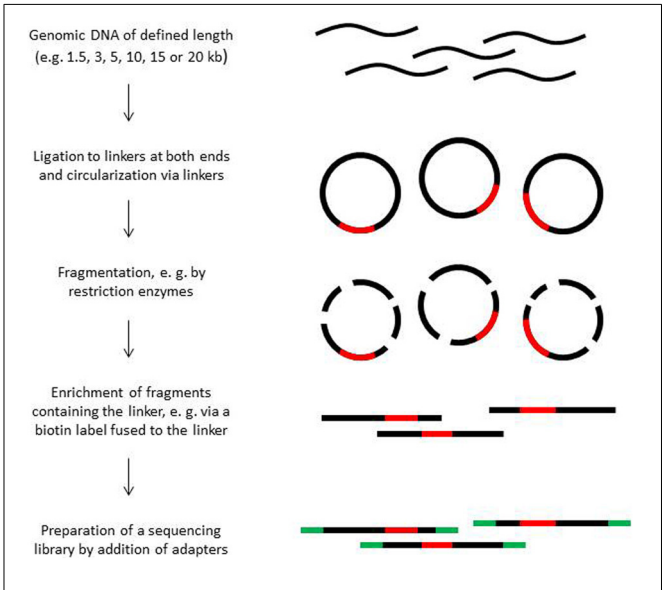

FIGURE 2 | Construction of mate pair libraries.

SINGLE MOLECULE SEQUENCING

Despite the fact that single molecule sequencing approaches are mostly still under development, they have already been described

in diverse review articles (e.g., Gupta, 2008; Xu et al., 2009; Schadt et al., 2010; Treffer and Deckert, 2010; Niedringhaus et al., 2011; Pareek et al., 2011; Zhang et al., 2011; Liu et al., 2012; Morey et al., 2013). Currently, the instrument from Pacific Biosciences is the only commercially available platform. Helicos Biosciences, the company that actually released the first single molecule sequencer, vanished from the market in 2012. The major goals that guide the development of single molecule sequencing platforms are longer read length, higher throughput, higher accuracy, faster turnaround time and lower costs per base (Schadt et al., 2010). It remains to be seen how well all these specifications can be met by one single instrument and which of the different systems currently under development will successfully establish on this highly competitive market.

SINGLE MOLECULE SEQUENCING WITH THE PacBio RS

The sequencing technology of the PacBio RS is described in detail in the above mentioned reviews about single molecule sequencing and in articles that introduce this sequencing system to the scientific community (Eid et al., 2009; Korlach et al., 2010). In brief, the principle of this single molecule real-time (SMRT) technology is to attach a DNA polymerase molecule on the bottom surface of a zero-mode waveguide detector (ZMW). The ZMW enables the detection of fluorescence of individual nucleotides that are incorporated by the polymerase into a single complementary DNA strand during the synthesis process. Each type of dNTP has a unique fluorescent label that is cleaved off during DNA synthesis. The ZMWs can be considered as densely arranged nano-chambers in a perforated metal film on top of a glass surface, enabling the parallelization of the sequencing process in 150,000 ZMWs within a SMRT cell (Levene et al., 2003). The ZMWs are scanned for fluorescent signals by a confocal imaging system, resulting in movies of up to 120 min or even 240 min in the near future that document the successive incorporation of nucleotides, from which the sequence is deduced. Nucleotide incorporation occurs continuously without intermittent washing steps, which accelerates sequencing substantially compared to second generation sequencing systems.

Initially, the DNA synthesis reaction could be monitored only in half of the ZMWs on the PacBio RS system at the same time, but a recent upgrade to RS II enables parallel recording of all ZMWs. However, not all ZMWs produce usable reads, so that the expected number of reads for a SMRT cell is approximately 50,000 for the RS II system. Currently, sequencing is done with the C2/P4 chemistry, but will soon be changed to C3/P5, which will support longer movies and thus the generation of longer reads. The mean read length of the instrument is around 8000 bases, probably increasing to 8500 bases with the new chemistry. A maximum read length of more than 20 kb was observed in different projects, reads of 16 kb are regularly obtained in runs with good quality libraries. In comparison to other sequencing platforms, read length and sequencing time are superior, while output per run is clearly lower and the costs per base are rather high. However, the costs for one SMRT cell are relatively low. These specifications suit in particular bacterial genome sequencing projects.

To improve sequence read quality, a circular consensus sequencing (CCS) strategy was developed. It is based on the fact that PacBio libraries have a circular molecule structure, referred to as SMRTbell template (Travers et al., 2010). These libraries are constructed by ligating hairpin loop adapters to the DNA fragments. The circular structure allows a continuous and repeated sequencing of sense and antisense strand, which can be used to generate single consensus reads with very high accuracy (>99%). The accuracy comes at the expense of read length, since the maximum recording time is limited. Thus, the length of the library molecules determine how often a strand is sequenced within the given time. The higher the desired accuracy of the reads the shorter the reads should be. It depends on the project whether high accuracy reads or longer reads are more valuable. In *de novo* genome sequencing projects the length of the reads is of higher relevance to support genome assembly. In contrast, high-accuracy single consensus sequencing can be useful in metagenomic and especially in amplicon sequencing projects, as higher accuracy prevents an overestimation of biological diversity due to sequencing errors.

FUTURE SINGLE MOLECULE SEQUENCING TECHNOLOGIES

Nucleotide identification of currently available sequencing platforms is mostly based on optical systems that detect incorporation of fluorescently labeled nucleotides or reaction products during DNA synthesis. Future sequencing methods aim at real-time label-free sequencing, e.g., by direct analysis of the DNA molecule using electron microscopic techniques, scanning tunneling microscopy and spectroscopy, or analysis by Raman spectroscopy. Nanopore sequencing is another strategy that has gained much attention and has already been addressed in a couple of reviews (Bayley, 2006; Branton et al., 2008; Xu et al., 2009; Timp et al., 2010; Maitra et al., 2012). The different nanopore sequencing strategies that are under development enable individual base detection based on the measurement of conductivity changes across a lipid membrane while a DNA fragment is pulled through a nano-scale pore by an electric current. Conductivity changes are nucleotide-specific, enabling the identification of nucleotides as they traverse the pore. Biological nanopores are either constructed from engineered proteins, e.g., α-hemolysin (originally from *Staphylooccus aureus*) or MspA (*Mycobacterium smegmatis* porin A), or are entirely synthetic, e.g., graphene (Schadt et al., 2010; Thompson and Milos, 2011; Maitra et al., 2012). One of the major challenges in nanopore sequencing is reliable signal detection of each individual nucleotide at the high speed at which the DNA molecule traverses the pore and against a background of stochastic alterations in translocation rate (Branton et al., 2008; Morey et al., 2013).

As single molecule sequencing technologies do no longer depend on a PCR amplification step for signal detection, they overcome any bias introduced during emPCR or bridge PCR as well as dephasing problems (see Section Error Accumulation toward the End of Reads) that result in signal decay, which largely limits read length of current second generation instruments. These advantages come along with a higher sequencing error rate in individual reads, as errors cannot be compensated by the consensus read-out of clonal molecules in a cluster or on a bead. Future improvements of the sequencing technologies and the generation of consensus sequences, as explained for

the PacBio instrument, have the potential to compensate these errors.

SEQUENCING ERRORS

ESTIMATED ERROR RATES OF SECOND GENERATION SEQUENCING PLATFORMS

In comparison to Sanger sequencing, NGS technologies are known for higher error rates and different types of errors in the generated sequence reads. A direct comparison of error rates from different sequencing platforms and studies is difficult due to differences with regard to the sequenced sample material, the library preparation method, data filtering, and error calculation methods, and the fact that reads of different length (not necessarily the maximum possible length of a platform) are analyzed. Nevertheless, some values are compiled and provided as Table S1 for orientation. They are mostly in the range of 0.4–1% for Roche 454, Illumina and the Ion PGM platforms. Clear differences between these platform are not evident from the data. The quality of Ion PGM data, which is discussed quite controversially in the literature, is often slightly lower in direct comparison to Illumina and 454 platforms (Liu et al., 2012; Loman et al., 2012; Quail et al., 2012; Jünemann et al., 2013; Perkins et al., 2013). Read quality of HiSeq data was mostly reported to be slightly better compared to GAIIx data (Meacham et al., 2011; Minoche et al., 2011; Quail et al., 2012). The error profiles for the Illumins GA, HiSeq, and MiSeq instruments remain principally the same (Minoche et al., 2011; Quail et al., 2012). The quality of sequencing data from different 454 platforms appears to be similar. Likewise differences in dependence of the used chemistry or the analyzed library type (shotgun or amplicon) are not evident.

Substantial effort has been made to identify different types and sources of sequencing errors with the aim to reduce these either during the sequencing process or afterwards by applying improved analyses and correction algorithms. Some sequencing errors are observed on all sequencing platforms, while others are platform-specific. The following discussion about sequencing errors is largely focused on two sequencing platforms, 454 and Illumina, since error evaluation has been most intensively done for these platforms and these are the most frequently used platforms.

ERROR DISTRIBUTION WITHIN READS OF A LIBRARY

If the distribution of errors among 454 reads would be completely random, an error rate of 0.5% would mean that each read of 500 bp has on average 2.5 errors. But sequencing errors occur only in a certain percentage of reads; most studies report around 70% error-free reads (Huse et al., 2007; Kunin et al., 2010; Niu et al., 2010; Prabakaran et al., 2011; Niklas et al., 2013). Huse et al. (2007) observed that many of the erroneous reads in an amplicon dataset were characterized by the simultaneous presence of ambiguous base calls and explained this with multitemplated beads that carry similar library fragments.

In Illumina datasets, an increasing number of errors is observed in a successively decreasing number of reads (Dohm et al., 2008; Hillier et al., 2008; Nguyen et al., 2011). The percentage or error free reads was reported to be 57% for the GAIIx platform and 76% for the MiSeq platform in two available reports (Hillier et al., 2008; Quail et al., 2012). During paired end sequencing, the forward read was usually of slightly better quality than the reverse read (Quail et al., 2008; Minoche et al., 2011).

TYPES OF SEQUENCING ERRORS AND THEIR FREQUENCY

Insertions are the most frequent type of error during 454 sequencing (e.g., Margulies et al., 2005; Prabakaran et al., 2011; Vandenbroucke et al., 2011; Skums et al., 2012; Niklas et al., 2013). Several studies have reported deletions to be the second-most frequent type of error, followed by substitution errors (Huse et al., 2007; Gilles et al., 2011; Schloss et al., 2011; Niklas et al., 2013). The majority of indel errors occurs in homopolymeric regions (Margulies et al., 2005; Huse et al., 2007; Rozera et al., 2009; Kunin et al., 2010; Gilles et al., 2011; Shao et al., 2013). The longer the homopolymeric region, the higher the probability of an indel error and the lower the quality scores of the bases toward the end of this region (Quinlan et al., 2008; Luo et al., 2012b; Skums et al., 2012; Niklas et al., 2013). Indel errors are explained by the underlying sequencing principle. The preciseness of the proportionality of the detected light signal decreases with increasing number of identical bases (Margulies et al., 2005). Due to an analogous sequencing principle, the Ion PGM sequencer shows a similar error profile, dominated by indel errors in homopolymeric regions and clearly less substitution errors (Loman et al., 2012; Merriman et al., 2012; Bragg et al., 2013).

In contrast, substitution errors are the most frequent error type in Illumina sequencing (Dohm et al., 2008; Hillier et al., 2008; Hoffmann et al., 2009; Minoche et al., 2011; Nguyen et al., 2011) and for SOLiD sequencers (Shendure and Ji, 2008; Ratan et al., 2013). For the Illumina platform, Nguyen et al. (2011) identified 79–88% of all errors as substitution errors. Hillier et al. (2008) reported a 3.7-fold higher substitution error rate than indel error rate. Deletions are more frequent than insertions and insertions are likely to occur in homopolymeric regions (Dohm et al., 2008; Minoche et al., 2011). The lower rate of indel errors compared to 454 sequencing is achieved by the terminal blocking strategy during the sequencing process, which allows the incorporation of only one base per sequencing cycle, so that a homopolymeric region is sequenced base by base.

ERROR ACCUMULATION TOWARD THE END OF READS

Sequencing errors accumulate toward the end of reads, along with decreasing quality of the called bases. This is well known for Illumina reads, but has also been reported for 454 and Ion PGM data (Campbell et al., 2008; Lind et al., 2010; Schröder et al., 2010; Huse and Welch, 2011; Schloss et al., 2011; Loman et al., 2012; Bragg et al., 2013; Perkins et al., 2013). This accumulation of errors is the result of a decreasing signal-to-noise ratio during the sequencing process, which largely determines the maximum read length of all sequencing platforms.

Errors in 454 reads occur more likely beyond base 200–300 under FLX run conditions on the FLX and the GS Junior platform (Campbell et al., 2008; Gilles et al., 2011; Schloss et al., 2011; Niklas et al., 2013). In particular substitutions and ambiguous base calls accumulate (Gilles et al., 2011). Such an error profile is the result of a loss of synchronism during the sequencing

process on the multitemplated beads. Even though the basecalling software accounts for this artifact and reads are trimmed, it does not fully eliminate these effects (Margulies et al., 2005; Gilles et al., 2011). Another reason for a decreasing signal-to-noise ratio toward the end of a read is signal drooping due to premature termination of the sequencing process on templates. This was reported for Ion PGM sequencing (Merriman et al., 2012; Golan and Medvedev, 2013).

In Illumina reads, an accumulation of errors toward the end mainly affects long reads. It becomes obvious in the last third to fourth of 100 or 150 bp reads (Dohm et al., 2008; Claesson et al., 2010; Minoche et al., 2011; Nakamura et al., 2011; Liu et al., 2012). The result of this accumulation are lower overall quality values for longer reads. Also on Illumina platforms, the decreasing signal-to-noise ratio is largely a problem of signal dephasing during the sequencing process (Erlich et al., 2008; Kircher et al., 2009; Metzker, 2010; Schadt et al., 2010). Dephasing occurs when part of the clonal fragments in a cluster on the flow cell lag behind or are advanced compared to the overall sequencing procedure. The signal-to-noise ratio also decreases when the fluorescent label is not efficiently cleaved from the nucleotides added in the previous cycle (Dohm et al., 2008), and due to fluorescent dye decay during the sequencing process over several days (Kircher et al., 2009).

SEQUENCING ERROR CONTEXT DEPENDENCE

Substitution errors in Illumina reads were analyzed in more detail to identify possible error sources (Dohm et al., 2008; Meacham et al., 2011; Minoche et al., 2011; Nakamura et al., 2011; Nguyen et al., 2011; Abnizova et al., 2012; Luo et al., 2012b; Quail et al., 2012). Certain types of substitutions were found to occur more frequently than others and accumulate at specific positions. They are sequence context dependent, for instance after G-rich regions (Dohm et al., 2008; Minoche et al., 2011). Moreover, many substitution errors occur strand-specific, i.e., either predominantly in reads that cover a genomic region in forward direction or in those of reverse direction (Meacham et al., 2011; Nguyen et al., 2011). Such errors can be identified during data assembly or read mapping based on their strand-specificity and the fact that they are associated with low quality values for the respective erroneous base (Minoche et al., 2011). Abnizova et al. (2012) observed that the correct base was frequently detected with the second most intensive sequencing signal at erroneous positions, providing a possibility for correction. That errors tend to accumulate at specific positions within a genome was also observed for SOLiD data (Meacham et al., 2011).

EVENNESS OF READ COVERAGE AND GC BIAS

Early NGS studies already reported uneven read coverage when Illumina reads were mapped to existing genomes (Dohm et al., 2008; Hillier et al., 2008). The extent of this variation appears to vary largely from only 2- or 4-fold (Dohm et al., 2008; Minoche et al., 2011) to more than 100-fold (Harismendy et al., 2009). It can also occur in SOLiD, 454 and Ion PGM datasets (Suzuki et al., 2011; Meglecz et al., 2012; Merriman et al., 2012; Balzer et al., 2013; Gori et al., 2013; Ratan et al., 2013). In comparative studies, each platform produced a specific coverage pattern (Harismendy et al., 2009; Quail et al., 2012; Rieber et al., 2013). Depending on the coverage with which a sample is sequenced, this bias can result in gaps and affect quantitative assessments, e.g., in metagenomic or (meta)transcriptomic studies (Tariq et al., 2011; Gori et al., 2013).

A detailed analysis revealed an underrepresentation of reads in AT-rich regions (Bentley et al., 2008; Dohm et al., 2008; Hillier et al., 2008; Harismendy et al., 2009; Kozarewa et al., 2009; Minoche et al., 2011; Quail et al., 2012) and GC-rich regions (Bentley et al., 2008; Kozarewa et al., 2009; Quail et al., 2012; Ratan et al., 2013). It is the GC content of the complete library molecule and not only of the sequenced region that affects GC bias (Benjamini and Speed, 2012).

PCR steps were identified as a major cause introducing GC bias (Hillier et al., 2008; Aird et al., 2011; Quail et al., 2012). Standard Illumina and Ion PGM library preparation protocols include a PCR amplification step prior to bridge PCR or emPCR. To reduce GC bias, PCR free protocols have been developed for Illumina library construction (Kozarewa et al., 2009; Mamanova and Turner, 2011) and have meanwhile also been implemented in dedicated Illumina kits. Since PCR-free library preparation methods are problematic when the available input material is limited, PCR protocols were also optimized, as well as other library preparation steps that may introduce such bias (Van Dijk et al., 2014). High cluster densities on the Illumina flow-cell were also discussed to suppress GC-rich reads (Aird et al., 2011). Error correction algorithms were developed and can be applied to account for GC-bias in projects where quantitative information is inferred from the sequencing data such as transcriptomic studies (Hansen et al., 2010; Li et al., 2010; Benjamini and Speed, 2012).

DUPLICATE READS

Another artifact that has been reported in particular for 454 sequencing data is the occurrence of duplicate reads in shotgun (meta-)genomic sequencing projects. These start at the same base position and, depending on the strictness of the definition, are fully identical or different in only few positions and/or read length. Such sequence reads can be true duplicates that arise when genomic DNA is sequenced at very high coverage, or they are artificial duplicates. The source of this type of error is not fully known. It was speculated that duplicates are generated during emPCR, when amplified DNA is attaching to empty beads (Briggs et al., 2007). However, emPCR is also used to amplify library fragments during Ion PGM sequencing, but duplicate reads appeared not to be a major problem in one study in which this issue was specifically assessed (Bragg et al., 2013).

The analysis of several metagenomic sequencing projects revealed between 10 and 45% of duplicate reads (Gomez-Alvarez et al., 2009; Niu et al., 2010; Balzer et al., 2013). Duplicate reads can affect quantitative data analyses, e.g., species or gene abundance analyses in metagenomic studies. To identify and remove duplicates, software tools such as cd-hit-454 (Niu et al., 2010), 454 Replicate Filter (Gomez-Alvarez et al., 2009), PyroCleaner (Mariette et al., 2011), the duplicate removal tool of the GATK package (McKenna et al., 2010), or JATAC (Balzer et al., 2013) can be applied. Criteria that define artificial duplicates can be defined in such software tools. Nevertheless, some true duplicate

reads may also be eliminated by these filters. The percentage of true duplicates among all identified duplicates can vary largely between 2 and 72% (Niu et al., 2010).

REPRODUCIBILITY ACROSS RUNS AND BETWEEN REGIONS OR LANES

The overall reproducibility between 454 runs and samples from different regions of the picotiter plate is usually high (Vandenbroucke et al., 2011; Niklas et al., 2013). However, variation in error rates, in particular for indel errors, was seen between different 454 sequencing runs (Gilles et al., 2011; Prabakaran et al., 2011; Shao et al., 2013). Variation in terms of read composition of a sample may also occur, as observed in a study in which the same 16S rRNA gene PCR products were sequenced at different sequencing centers and in different runs (Schloss et al., 2011). A similarity analysis of the datasets revealed a clustering according to sequencing centers and, to lesser extent, to runs.

For Illumina, some studies report variation between runs and from lane to lane, e.g. with regard to sequencing errors (He et al., 2010; Aird et al., 2011; Nguyen et al., 2011; Chen et al., 2013), but also in this case it seems not to be a consistent problem (Abnizova et al., 2012; Benjamini and Speed, 2012). Nguyen et al. (2011) reported that variation with regard to sequencing errors largely diminished after data quality filtering. Highly reproducible results were also obtained in a study by Caporaso et al. (2012) across lanes and even on different platforms (i.e., HiSeq 2000 and MiSeq), showing that cross-platform data handling is possible (Bokulich et al., 2013).

It will depend on the project whether possible variation in sequencing performance is acceptable or will negatively affect results and conclusions. It can be a relevant issue when highly similar samples are comparatively analyzed, e.g., in amplicon sequencing projects. To identify method related variation in such critical studies, the inclusion of a standardized reference sample is highly recommended (Schloss et al., 2011; Bokulich et al., 2013).

SEQUENCING ERRORS OF THE PacBio RS SYSTEM

Sequencing errors of PacBio single reads are reported in the range of 13–20% (Thompson and Milos, 2011; Quail et al., 2012) but this high error rate can be reduced to 1% or less by CCS (Metzker, 2010). Sequencing errors on the PacBio system are mostly insertions and deletions (Eid et al., 2009). During single molecule sequencing, dephasing is not an issue, so that errors do not accumulate toward the end of the reads. Moreover, sequencing errors appear not to be sequence context specific (Carneiro et al., 2012; Koren et al., 2012) contributing to the high consensus accuracy that can be achieved when sequencing is done with high coverage (>20-fold) or by using the CCS strategy. Good performance was reported in difficult to sequence regions and GC-rich samples, resulting in more even coverage (Quail et al., 2012; Ross et al., 2013; Shin et al., 2013).

COMPENSATING AND CORRECTING SEQUENCING ERRORS

Once the types and sources of sequencing errors are known, different strategies and tools can be developed to compensate and correct errors. As a general strategy, accuracy is improved by sequencing with high coverage, usually 20- to 60-fold, depending on the sequencing purpose (Margulies et al., 2005; Voelkerding et al., 2009; Luo et al., 2012b). Also, the combination of sequencing data generated from different sequencing platforms with different error profiles was suggested and has been applied to identify and eliminate sequencing errors (Nakamura et al., 2011; Koren et al., 2012). These strategies are effective in *de novo* genomic sequencing and resequencing projects, but they are of limited use in metagenomic or metatranscriptomic studies that deal with biological variation. Each different read can represent a distinct genotype in such studies or is the result of a sequencing error. Sophisticated methods are needed to distinguish between natural sequence variation and sequencing errors in order not to overestimate diversity.

One way to reduce error rates is to apply alternative basecallers that show superior performance compared to the standard basecalling algorithms (e.g., Ledergerber and Dessimoz, 2011; Das and Vikalo, 2013; Golan and Medvedev, 2013). However, their application is often limited, as it comes along with a transfer of massive amounts of raw signal data from the sequencing service center to the customer and the need for high computational power to perform basecalling, in particular for large Illumina datasets.

In order to improve data quality after basecalling, filtering algorithms were developed. Such filters discard reads with low-quality bases or with uncalled/ambiguous bases, or they clip the lower quality 3′-ends of reads. Many of these filters use the information contained in quality values that are calculated for each base during the base calling process. Minoche et al. (2011) studied the effect of different filtering methods on Illumina data and could reduce the error rate to <0.2% by eliminating approximately 15–20% of the low-quality bases, mostly via 3′-end trimming. Nguyen et al. (2011) reported a 5-fold decrease of the error rate by applying a filter that eliminated reads with low quality bases (<Q30; i.e., with 0.1% likelihood of a false basecall), which resulted in a loss of 24–35% of sequence reads. It has to be kept in mind that low quality bases are to certain extent localized in specific regions of a genome. Discarding such reads can result in a more uneven coverage, introducing potential bias in quantitative studies (Minoche et al., 2011; Nakamura et al., 2011).

An alternative strategy to read clipping and exclusion of low quality reads is error correction. Several tools (e.g., Coral, HiTEC, Musket, Quake, RACER, Reptile, or SHREC) have been developed for this purpose, in particular for the correction of substitution errors in Illumina data (Ilie and Molnar, 2013; Liu et al., 2013; Yang et al., 2013). Some of these tools (Coral, HSHREC, KEC, and ET) have implemented indel correction algorithms and are thus suited for the analysis of 454 and Ion PGM data (Salmela, 2010; Salmela and Schröder, 2011; Skums et al., 2012). Error correction methods make use of the high sequence coverage in order to identify and correct errors. Moreover, most algorithms take into account the quality scores given for the individual bases and/or analyze the neighboring contextual sequence information. The application of error correction tools has been proven useful in *de novo* genome sequencing projects, resequencing and amplicon sequencing projects (e.g., Skums et al., 2012; Yang et al., 2013). At the same time, Yang et al. (2013) pointed out a need for improved algorithms, in particular for non-uniform data sets, such as metagenomic or (meta-)transcriptomic data. A strategy that can be applied in metagenomics studies to correct sequencing

errors is the generation of overlapping paired end reads that are assembled prior to further analyses (Zhou et al., 2011; Masella et al., 2012; Eren et al., 2013).

METAGENOMIC SEQUENCING OF THE PLANT ASSOCIATED MICROBIOTA

SEQUENCING AND ANALYSIS STRATEGIES FOR METAGENOMICS STUDIES

The optimal sequencing strategy for a metagenomics project will largely depend on the aim of the project. For a functional description of a microbial community, the Illumina HiSeq sequencing platform will be a good choice due to the low costs per sequenced base, which allows sequencing to high depth in order to gain as much information as possible, even from less-abundant microorganisms that may nevertheless play important roles for ecosystem functioning. Initially, the rather short read length of this platform was considered to be a critical issue (Wommack et al., 2008), but it appears that this is not necessarily a problem. A comparative study of a metagenomic analysis based on 454 and Illumina reads revealed that assembled data derived from both methods reflected the genomic composition of the sample equally well, with the Illumina dataset showing even a slightly better assembly result (using a 5-fold higher volume of data) (Luo et al., 2012b). Annotation of unassembled reads was slightly better for the longer 454 reads. In general, short reads will not allow the generation of a high number of large contigs, in particular for complex samples. As an example, assembly success for a metagenomic sample from the soybean phyllosphere microbiota, which showed medium complexity, was only moderate. The assembly of approximately 1 mio 454 reads with a mean read length of 235 bp resulted in 140,000 contigs with a mean length of 276 bp and left 30% of the reads unassembled. The largest contig had a length of 12,888 bp (Delmotte et al., 2009). In another study with datasets from complex freshwater microbial communities between 50 and 60% of 454 and Illumina reads remained unassembled (Luo et al., 2012b). Despite this moderate success, gene prediction or identification of protein domains is possible. This is even the case for unassembled short reads, though it becomes more difficult when no close homolog is present in the reference database (Scholz et al., 2012; Luo and Moran, 2013). Moreover, annotation of several million unassembled short reads can become a very time-consuming step, depending on the algorithm that is used.

An alternative to assembly and/or direct annotation of short sequence reads is the mapping of reads to existing genomes. The prerequisite for this strategy is that the genomes of the organisms of interest have been genome sequenced. This is currently still a limiting factor (Weinstock, 2011), although the entries in public databases are much more strongly growing since NGS technologies became available. Currently, there are nearly 3000 complete genome sequences of microorganisms deposited in the NCBI database and genomic information of approximately 16,000 microorganisms is available as scaffolds or contigs. It can be a very valuable step to enrich, isolate and sequence the dominant community members, as it is for instance done in the Human Microbiome Project (Turnbaugh et al., 2007), or was already done for 21 bacterial isolates from the *Populus* rhizosphere (Brown et al., 2012). Such attempts will be of value for diverse studies of plant associated microorganisms, as the plant associated microbiota appears to show certain degree of consistency in terms of colonizing taxa (Bulgarelli et al., 2012; Lundberg et al., 2012; Vorholt, 2012), so that stains sequenced in one study may support data analysis of another study using plants grown under different conditions or even different model plants. Thus, the generation of further individual genome sequences will improve data analysis of future metagenomics, metatransriptomics, and metaproteomics studies of plant-associated microorganisms.

As several microbial taxa remain unculturable, some metagenomic studies aim at the reconstruction of individual genomes to obtain information from these organisms. In such studies sequence read assembly is a key step and challenging due to the complexity and uneven composition of microbial communities (Scholz et al., 2012). Assembly will be most successful if the complexity of the microbial community is rather low and dominated by one or a few phylogenetically distinct bacterial taxa. Different studies have meanwhile demonstrated that genome reconstruction of individual members in metagenomic samples is possible, even when rather short Illumina reads are generated (Mackelprang et al., 2011; Albertsen et al., 2013).

Assembly success also depends on sequence read length and the coverage with which the genome(s) of interest are sequenced (Kunin et al., 2008; Schatz et al., 2010; Weinstock, 2011; Luo et al., 2012a); parameters that can be considered in the design of the sequencing strategy. In an *in silico* study, Luo et al. (2012a) demonstrated that a 20-fold coverage was sufficient to reconstruct the genome of a dominant member in a metagenomic sample and that a higher coverage did not substantially improve the assembly result. Strategies that are frequently applied in pure culture genome sequencing projects to improve assembly are the inclusion of longer reads, paired end reads or reads from mate pair libraries (Schatz et al., 2010). This strategy can also be useful in metagenomic sequencing projects. The combination of sequencing data from different platforms that generate reads of different lengths and with different error profiles was reported multiple times as a successful strategy to improve genome assembly of individual bacterial strains (Aury et al., 2008; Reinhardt et al., 2009; Koren et al., 2012). In particular the PacBio instrument holds potential to fulfill the need for long reads in order to bridge larger gaps or repetitive regions (English et al., 2012; Mavromatis et al., 2012). These strategies have not yet been widely applied in metagenomics projects, but it appears likely that they are of value (Niedringhaus et al., 2011).

Assemblies may also be improved by using new assembly strategies, e.g., a nested strategy, in which the short reads are assembled to longer reads in a first step, before those are further assembled. The *in silico* generation of Sanger-like reads from Illumina reads by filling the gaps between paired end reads can be done by searching for reads within the same library that fill the gap between a read pair or by constructing paired end libraries of successively decreasing insert length, which are searched for suitable paired end reads to close the gaps between those paired end reads that are contained in the library with the largest library molecules (Rodrigue et al., 2010; Nadalin et al., 2012; Ruan et al., 2013). This strategy may be of particular help to fill small gaps, i.e., of a distance smaller than the size of the largest library

molecules, but will not help to bridge repetitive regions that are larger than the largest library molecules.

BIOINFORMATICS TOOLS FOR METAGENOMIC DATA ANALYSIS

The massive amount of sequence data that are generated in metagenomic projects demand new and efficient computational methods for data processing, analysis, and storage (Pop and Salzberg, 2008; Tautz et al., 2010). Substantial progress has been made in this field, as evident from the many different tools that are meanwhile available, e.g., for sequence read assembly, read mapping, or gene prediction (for an overview of available tools see for instance Voelkerding et al., 2009; Guazzaroni and Ferrer, 2011; Zhang et al., 2011; Thomas et al., 2012). New tools become available that are specifically designed for the analysis of metagenomic data, including assemblers such as MetaVelvet or Meta-IDBA (Peng et al., 2011; Namiki et al., 2012), annotation tools such as MG-RAST or CAMERA (Glass et al., 2010; Sun et al., 2011), tools for read mapping and alignment and for further data analysis, e.g., taxon identification and analysis of the microbial community composition based on phylogenetic marker genes (e.g., Stark et al., 2010; Scholz et al., 2012; Sunagawa et al., 2013). It would go beyond the scope of this review to discuss the diverse options for the analysis of metagenomic data along with the available software tools. Several recent reviews have addressed this aspect in detail (Kunin et al., 2008; De Filippo et al., 2012; Hunter et al., 2012; Logares et al., 2012; Scholz et al., 2012; Teeling and Glöckner, 2012; Davenport and Tümmler, 2013; Kim et al., 2013; Luo et al., 2013; Preheim et al., 2013; Segata et al., 2013).

Not only powerful software tools are required for the analysis of NGS data, but also high-performance computing capacity, in particular for large metagenomics datasets. This may pose a problem to research laboratories that are not specialized on NGS data analysis. Cloud computing, i.e., the rental of processing time on a computer cluster on demand over a network, is discussed and developing as a possible solution to this problem (Angiuoli et al., 2011; Wilke et al., 2011; Zhang et al., 2011; Dai et al., 2012; Nagasaki et al., 2013), though it has to be considered that this is often not free of costs and may pose security issues related to data transfer (Angiuoli et al., 2011; Hunter et al., 2012).

TARGETED GENE SEQUENCING OF AMPLICONS FROM METAGENOMIC DNA

SELECTING THE APPROPRIATE SEQUENCING STRATEGY FOR AMPLICON SEQUENCING

Targeted sequencing approaches of metagenomic DNA are mostly applied to identify the members of microbial communities or to compare their composition in different samples. Diversity studies are usually based on the 16S rRNA gene as bacterial marker and 18S rRNA or ITS as fungal markers (Table S2), while functional marker genes are analyzed when microorganisms with specific metabolic functions such as chitin degradation are addressed (Cretoiu et al., 2012). Until now the fast majority of amplicon sequencing studies have been performed using 454 technology (Table S2), mostly due to the fact that this was the first available NGS platform and due to the relatively long reads, that can be obtained from this platform. However, a shift toward the Illumina platform is currently noticable. First studies were already performed on the GAIIx platform with 76 bp paired end reads and later on with longer paired end reads up to 150 bp, followed by analysis on the HiSeq instrument and recently also on the MiSeq platform (Claesson et al., 2010; e.g., Gloor et al., 2010; Hummelen et al., 2010; Caporaso et al., 2011, 2012; Jogler et al., 2011; Degnan and Ochman, 2012; Kozich et al., 2013; Bokulich et al., 2014). The generation of overlapping paired end reads is recommended on these platforms as it will help to minimize the error rate (Eren et al., 2013; Kozich et al., 2013). As outlined above, errors accumulate toward the end of the reads, so that they can be corrected if consensus reads are generated from the read pairs. In particular the MiSeq instrument is a suitable platform for such studies, as it produces reads with a length comparable to those of the first 454 instruments, but at much lower costs. The read number obtained from MiSeq runs will in many cases be sufficient to obtain a sequencing depth that allows to answer a research question. In a few studies, the Ion Torrent PGM was used to analyze bacterial or fungal communities based on reads with a length of approximately 100 or 200 bp (Whiteley et al., 2012; Kemler et al., 2013). Longer reads are meanwhile possible on this sequencer and a protocol for paired end sequencing is available (though not yet officially supported by the company), so that this platform can be an alternative to the previously mentioned systems for amplicon sequencing.

The taxonomic resolution that is achieved with reads from these sequencers is clearly lower compared to Sanger reads. Nearly full length 16S rRNA gene sequences were Gold standard for clone library analysis based on Sanger reads and have led to the comprehensive sequence databases we have today. They enable species differentiation and often even the distinction of different strains. In contrast, the short NGS reads provide a resolution at maximum down to genus level. It turned out that this is frequently sufficient, in particular if the method is used for comparative purposes and microbial communities in the samples of interest do not contain many closely related species. Compared to clone library analysis, DGGE or T-RFLP, NGS amplicon sequencing allows analysis at greater depth so that many more low-abundant taxa can be detected. Thus, despite the lower taxonomic resolution, sensitivity of the method is reached here due to sequencing depth. It is up to the researcher to decide which information, resolution of taxa or sequencing depth will be more important for a project.

In case taxon resolution is important, sequence information of longer reads is needed, and the Roche 454 sequencer is a better choice. With the latest software update to version 2.9, amplicon sequencing is supported under FLX+ run conditions. Under these conditions, 16S rRNA gene and ITS sequence reads with a mean length of 650 and 750 bp were obtained (Perazzolli et al., 2014). Even longer amplicons can be sequenced when using the PacBio RS platform. A recent study demonstrated the feasibility of amplicon sequencing for community analysis on this platform (Marshall et al., 2012; Fichot and Norman, 2013), although another study reported higher error rates for PacBio amplicon sequence reads compared to 454 reads of equal length, despite that fact that the CCS strategy was used (Mosher et al., 2013). Rather short movies of only 45 min were recorded in that study. By increasing the recording time higher quality sequences can be

obtained. The current release of new sequencing chemistry and future improvements will enable the generation of higher quality sequences that will probably allow resolution even below genus level.

SEQUENCE READ ANALYSIS OF AMPLICON DATA

Diverse tools have been developed specifically for the analysis of amplicon data derived from metagenomic DNA, in particular for 454 data. This is largely due to the fact that many projects aim at an estimation of the microbial diversity within samples and along with this the indispensable need to differentiate between true diversity and sequencing errors (Sogin et al., 2006; Quince et al., 2009; Kunin et al., 2010). The fact that amplicon sequencing on NGS platforms is more and more widely applied has expedited the development of specific data analysis tools.

Based on the initial findings of Huse et al. (2007), who reported an accumulation of errors within a rather small subset of 454 reads, it became common to discard reads with one or more errors in the index and the target gene specific primer region. Likewise, reads with ambiguous basecalls (Ns), of unexpected length, with low quality scores or those that cannot be aligned to the gene of interest are assumed to be unspecific PCR products and are often removed (Huse et al., 2007, 2010; Kunin et al., 2010; Huse and Welch, 2011; Schloss et al., 2011; Zhou et al., 2011). Read trimming based on quality scores has also been applied to improve quality of 454 and Illumina data (Kunin et al., 2010; Caporaso et al., 2011; Schloss et al., 2011; Bokulich et al., 2013). In some studies singletons, i.e., sequence reads that occur only once, are removed from the datasets to further reduce the error rate (Caporaso et al., 2011; Shade et al., 2013).

Besides this quality filtering, specific algorithms are applied to improve quality. These aim at the correction of errors and the selection of representative sequence reads (=denoising), so that the number of reads or bases is not further decreased. The methods are based on the assumption that erroneous reads are representatives of more abundant error-free reads. Representative error free reads are identified and selected based on comparative sequence analysis, e.g., in the single-linkage preclustering (SLP) approach of Huse et al. (2010) or by the Pyrotagger tool (Kunin and Hugenholtz, 2010). Denoising algorithms such as PyroNoise, its successor AmpliconNoise or the DeNoiser analyze 454 flow grams (Quince et al., 2009; Reeder and Knight, 2010; Quince et al., 2011). The latter two algorithms have been reported to be very efficient, but demand much computational power, which has limited their application (Quince et al., 2011; Bragg et al., 2012). The SeqNoise algorithm, implemented in the software package Mothur, is less computationally demanding and therefore more often used. In comparative studies, the AmpliconNoise algorithm performed very well for OTU estimation (Quince et al., 2011; Bragg et al., 2012; Gaspar and Thomas, 2013). Critical analyses of different denoising tools demonstrated that parameters have to be chosen very carefully in order not to introduce bias by read modification during the generation of representative consensus reads. Default settings did not necessarily provide the best results (Bragg et al., 2012; Gaspar and Thomas, 2013).

The identification and elimination of chimeric sequences is another type of error that needs to be accounted for. Chimeric sequences originate during PCR and have been reported to contribute between 5 and 45% of a PCR product (Lahr and Katz, 2009; Haas et al., 2011). Available algorithms to eliminate these artifacts are Perseus, which was developed together with AmpliconNoise (Quince et al., 2011), ChimeraSlayer (Haas et al., 2011), or UCHIME (Edgar et al., 2011). While ChimeraSlayer needs a chimera-free reference database for chimera detection, Perseus is used without reference database. UCHIME offers both options and was reported to be faster compared to the other two methods (Edgar et al., 2011). UCHIME performed best in a comparative study when a reference database was used. Without reference database, UCHIME and Perseus performed equally well (Schloss et al., 2011). Considering that the use of database-independent methods is not limited by the quality and diversity of data in the reference database, database-free methods may be preferred.

Not all tools can be applied to Illumina datasets, for instance denoising algorithms that use 454 flow grams as input data. Moreover, some tools are computationally too demanding to be used for large Illumina datasets. A specific quality filtering approach for Illumina data was recently described using the "Quantitative Insights Into Microbial Ecology" (QIIME) toolkit (Bokulich et al., 2013). Other packages that combine the above mentioned analysis steps for error reduction with further analyses such as OTU clustering, taxonomy assignment or multiple sample comparison, are Mothur or the UPARSE pipeline (Caporaso et al., 2010; Schloss et al., 2011; Edgar, 2013).

APPLICATION OF NGS TECHNOLOGIES IN PRESENT STUDIES OF PLANT ASSOCIATED MICROORGANISMS

SHOTGUN METAGENOMIC STUDIES

Until today, only a limited number of shotgun metagenomic studies of plant associated microorganisms exist (**Table 3**). Most of the studies are based on Roche 454 sequencing technology and generated a few hundred Mb of sequence data. In a very recent study of Mendes et al. (2014) the epiphytic rhizosphere microbiome of soybean was compared to that in bulk soil with regard to taxonomic and functional composition. A specific rhizosphere microbiota was observed, representing a subset of the taxonomic and functional diversity present in bulk soil. Moreover, functions that may be of benefit for the plant in terms of growth promotion and nutrition were detected, likewise as in a study of Sessitsch et al. (2012), who performed the first extensive metagenomic study of plant associated microorganisms, still using Sanger sequencing technology. In two other rhizosphere studies, the genomic basis for phosphorous acquisition was addressed. Unno and Shinano (2013) analyzed the rhizosphere metagenome of plants that showed enhanced growth in the presence of phytic acid and detected genes encoding enzymes related to phytic acid utilization such as alkaline phosphatase or citrate synthase. Chhabra et al. (2013) applied a targeted metagenomic approach by constructing a fosmid library in *Escherichia coli*, which was screened in an assay for mineral phosphate solubilization activity. Six positive clones were shotgun sequenced using 454 technology. Genes and operons with homology to phosphorous uptake systems, regulatory, and solubilization mechanisms were identified.

Table 3 | Metagenomic studies based on NGS technology that target the plant-associated microbiota.

Sequencing technology	Sequencing statistics	Plant compartment	Plant species and type of sample	Major findings	References
Roche 454	3.2 million raw reads 2,472,359 filtered reads Mean read number per sample 103,014 Mean read length 523 bp	Rhizosphere	Soybean (*Glycine max*) rhizosphere and bulk soil samples taken from mesocosm experiments with soil from soybean fields in Brazil	The rhizosphere community is selected from the bulk soil based on functions related to N, Fe, P, and K metabolism	Mendes et al., 2014
Roche 454	Not specified	Rhizosphere	Barley rhizosphere samples collected from an experimental field in Ireland with 15 years of barley monoculture under low-input mineral management regime	Identification of genes and operons involved in mineral phosphate solubilization in the rhizosphere	Chhabra et al., 2013
Illumina Miseq	15 million paired end reads 2.6 Gbp	Phyllosphere	Samples from *Salmonella* enrichment cultures from outdoor grown tomato (*Solanum lycopersicum*) and tomato leaves and fruits	Differences in metagenomic composition of replicate phyllosphere enrichment cultures; enrichment of *Paenibacillus* on *Salmonella*-selective media	Ottesen et al., 2013a
Roche 454	Not specified	Phyllosphere Rhizosphere	Leaves, stems, roots, flowers, and fruits from outdoor grown tomato (*S. lycopersicum*)	Distinct microbial communities detected on different tomato plant organs	Ottesen et al., 2013b
Roche 454	8445 and 3799 filtered reads Mean read length 228 and 226 bp	Rhizosphere	Rhizosphere samples from greenhouse grown *Lotus japonicus*; plants of the same age but two different developmental stages grown in presence of phytic acid	Differences in microbial community composition in the rhizosphere of the differently developed plants; identification of genes related to phytic acid utilization	Unno and Shinano, 2013
Roche 454	448 Mb sequence data Mean read length 357 bp	Phyllosphere	Leaf samples of tamarisk (*Tamarix nilotica*); datasets from soybean, (*G. max*), *Arabidopsis thaliana*, clover (*Trifolium repens*), and rice (*Oryza sativa*) included in analyses (Delmotte et al., 2009; Knief et al., 2012; Vorholt, 2012)	Diverse microbial rhodopsins detected in phyllosphere bacteria Detection of genes encoding proteins involved in anoxygenic photosynthesis (*bchY*, *pufM*, and *pufL*)	Atamna-Ismaeel et al., 2012b Atamna-Ismaeel et al., 2012a
Roche 454	832 and 396 Mb of sequence data per sample	Phyllosphere Rhizosphere	Phyllosphere and rhizosphere sample of field grown rice (*O. sativa*), Philippines	Contrasting proteome patterns in phyllosphere and rhizosphere of rice	Knief et al., 2012
Roche 454	1,109,816 reads 260 Mb of sequence data 235 bp mean read length	Phyllosphere	Leaf samples from field grown soybean (*G. max*), Switzerland	High consistency in the microbial community composition and their proteomes on different host plants	Delmotte et al., 2009
Roche 454	419,571 reads 216 bp mean read length 90,813,125 bp of sequence data	(Phyllosphere)	Psyllid infected with the endophyte "*Candidatus* Liberibacter asiaticus"	Complete genome sequence of the uncultured plant pathogen and insect symbiont "*Candidatus* Liberibacter asiaticus"	Duan et al., 2009

Metagenomic data of phyllosphere associated microbial communities are available from soybean, rice, clover, *Arabidopsis thaliana, Tamarix*, and tomato (Delmotte et al., 2009; Atamna-Ismaeel et al., 2012a; Knief et al., 2012; Ottesen et al., 2013b). Some of these datasets were analyzed in combination with metaproteomic data obtained from the same sampling material (Delmotte et al., 2009; Knief et al., 2012). These analyses revealed high consistency in the metaproteomes of phyllosphere bacteria from different plant species. In agreement, microbial community composition as inferred from these phyllosphere

metagenomic datasets revealed consistency in microbial community composition at phylum level (Vorholt, 2012). Comparative analyses of metagenomic and metaproteomic data of rice phyllosphere and rhizosphere samples revealed a higher complexity of the rhizosphere microbiota and a clearly distinct metagenomic and -proteomic composition (Knief et al., 2012). The phyllosphere metagenomic datasets generated in these studies were further used in combination with a metagenomic dataset from *Tamarix* associated phyllosphere bacteria to screen for photosynthetic genes that are known from other microorganisms to be involved in light-driven energy generation (Atamna-Ismaeel et al., 2012a,b).

Another kind of metagenomic project was performed with the aim to obtain a complete sequence of an unculturable plant pathogen, "*Candidatus* Liberibacter asiaticus," which causes citrus huanglongbing (Duan et al., 2009). This pathogen is transmitted by phloem-feeding insects. Metagenomic DNA was extracted from a single Asian citrus psyllid and not from an infected plant, due to the fact that the natural enrichment of the target organism is higher in the insect. Extracted DNA was subjected to multiple displacement amplification prior to sequencing using 454 technology. Sequence read assembly resulted in 38 contigs for "*Candidatus* L. asiaticus," which were identified by PCR confirmation reactions from a total of 1475 generated contigs. Gap closure was achieved by sequencing gap bridging PCR products. Genome analysis revealed a heavily reduced genome of this highly divergent member of the family Rhizobiacea, as it is seen frequently for microorganisms with a predominantly intracellular lifestyle.

AMPLICON SEQUENCING STUDIES

NGS technologies are increasingly often used for amplicon sequencing of bacterial and fungal marker genes in order to characterize the communities in the phyllosphere and rhizosphere. There are more than 100 rhizosphere and at least 37 phyllosphere articles published until now that have used these techniques (see Supplementary Material for a compilation of studies). The fast majority of these studies applied Roche 454 sequencing technology. Only few used the Ion PGM platform (Kavamura et al., 2013; Kemler et al., 2013; Yergeau et al., 2014) or the Illumina MiSeq (Jiang et al., 2013). A detailed look at the phyllosphere studies (Table S2) reveals that the generated read numbers in amplicon studies are mostly in a range from a few thousand to ten thousand reads per sample (Table S2). The obtained read length increased successively over time, along with the development of the Roche 454 sequencing platform. With the 454 FLX+ instrument a mean read length of 750 bp was recently obtained for 16S rRNA gene amplicons (Perazzolli et al., 2014).

NGS amplicon sequencing was so far almost exclusively applied for the analysis of bacterial or fungal communities. Bacterial phyllosphere communities were studied based on the 16S rRNA gene without a preference for one specific region within this gene (Table S2). Fungal communities were mostly analyzed based on the ITS region. The only functional marker gene that has been studied so far in plant associated microorganisms via amplicon sequencing is *chiA*, encoding a chitinase (Cretoiu et al., 2012). The aim of that particular study was an assessment of *chiA* gene diversity in different habitats, including rhizosphere samples from two arctic plant species. Analysis revealed that the rhizosphere of *Oxyria digyna* was among the samples with the highest *chiA* diversity.

Most amplicon sequencing studies in the phyllosphere were performed to describe and understand plant colonization by microorganisms. In particular biogeographic patterns, the role of the plant taxon for shaping communities and the temporal succession of the microbiota were addressed (e.g., Redford et al., 2010; Rastogi et al., 2012; Bokulich et al., 2014; Maignien et al., 2014). Also differences in the colonization of different plant compartments were analyzed (Bodenhausen et al., 2013; Ottesen et al., 2013b). The impact of specific treatments during plant cultivation such as irrigation were also addressed in some studies (Williams et al., 2013).

Amplicon sequencing projects performed in the rhizosphere addressed basically the same questions, i.e., aspects of biogeographical dispersal of rhizosphere microorganisms, or the impact of factors such as season, host plant species, soil type, or plant growth conditions (Gottel et al., 2011; Lundberg et al., 2012; Navarrete et al., 2013; Peiffer et al., 2013; Zhang et al., 2013). A major additional focus of rhizosphere studies is the analysis of endo- and ectomycorrhiza (Lumini et al., 2010; Dumbrell et al., 2011; Yu et al., 2012). It has become clear that the plant plays a significant role in shaping the associated microbiota and that root exudates are involved in this process (Badri et al., 2013), but to better understand how plants affect this process, plant mutant strains altered in root exudation or, in case of the phyllosphere with altered leaf surface properties, were analyzed (Badri et al., 2009; Reisberg et al., 2013). Furthermore, aspects of bioremediation, disease suppressiveness or possible impacts of herbicide application or of genetically modified plants have been addressed in rhizosphere studies (Barriuso et al., 2010; Rosenzweig et al., 2012; Dohrmann et al., 2013; Bell et al., 2014). All these exemplarily selected publications demonstrate the usefulness of NGS amplicon sequencing projects for studying microbial plant colonization. Future studies in this field will lead to an even better understanding of the factors that determine microbial plant colonization.

TRANSCRIPTOMIC AND METATRANSCRIPTOMIC STUDIES

NGS technologies have not only stimulated research in the field of (meta-)genomics, but are also excellent tools to perform (meta-)transcriptomic analyses. The appearance of these technologies has boosted transcriptomic studies of plant associated microorganisms, until now in particular of pathogenic fungi (e.g., Tremblay et al., 2012; Weßling et al., 2012; Thakur et al., 2013). Both, Illumina and 454 technology have been used in such studies. NGS is of particular advantage when the organisms of interest have not been genome sequenced, which is a prerequisite for the alternative microarray analyses. In some studies, the transcriptome of the host and the pathogen were even analyzed in parallel (e.g., Fernandez et al., 2012; Zhuang et al., 2012). The success of such parallel analyses depends on the ratio of plant to fungal mRNA in the sequenced sample.

First metatranscriptomic studies of the whole plant associated microbial communities appeared just recently. Chaparro

et al. (2014) analyzed the microbial metatranscriptome of the *Arabidopsis thaliana* rhizosphere at different plant development stages. They observed that microbial genes involved in metabolism of carbohydrates, amino acids and secondary metabolites changed over time in correspondence to root exudate patterns, which also changed over time. Yergeau et al. (2014) compared the microbial metatranscriptomic composition in the rhizosphere of willow with that in bulk soil in soils contaminated with organic pollutants. Different genes involved in hydrocarbon degradation were expressed in rhizosphere and bulk soil microbial communities. Genes related to carbon and amino-acid uptake and utilization were in general up-regulated in the rhizosphere.

Instead of an mRNA analysis, Turner et al. (2013) performed rRNA sequencing to characterize the active microbiota in the rhizosphere of different crops (wheat, oat, pea). Analyzing microbial communities based on rRNA instead of their rRNA genes is assumed to reflect the physiologically active microbiota in a sample and does not necessarily need extensive PCR amplification of the target molecules prior to library preparation, as demonstrated in that study. Clear differences were observed in the composition of the active prokaryotic and eukaryotic communities compared to bulk soil samples and between the different plant species. A strong response in the fungal community to plant produced anti-fungal avenacins was observed in the rhizosphere.

APPLICATION OF NGS TECHNOLOGIES IN FUTURE METAGENOMICS STUDIES WILL ADVANCE UNDERSTANDING IN PLANT-MICROBE ASSOCIATIONS

With the availability of second generation sequencing platforms many of the limitations metagenomic studies had to deal with at the time when Sanger sequencing was the predominant technology have been overcome. In particular the preparation of metagenomic/sequencing libraries can be done much faster and the sequencing costs per base are drastically reduced. The new technologies allow much deeper sequencing of microbial communities, providing more information about identity and physiological potential of microbial communities associated to plants. Limitations of NGS approaches such as shorter reads and higher sequencing error rates can be largely compensated by using specifically designed sequence data analyses methods. Future developments of the sequencing technology will enable us to obtain even more and longer reads; the generation of sequence information will thus most likely not be a limiting factor in future studies, but enable to address the open questions in phyllosphere and rhizosphere research, as outlined in the introduction, in even more detail.

A current limitation of metagenomic sequencing studies is a high ratio of sequences that represent unknown genes of known or unknown organisms, and of sequences for which no homolog is found in public databases that would enable to infer further information. To improve the still challenging task of linking genes and thus function to phylogeny, genomic sequencing of representative pure cultures and the genetic and physiological characterization of strains will remain an important task. Genome sequencing projects of strain collections from the ecosystems of interest are one step further to overcome this limitation (Turnbaugh et al., 2007; Brown et al., 2012). Concerted sequencing of currently underrepresented organisms in databases, e.g., based on evolutionary relationship as in the GEBA project, will further improve databases (Wu et al., 2009). Likewise, advance in single cell genome sequencing has recently enabled the sequencing of yet uncultivated microorganisms; 200 bacterial and archaeal cells representing diverse largely uncharacterized phyla were successfully sequenced (Rinke et al., 2013). This genomic information will enable a more specific assignment of metagenome reads to taxa. (Meta-)transcriptomic and -proteomic studies based on known and well characterized representative model organisms under controlled conditions will contribute to a deeper understanding of microbial life in the phyllosphere.

The complementation of metagenomics data with metatranscriptomic, metaproteomic, and (meta-)metabolomic data will be one of the future goals to obtain a more complete view of the activities and the physiological potential of plant associated microbial communities under given conditions at systems level (Zhang et al., 2010; Knief et al., 2011; Segata et al., 2013). Such information is inevitable to build up models that can explain and predict microbially mediated processes and interactions in the phyllosphere and rhizosphere under different environmental conditions, including agricultural practices, responses to pathogen attack and disease, or to climate change.

ACKNOWLEDGMENTS

I thank Richard Reinhardt (Max Planck Genomecentre Cologne, Germany) for constructive discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Journal/10.3389/fpls.2014.00216/abstract

REFERENCES

Abnizova, I., Leonard, S., Skelly, T., Brown, A., Jackson, D., Gourtovaia, M., et al. (2012). Analysis of context-dependent errors for Illumina sequencing. *J. Bioinform. Comput. Biol.* 10:1241005. doi: 10.1142/S0219720012410053

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* 11:R119. doi: 10.1186/gb-2010-11-12-r119

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12:R18. doi: 10.1186/gb-2011-12-2-r18

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579

Angiuoli, S. V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D. R., et al. (2011). CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12:356. doi: 10.1186/1471-2105-12-356

Atamna-Ismaeel, N., Finkel, O., Glaser, F., Von Mering, C., Vorholt, J. A., Koblizek, M., et al. (2012a). Bacterial anoxygenic photosynthesis on plant leaf surfaces. *Environ. Microbiol. Rep.* 4, 209–216. doi: 10.1111/j.1758-2229.2011.00323.x

Atamna-Ismaeel, N., Finkel, O. M., Glaser, F., Sharon, I., Schneider, R., Post, A. F., et al. (2012b). Microbial rhodopsins on leaf surfaces of terrestrial plants. *Environ. Microbiol.* 14, 140–146. doi: 10.1111/j.1462-2920.2011.02554.x

Aury, J. M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., et al. (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* 9:603. doi: 10.1186/1471-2164-9-603

Badri, D. V., Chaparro, J. M., Zhang, R., Shen, Q., and Vivanco, J. M. (2013). Application of natural blends of phytochemicals derived from the root exudates of *Arabidopsis* to the soil reveal that phenolic-related compounds predominantly modulate the soil microbiome. *J. Biol. Chem.* 288, 4502–4512. doi: 10.1074/jbc.M112.433300

Badri, D. V., Quintana, N., El Kassis, E. G., Kim, H. K., Choi, Y. H., Sugiyama, A., et al. (2009). An ABC transporter mutation alters root exudation of phytochemicals that provoke an overhaul of natural soil microbiota. *Plant Physiol.* 151, 2006–2017. doi: 10.1104/pp.109.147462

Balzer, S., Malde, K., Grohme, M. A., and Jonassen, I. (2013). Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics* 29, 830–836. doi: 10.1093/bioinformatics/btt047

Barriuso, J., Marin, S., and Mellado, R. P. (2010). Effect of the herbicide glyphosate on glyphosate-tolerant maize rhizobacterial communities: a comparison with pre-emergency applied herbicide consisting of a combination of acetochlor and terbuthylazine. *Environ. Microbiol.* 12, 1021–1030. doi: 10.1111/j.1462-2920.2009.02146.x

Bayley, H. (2006). Sequencing single molecules of DNA. *Curr. Opin. Chem. Biol.* 10, 628–637. doi: 10.1016/j.cbpa.2006.10.040

Bell, T. H., Hassan, S. E., Lauron-Moreau, A., Al-Otaibi, F., Hijri, M., Yergeau, E., et al. (2014). Linkage between bacterial and fungal rhizosphere communities in hydrocarbon-contaminated soils is related to plant phylogeny. *ISME J.* 8, 331–343. doi: 10.1038/ismej.2013.149

Benjamini, Y., and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72. doi: 10.1093/nar/gks001

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Berry, D., Ben Mahfoudh, K., Wagner, M., and Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl. Environ. Microbiol.* 77, 7846–7849. doi: 10.1128/AEM.05220-11

Bodenhausen, N., Horton, M. W., and Bergelson, J. (2013). Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS ONE* 8:e56329. doi: 10.1371/journal.pone.0056329

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Bokulich, N. A., Thorngate, J. H., Richardson, P. M., and Mills, D. A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proc. Natl. Acad. Sci. U.S.A.* 111, E139–E148. doi: 10.1073/pnas.1317377110

Bowman, S. K., Simon, M. D., Deaton, A. M., Tolstorukov, M., Borowsky, M. L., and Kingston, R. E. (2013). Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* 14:466. doi: 10.1186/1471-2164-14-466

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., and Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* 9:e1003031. doi: 10.1371/journal.pcbi.1003031

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., and Tyson, G. W. (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat. Methods* 9, 425–426. doi: 10.1038/nmeth.1990

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., et al. (2008). The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26, 1146–1153. doi: 10.1038/nbt.1495

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14616–14621. doi: 10.1073/pnas.0704665104

Brown, S. D., Utturkar, S. M., Klingeman, D. M., Johnson, C. M., Martin, S. L., Land, M. L., et al. (2012). Twenty-one genome sequences from *Pseudomonas* species and 19 genome sequences from diverse bacteria isolated from the rhizosphere and endosphere of *Populus deltoides*. *J. Bacteriol.* 194, 5991–5993. doi: 10.1128/JB.01243-12

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren Van Themaat, E., Ahmadinejad, N., Assenza, F., et al. (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* 488, 91–95. doi: 10.1038/nature11336

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren Van Themaat, E., and Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* 64, 807–838. doi: 10.1146/annurev-arplant-050312-120106

Buschmann, T., and Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* 14:272. doi: 10.1186/1471-2105-14-272

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., et al. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13081–13086. doi: 10.1073/pnas.0801523105

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080107

Carlsen, T., Aas, A. B., Lindner, D., Vralstad, T., Schumacher, T., and Kauserud, H. (2012). Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol.* 5, 747–749. doi: 10.1016/j.funeco.2012.06.003

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., and Depristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 13:375. doi: 10.1186/1471-2164-13-375

Caruccio, N. (2011). Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition. *Methods Mol. Biol.* 733, 241–255. doi: 10.1007/978-1-61779-089-8_17

Chaparro, J. M., Badri, D. V., and Vivanco, J. M. (2014). Rhizosphere microbiome assemblage is affected by plant development. *ISME J.* 8, 790–803. doi: 10.1038/ismej.2013.196

Chen, Y. C., Liu, T. L., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS ONE* 8:e62856. doi: 10.1371/journal.pone.0062856

Chhabra, S., Brazil, D., Morrissey, J., Burke, J. I., O'Gara, F., and Dowling, N. D. (2013). Characterization of mineral phosphate solubilization traits from a barley rhizosphere soil functional metagenome. *Microbiologyopen* 2, 717–724. doi: 10.1002/mbo3.110

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., 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 Res.* 38:e200. doi: 10.1093/nar/gkq873

Costea, P. I., Lundeberg, J., and Akan, P. (2013). TagGD: fast and accurate software for DNA tag generation and demultiplexing. *PLoS ONE* 8:e57521. doi: 10.1371/journal.pone.0057521

Cretoiu, M. S., Kielak, A. M., Abu Al-Soud, W., Sørensen, S. J., and Van Elsas, J. D. (2012). Mining of unexplored habitats for novel chitinases - *chiA* as a helper gene proxy in metagenomics. *Appl. Microbiol. Biotechnol.* 94, 1347–1358. doi: 10.1007/s00253-012-4057-5

Dai, L., Gao, X., Guo, Y., Xiao, J. F., and Zhang, Z. (2012). Bioinformatics clouds for big data manipulation. *Biol. Direct* 7:43. doi: 10.1186/1745-6150-7-43

Das, S., and Vikalo, H. (2013). Base calling for high-throughput short-read sequencing: dynamic programming solutions. *BMC Bioinformatics* 14:129. doi: 10.1186/1471-2105-14-129

Davenport, C. F., and Tümmler, B. (2013). Advances in computational analysis of metagenome sequences. *Environ. Microbiol.* 15, 1–5. doi: 10.1111/j.1462-2920.2012.02843.x

De Filippo, C., Ramazzotti, M., Fontana, P., and Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief. Bioinform.* 13, 696–710. doi: 10.1093/bib/bbs070

Degnan, P. H., and Ochman, H. (2012). Illumina-based analysis of microbial community diversity. *ISME J.* 6, 183–194. doi: 10.1038/ismej.2011.74

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., et al. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16428–16433. doi: 10.1073/pnas.0905240106

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36:e105. doi: 10.1093/nar/gkn425

Dohrmann, A. B., Küting, M., Jünemann, S., Jaenicke, S., Schlüter, A., and Tebbe, C. C. (2013). Importance of rare taxa for bacterial diversity in the rhizosphere of Bt- and conventional maize varieties. *ISME J.* 7, 37–49. doi: 10.1038/ismej.2012.77

Duan, Y., Zhou, L., Hall, D. G., Li, W., Doddapaneni, H., Lin, H., et al. (2009). Complete genome sequence of citrus huanglongbing bacterium, *'Candidatus Liberibacter asiaticus*' obtained through metagenomics. *Mol. Plant Microbe Interact.* 22, 1011–1020. doi: 10.1094/MPMI-22-8-1011

Dumbrell, A. J., Ashton, P. D., Aziz, N., Feng, G., Nelson, M., Dytham, C., et al. (2011). Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytol.* 190, 794–804. doi: 10.1111/j.1469-8137.2010.03636.x

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J. X., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* 7:e47768. doi: 10.1371/journal.pone.0047768

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013). A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS ONE* 8:e66643. doi: 10.1371/journal.pone.0066643

Erlich, Y., Mitra, P. P., Delabastide, M., Mccombie, W. R., and Hannon, G. J. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods* 5, 679–682. doi: 10.1038/nmeth.1230

Faircloth, B. C., and Glenn, T. C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS ONE* 7:e42543. doi: 10.1371/journal.pone.0042543

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Fichot, E. B., and Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* 1:10. doi: 10.1186/2049-2618-1-10

Gaspar, J. M., and Thomas, W. K. (2013). Assessing the consequences of denoising marker-based metagenomic data. *PLoS ONE* 8:e60458. doi: 10.1371/journal.pone.0060458

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J. F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12:245. doi: 10.1186/1471-2164-12-245

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.* 2010:pdb prot5368. doi: 10.1101/pdb.prot5368

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., Macphee, R., et al. (2010). Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* 5:e15406. doi: 10.1371/journal.pone.0015406

Golan, D., and Medvedev, P. (2013). Using state machines to model the Ion Torrent sequencing process and to improve read error rates. *Bioinformatics* 29, 344–351. doi: 10.1093/bioinformatics/btt212

Gomez-Alvarez, V., Teal, T. K., and Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* 3, 1314–1317. doi: 10.1038/ismej.2009.72

Gori, F., Tringe, S. G., Folino, G., Hijum, S. A. F. T., Den Camp, H. J. M. O., Jetten, M. S. M., et al. (2013). Differences in sequencing technologies improve the retrieval of anammox bacterial genome from metagenomes. *BMC Genomics* 14:7. doi: 10.1186/1471-2164-14-7

Gottel, N. R., Castro, H. F., Kerley, M., Yang, Z. M., Pelletier, D. A., Podar, M., et al. (2011). Distinct microbial communities within the endosphere and rhizosphere of *Populus deltoides* roots across contrasting soil types. *Appl. Environ. Microbiol.* 77, 5934–5944. doi: 10.1128/AEM.05255-11

Guazzaroni, M. E., and Ferrer, M. (2011). "Metagenomic approaches in systems biology," in *Handbook of Molecular Microbial Ecology, Volume 1: Metagenomics and Complementary Approaches*, ed F. J. De Bruijn (Hoboken, NJ: Wiley-Blackwell), 475–489.

Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602–611. doi: 10.1016/j.tibtech.2008.07.003

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38:e131. doi: 10.1093/nar/gkq224

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10:R32. doi: 10.1186/gb-2009-10-3-r32

He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., et al. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* 7, 807–812. doi: 10.1038/nmeth.1507

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., et al. (2008). Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* 5, 183–188. doi: 10.1038/nmeth.1179

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., et al. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* 5:e1000502. doi: 10.1371/journal.pcbi.1000502

Hummelen, R., Fernandes, A. D., Macklaim, J. M., Dickson, R. J., Changalucha, J., Gloor, G. B., et al. (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE* 5:e12078. doi: 10.1371/journal.pone.0012078

Hunter, C. I., Mitchell, A., Jones, P., Mcanulla, C., Pesseat, S., Scheremetjew, M., et al. (2012). Metagenomic analysis: the challenge of the data bonanza. *Brief. Bioinform.* 13, 743–746. doi: 10.1093/bib/bbs020

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143. doi: 10.1186/gb-2007-8-7-r143

Huse, S. M., and Welch, D. M. (2011). "Accuracy and quality of massively parallel DNA pyrosequencing," in *Handbook of Molecular Microbial Ecology, Volume I: Metagenomics and Complementary Approaches,* ed F. J. De Bruijn (Hoboken, NJ: Wiley-Blackwell), 149–155.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Ilie, L., and Molnar, M. (2013). RACER: rapid and accurate correction of errors in reads. *Bioinformatics* 29, 2490–2493. doi: 10.1093/bioinformatics/btt407

Jiang, X. T., Peng, X., Deng, G. H., Sheng, H. F., Wang, Y., Zhou, H. W., et al. (2013). Illumina sequencing of 16S rRNA tag revealed spatial variations of bacterial communities in a mangrove wetland. *Microb. Ecol.* 66, 96–104. doi: 10.1007/s00248-013-0238-8

Jogler, M., Siemens, H., Chen, H., Bunk, B., Sikorski, J., and Overmann, J. (2011). Identification and targeted cultivation of abundant novel freshwater sphingomonads and analysis of their population substructure. *Appl. Environ. Microbiol.* 77, 7355–7364. doi: 10.1128/AEM.05832-11

Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., et al. (2013). Updating benchtop sequencing performance comparison. *Nat. Biotechnol.* 31, 294–296. doi: 10.1038/nbt.2522

Kavamura, V. N., Taketani, R. G., Lançoni, M. D., Andreote, F. D., Mendes, R., and De Melo, I. S. (2013). Water regime influences bulk soil and rhizosphere of *Cereus jamacaru* bacterial communities in the Brazilian Caatinga biome. *PLoS ONE* 8:e73606. doi: 10.1371/journal.pone.0073606

Kemler, M., Garnas, J., Wingfield, M. J., Gryzenhout, M., Pillay, K. A., and Slippers, B. (2013). Ion Torrent PGM as tool for fungal community analysis: a case study of endophytes in *Eucalyptus grandis* reveals high taxonomic diversity. *PLoS ONE* 8:e81718. doi: 10.1371/journal.pone.0081718

Kim, M., Lee, K. H., Yoon, S. W., Kim, B. S., Chun, J., and Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform.* 11, 102–113. doi: 10.5808/GI.2013.11.3.102

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* 10:R83. doi: 10.1186/gb-2009-10-8-r83

Knief, C., Delmotte, N., Chaffron, S., Stark, M., Innerebner, G., Wassmann, R., et al. (2012). Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J.* 6, 1378–1390. doi: 10.1038/ismej.2011.192

Knief, C., Delmotte, N., and Vorholt, J. A. (2011). Bacterial adaptation to life in association with plants - a proteomic perspective from culture to *in situ* conditions. *Proteomics* 11, 3086–3105. doi: 10.1002/pmic.201000818

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 692–700. doi: 10.1038/nbt.2280

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., et al. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* 472, 431–455. doi: 10.1016/S0076-6879(10)72001-2

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* 6, 291–295. doi: 10.1038/nmeth.1311

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/AEM.01043-13

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578. doi: 10.1128/MMBR.00009-08

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12, 118–123. doi: 10.1111/j.1462-2920.2009.02051.x

Kunin, V., and Hugenholtz, P. (2010). PyroTagger: a fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *Open J.* 1, 1–8.

Lahr, D. J., and Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47, 857–866. doi: 10.2144/000113219

Langevin, S. A., Bent, Z. W., Solberg, O. D., Curtis, D. J., Lane, P. D., Williams, K. P., et al. (2013). Peregrine: a rapid and unbiased method to produce strand-specific RNA-Seq libraries from small quantities of starting material. *RNA Biol.* 10, 502–515. doi: 10.4161/rna.24284

Ledergerber, C., and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* 12, 489–497. doi: 10.1093/bib/bbq077

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682–686. doi: 10.1126/science.1079700

Li, J., Jiang, H., and Wong, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* 11:R50. doi: 10.1186/gb-2010-11-5-r50

Lind, C., Ferriola, D., Mackiewicz, K., Heron, S., Rogers, M., Slavich, L., et al. (2010). Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Hum. Immunol.* 71, 1033–1042. doi: 10.1016/j.humimm.2010.06.016

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364. doi: 10.1155/2012/251364

Liu, Y. C., Schroder, J., and Schmidt, B. (2013). Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* 29, 308–315. doi: 10.1093/bioinformatics/bts690

Logares, R., Haverkamp, T. H. A., Kumar, S., Lanzen, A., Nederbragt, A. J., Quince, C., et al. (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J. Microbiol. Methods* 91, 106–113. doi: 10.1016/j.mimet.2012.07.017

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439. doi: 10.1038/nbt.2198

Lumini, E., Orgiazzi, A., Borriello, R., Bonfante, P., and Bianciotto, V. (2010). Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach. *Environ. Microbiol.* 12, 2165–2179. doi: 10.1111/j.1462-2920.2009.02099.x

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488, 86–90. doi: 10.1038/nature11237

Luo, C., Rodriguez, R. L., and Konstantinidis, K. T. (2013). A user's guide to quantitative and comparative analysis of metagenomic datasets. *Methods Enzymol.* 531, 525–547. doi: 10.1016/B978-0-12-407863-5.00023-X

Luo, C., Tsementzi, D., Kyrpides, N. C., and Konstantinidis, K. T. (2012a). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* 6, 898–901. doi: 10.1038/ismej.2011.147

Luo, C. W., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012b). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* 7:e30087. doi: 10.1371/journal.pone.0030087

Luo, H. W., and Moran, M. A. (2013). Assembly-free metagenomic analysis reveals new metabolic capabilities in surface ocean bacterioplankton. *Environ. Microbiol. Rep.* 5, 686–696. doi: 10.1111/1758-2229.12068

Mackelprang, R., Waldrop, M. P., Deangelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., et al. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480, 368–371. doi: 10.1038/nature10576

Maignien, L., Deforce, E. A., Chafee, M. E., Eren, A. M., and Simmons, S. L. (2014). Ecological succession and stochastic variation in the assembly of *Arabidopsis thaliana* phyllosphere communities. *MBio* 5:e00682–13. doi: 10.1128/mBio.00682-13

Maitra, R. D., Kim, J., and Dunbar, W. B. (2012). Recent advances in nanopore sequencing. *Electrophoresis* 33, 3418–3428. doi: 10.1002/elps.201200272

Mamanova, L., and Turner, D. J. (2011). Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). *Nat. Protoc.* 6, 1736–1747. doi: 10.1038/nprot.2011.399

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* 6, 287–303. doi: 10.1146/annurev-anchem-062012-092628

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi: 10.1038/nature03959

Mariette, J., Noirot, C., and Klopp, C. (2011). Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Res. Notes* 4:149. doi: 10.1186/1756-0500-4-149

Marshall, C. W., Ross, D. E., Fichot, E. B., Norman, R. S., and May, H. D. (2012). Electrosynthesis of commodity chemicals by an autotrophic microbial community. *Appl. Environ. Microbiol.* 78, 8412–8420. doi: 10.1128/AEM.02401-12

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. (2012). PANDAseq: PAired-eND Assembler for Illumina sequences. *BMC Bioinformatics* 13:31. doi: 10.1186/1471-2105-13-31

Mavromatis, K., Land, M. L., Brettin, T. S., Quest, D. J., Copeland, A., Clum, A., et al. (2012). The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLoS ONE* 7:e48837. doi: 10.1371/journal.pone.0048837

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12:451. doi: 10.1186/1471-2105-12-451

Meglecz, E., Pech, N., Gilles, A., Martin, J. F., and Gardner, M. G. (2012). A shot in the genome: how accurately do shotgun 454 sequences represent a genome? *BMC Res. Notes* 5:259. doi: 10.1186/1756-0500-5-259

Mendes, L. W., Kuramae, E. E., Navarrete, A. A., Van Veen, J. A., and Tsai, S. M. (2014). Taxonomical and functional microbial community selection in soybean rhizosphere. *ISME J.* doi: 10.1038/ismej.2014.17. [Epub ahead of print].

Merriman, B., Rothberg, J. M., and Ion Torrent R&D Team. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* 33, 3397–3417. doi: 10.1002/elps.201200424

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 12:R112. doi: 10.1186/gb-2011-12-11-r112

Morey, M., Fernandez-Marmiesse, A., Castineiras, D., Fraga, J. M., Couce, M. L., and Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* 110, 3–24. doi: 10.1016/j.ymgme.2013.04.024

Mosher, J. J., Bernberg, E. L., Shevchenko, O., Kan, J., and Kaplan, L. A. (2013). Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *J. Microbiol. Methods* 95, 175–181. doi: 10.1016/j.mimet.2013.08.009

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13:S8. doi: 10.1186/1471-2105-13-S14-S8

Nagasaki, H., Mochizuki, T., Kodama, Y., Saruhashi, S., Morizaki, S., Sugawara, H., et al. (2013). DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.* 20, 383–390. doi: 10.1093/dnares/dst017

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., et al. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39:e90. doi: 10.1093/nar/gkr344

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40:e155. doi: 10.1093/nar/gks678

Navarrete, A. A., Kuramae, E. E., De Hollander, M., Pijl, A. S., Van Veen, J. A., and Tsai, S. M. (2013). Acidobacterial community responses to agricultural management of soybean in Amazon forest soils. *FEMS Microbiol. Ecol.* 83, 607–621. doi: 10.1111/1574-6941.12018

Neiman, M., Sundling, S., Gronberg, H., Hall, P., Czene, K., Lindberg, J., et al. (2012). Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. *PLoS ONE* 7:e48616. doi: 10.1371/journal.pone.0048616

Newton, A. C., Fitt, B. D. L., Atkins, S. D., Walters, D. R., and Daniell, T. J. (2010). Pathogenesis, parasitism and mutualism in the trophic space of microbe-plant interactions. *Trends Microbiol.* 18, 365–373. doi: 10.1016/j.tim.2010.06.002

Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., and Geiger, T. L. (2011). Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 12:106. doi: 10.1186/1471-2164-12-106

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., and Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Anal. Chem.* 83, 4327–4341. doi: 10.1021/ac2010857

Niklas, N., Proll, J., Danzer, M., Stabentheiner, S., Hofer, K., and Gabriel, C. (2013). Routine performance and errors of 454 HLA exon sequencing in diagnostics. *BMC Bioinformatics* 14:176. doi: 10.1186/1471-2105-14-176

Niu, B. F., Fu, L. M., Sun, S. L., and Li, W. Z. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11:187. doi: 10.1186/1471-2105-11-187

Ottesen, A. R., Gonzalez, A., Bell, R., Arce, C., Rideout, S., Allard, M., et al. (2013a). Co-enriching microflora associated with culture based methods to detect Salmonella from tomato phyllosphere. *PLoS ONE* 8:e73079. doi: 10.1371/journal.pone.0073079

Ottesen, A. R., Peña, A. G., White, J. R., Pettengill, J. B., Li, C., Allard, S., et al. (2013b). Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum lycopersicum* (tomato). *BMC Microbiol.* 13:114. doi: 10.1186/1471-2180-13-114

Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., et al. (2012). Optimizing Illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* 13:1. doi: 10.1186/1471-2164-13-1

Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J. Appl. Genet.* 52, 413–435. doi: 10.1007/s13353-011-0057-x

Parkinson, N. J., Maslau, S., Ferneyhough, B., Zhang, G., Gregory, L., Buck, D., et al. (2012). Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res.* 22, 125–133. doi: 10.1101/gr.124016.111

Peiffer, J. A., Spor, A., Koren, O., Jin, Z., Tringe, S. G., Dangl, J. L., et al. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6548–6553. doi: 10.1073/pnas.1302837110

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2011). Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* 27, I94–I101. doi: 10.1093/bioinformatics/btr216

Perazzolli, M., Antonielli, L., Storari, M., Puopolo, G., Pancher, M., Giovannini, O., et al. (2014). Resilience of the natural phyllosphere microbiota of the grapevine to chemical and biological pesticides. *Appl. Environ. Microbiol.* doi: 10.1128/AEM.00415-00411. [Epub ahead of print].

Perkins, T. T., Tay, C. Y., Thirriot, F., and Marshall, B. (2013). Choosing a benchtop sequencing machine to characterise *Helicobacter pylori* genomes. *PLoS ONE* 8:e67539. doi: 10.1371/journal.pone.0067539

Pop, M., and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142–149. doi: 10.1016/j.tig.2007.12.006

Prabakaran, P., Streaker, E., Chen, W., and Dimitrov, D. S. (2011). 454 antibody sequencing - error characterization and correction. *BMC Res. Notes* 4:404. doi: 10.1186/1756-0500-4-404

Preheim, S. P., Perrotta, A. R., Friedman, J., Smilie, C., Brito, I., Smith, M. B., et al. (2013). Computational methods for high-throughput comparative analyses of natural microbial communities. *Methods Enzymol.* 531, 353–370. doi: 10.1016/B978-0-12-407863-5.00018-6

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., et al. (2008). A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5, 1005–1010. doi: 10.1038/nmeth.1270

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341. doi: 10.1186/1471-2164-13-341

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., et al. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641. doi: 10.1038/nmeth.1361

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12:38. doi: 10.1186/1471-2105-12-38

Quinlan, A. R., Stewart, D. A., Stromberg, M. P., and Marth, G. T. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* 5, 179–181. doi: 10.1038/nmeth.1172

Rastogi, G., Sbodio, A., Tech, J. J., Suslow, T. V., Coaker, G. L., and Leveau, J. H. J. (2012). Leaf microbiota in an agroecosystem: spatiotemporal variation in bacterial community composition on field-grown lettuce. *ISME J.* 6, 1812–1822. doi: 10.1038/ismej.2012.32

Ratan, A., Miller, W., Guillory, J., Stinson, J., Seshagiri, S., and Schuster, S. C. (2013). Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS ONE* 8:e55089. doi: 10.1371/journal.pone.0055089

Redford, A. J., Bowers, R. M., Knight, R., Linhart, Y., and Fierer, N. (2010). The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environ. Microbiol.* 12, 2885–2893. doi: 10.1111/j.1462-2920.2010.02258.x

Reeder, J., and Knight, R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods* 7, 668–669. doi: 10.1038/nmeth0910-668b

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., and Dangl, J. L. (2009). *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae. Genome Res.* 19, 294–305. doi: 10.1101/gr.083311.108

Reisberg, E. E., Hildebrandt, U., Riederer, M., and Hentschel, U. (2013). Distinct phyllosphere bacterial communities on *Arabidopsis* wax mutant leaves. *PLoS ONE* 8:e78613. doi: 10.1371/journal.pone.0078613

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., et al. (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS ONE* 8:e66621. doi: 10.1371/journal.pone.0066621

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Rodrigue, S., Materna, A. C., Timberlake, S. C., Blackburn, M. C., Malmstrom, R. R., Alm, E. J., et al. (2010). Unlocking short read sequencing for metagenomics. *PLoS ONE* 5:e11840. doi: 10.1371/journal.pone.0011840

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946. doi: 10.1101/gr.128124.111

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science* 281, 363, 365.

Rosenzweig, N., Tiedje, J. M., Quensen, J. F., Meng, Q. X., and Hao, J. J. J. (2012). Microbial communities associated with potato common scab-suppressive soil determined by pyrosequencing analyses. *Plant Dis.* 96, 718–725. doi: 10.1094/PDIS-07-11-0571

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51. doi: 10.1186/gb-2013-14-5-r51

Rozera, G., Abbate, I., Bruselles, A., Vlassi, C., D'Offizi, G., Narciso, P., et al. (2009). Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* 6:15. doi: 10.1186/1742-4690-6-15

Ruan, J., Jiang, L., Chong, Z. C., Gong, Q., Li, H., Li, C. Y., et al. (2013). Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. *BMC Genomics* 14:711. doi: 10.1186/1471-2164-14-711

Salmela, L. (2010). Correction of sequencing errors in a mixed set of reads. *Bioinformatics* 26, 1284–1290. doi: 10.1093/bioinformatics/btq151

Salmela, L., and Schröder, J. (2011). Correcting errors in short reads by multiple alignments. *Bioinformatics* 27, 1455–1461. doi: 10.1093/bioinformatics/btr170

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240. doi: 10.1093/hmg/ddq416

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Res.* 20, 1165–1173. doi: 10.1101/gr.101360.109

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310. doi: 10.1371/journal.pone.0027310

Scholz, M. B., Lo, C. C., and Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15. doi: 10.1016/j.copbio.2011.11.013

Schröder, J., Bailey, J., Conway, T., and Zobel, J. (2010). Reference-free validation of short read data. *PLoS ONE* 5:e12681. doi: 10.1371/journal.pone.0012681

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* 9:666. doi: 10.1038/msb.2013.22

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., et al. (2012). Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Mol. Plant Microbe Interact.* 25, 28–36. doi: 10.1094/MPMI-08-11-0204

Shade, A., Mcmanus, P. S., and Handelsman, J. (2013). Unexpected diversity during community succession in the apple flower microbiome. *MBio* 4:e00602-12. doi: 10.1128/mBio.00602-12

Shao, W., Boltz, V. F., Spindler, J. E., Kearney, M. F., Maldarelli, F., Mellors, J. W., et al. (2013). Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology* 10:18. doi: 10.1186/1742-4690-10-18

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi: 10.1038/nbt1486

Shin, S. C., Ahn Do, H., Kim, S. J., Lee, H., Oh, T. J., Lee, J. E., et al. (2013). Advantages of single-molecule real-time sequencing in high-GC content genomes. *PLoS ONE* 8:e68824. doi: 10.1371/journal.pone.0068824

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* 21, 1794–1805. doi: 10.1111/j.1365-294X.2012.05538.x

Skums, P., Dimitrova, Z., Campo, D. S., Vaughan, G., Rossi, L., Forbi, J. C., et al. (2012). Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13(Suppl. 10):S6. doi: 10.1186/1471-2105-13-S10-S6

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Stark, M., Berger, S. A., Stamatakis, A., and Von Mering, C. (2010). MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 11:461. doi: 10.1186/1471-2164-11-461

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., et al. (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.* 39, D546–D551. doi: 10.1093/nar/gkq1102

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Suzuki, S., Ono, N., Furusawa, C., Ying, B. W., and Yomo, T. (2011). Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE* 6:e19534. doi: 10.1371/journal.pone.0019534

Tariq, M. A., Kim, H. J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res.* 39:e120. doi: 10.1093/nar/gkr547

Tautz, D., Ellegren, H., and Weigel, D. (2010). Next generation molecular ecology. *Mol. Ecol.* 19(Suppl. 1), 1–3. doi: 10.1111/j.1365-294X.2009.04489.x

Teeling, H., and Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis-a bioinformatic perspective. *Brief. Bioinform.* 13, 728–742. doi: 10.1093/bib/bbs039

Thakur, K., Chawla, V., Bhatti, S., Swarnkar, M. K., Kaur, J., Shankar, R., et al. (2013). *De novo* transcriptome sequencing and analysis for *Venturia inaequalis*, the devastating apple scab pathogen. *PLoS ONE* 8:e53937. doi: 10.1371/journal.pone.0053937

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* 2:3. doi: 10.1186/2042-5783-2-3

Thompson, J. F., and Milos, P. M. (2011). The properties and applications of single-molecule DNA sequencing. *Genome Biol.* 12:217. doi: 10.1186/gb-2011-12-2-217

Timp, W., Mirsaidov, U. M., Wang, D., Comer, J., Aksimentiev, A., and Timp, G. (2010). Nanopore sequencing: electrical measurements of the code of life. *IEEE Trans. Nanotechnol.* 9, 281–294. doi: 10.1109/TNANO.2010.2044418

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38:e159. doi: 10.1093/nar/gkq543

Treffer, R., and Deckert, V. (2010). Recent advances in single-molecule sequencing. *Curr. Opin. Biotechnol.* 21, 4–11. doi: 10.1016/j.copbio.2010.02.009

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2012). Identification of genes expressed by *Phakopsora pachyrhizi*, the pathogen causing soybean rust, at a late stage of infection of susceptible soybean leaves. *Plant Pathol.* 61, 773–786. doi: 10.1111/j.1365-3059.2011.02550.x

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., et al. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME J.* 7, 2248–2258. doi: 10.1038/ismej.2013.119

Unno, Y., and Shinano, T. (2013). Metagenomic analysis of the rhizosphere soil microbiome with respect to phytic acid utilization. *Microbes Environ.* 28, 120–127. doi: 10.1264/jsme2.ME12181

Vandenbroucke, I., Van Marck, H., Verhasselt, P., Thys, K., Mostmans, W., Dumont, S., et al. (2011). Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. *Biotechniques* 51, 167–177. doi: 10.2144/000113733

Van Dijk, E. L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322, 12–20. doi: 10.1016/j.yexcr.2014.01.008

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55, 641–658. doi: 10.1373/clinchem.2008.112789

Vorholt, J. A. (2012). Microbial life in the phyllosphere. *Nat. Rev. Microbiol.* 10, 828–840. doi: 10.1038/nrmicro2910

Weßling, R., Schmidt, S. M., Micali, C. O., Knaust, F., Reinhardt, R., Neumann, U., et al. (2012). Transcriptome analysis of enriched *Golovinomyces orontii* haustoria by deep 454 pyrosequencing. *Fungal Genet. Biol.* 49, 470–482. doi: 10.1016/j.fgb.2012.04.001

Weinstock, G. M. (2011). "The impact of next-generation sequencing technologies on metagenomics," in *Handbook of Molecular Microbial Ecology, Volume 1: Metagenomics and Complementary Approaches*, ed F. J. De Bruijn (Hoboken, NJ: Wiley-Blackwell), 143–147.

Whiteley, A. S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., et al. (2012). Microbial 16S rRNA ion tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *J. Microbiol. Methods* 91, 80–88. doi: 10.1016/j.mimet.2012.07.008

Wilke, A., Wilkening, J., Glass, E. M., Desai, N. L., and Meyer, F. (2011). An experience report: porting the MG-RAST rapid metagenomics analysis pipeline to the cloud. *Concurr. Comp. Pract. Exp.* 23, 2250–2257. doi: 10.1002/cpe.1799

Williams, T. R., Moyne, A. L., Harris, L. J., and Marco, M. L. (2013). Season, irrigation, leaf age, and *Escherichia coli* inoculation influence the bacterial diversity in the lettuce phyllosphere. *PLoS ONE* 8: e68642. doi: 10.1371/journal.pone.0068642

Wommack, K. E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74, 1453–1463. doi: 10.1128/AEM.02181-07

Wu, D. Y., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 462, 1056–1060. doi: 10.1038/nature08656

Xu, M., Fujita, D., and Hanagata, N. (2009). Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small* 5, 2638–2649. doi: 10.1002/smll.200900976

Yang, X., Chockalingam, S. P., and Aluru, S. (2013). A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform.* 14, 56–66. doi: 10.1093/bib/bbs015

Yergeau, E., Sanschagrin, S., Maynard, C., St-Arnaud, M., and Greer, C. W. (2014). Microbial expression profiles in the rhizosphere of willows depend on soil contamination. *ISME J.* 8, 344–358. doi: 10.1038/ismej.2013.163

Yu, L., Nicolaisen, M., Larsen, J., and Ravnskov, S. (2012). Succession of root-associated fungi in *Pisum sativum* during a plant growth cycle as examined by 454 pyrosequencing. *Plant Soil* 358, 216–224. doi: 10.1007/s11104-012-1188-5

Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *J. Genet. Genomics* 38, 95–109. doi: 10.1016/j.jgg.2011.02.003

Zhang, W., Wu, X. K., Liu, G. X., Chen, T., Zhang, G. S., Dong, Z. B., et al. (2013). Pyrosequencing reveals bacterial diversity in the rhizosphere of three *Phragmites australis* ecotypes. *Geomicrobiol. J.* 30, 593–599. doi: 10.1080/01490451.2012.740145

Zhang, W. W., Li, F., and Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology* 156, 287–301. doi: 10.1099/mic.0.034793-0

Zhou, H. W., Li, D. F., Tam, N. F. Y., Jiang, X. T., Zhang, H., Sheng, H. F., et al. (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* 5, 741–749. doi: 10.1038/ismej.2010.160

Zhuang, X., McPhee, K. E., Coram, T. E., Peever, T. L., and Chilvers, M. I. (2012). Rapid transcriptome characterization and parsing of sequences in a non-model host-pathogen interaction: pea-*Sclerotinia sclerotiorum*. *BMC Genomics* 13:668. doi: 10.1186/1471-2164-13-668

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 February 2014; accepted: 30 April 2014; published online: 21 May 2014.
*Citation: Knief C (2014) Analysis of plant microbe interactions in the era of next generation sequencing technologies. Front. Plant Sci. **5**:216. doi: 10.3389/fpls.2014.00216*
This article was submitted to Plant Genetics and Genomics, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Knief. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.