Genomics Data 6 (2015) 159–163

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Metagenomics study of endophytic bacteria in *Aloe vera* using next-generation technology

Mushafau Adewale Akinsanya [a,b], Joo Kheng Goh [a], Siew Ping Lim [a], Adeline Su Yien Ting [a,*]

[a] School of Science, Monash University Malaysia, 46150 Bandar Sunway, Selangor, Malaysia
[b] Department of Medical Biochemistry, Faculty of Basic Medical Sciences, College of Medicine, Lagos State University, P.M.B 21266 Ikeja, Lagos, Nigeria

ARTICLE INFO

Article history:
Received 18 July 2015
Received in revised form 19 August 2015
Accepted 4 September 2015
Available online 10 September 2015

Keywords:
Aloe vera
α-Diversity
Bacterial endophytes
Illumina
Metagenomics
NGS

ABSTRACT

Next generation sequencing (NGS) enables rapid analysis of the composition and diversity of microbial communities in several habitats. We applied the high throughput techniques of NGS to the metagenomics study of endophytic bacteria in *Aloe vera* plant, by assessing its PCR amplicon of 16S rDNA sequences (V3–V4 regions) with the Illumina metagenomics technique used to generate a total of 5,199,102 reads from the samples. The analyses revealed *Proteobacteria*, *Firmicutes*, *Actinobacteria* and *Bacteriodetes* as the predominant genera. The roots have the largest composition with 23% not present in other tissues. The stems have more of the genus—*Pseudomonas* and the unclassified *Pseudomonadaceae*. The α-diversity analysis indicated the richness and inverse Simpson diversity index of the bacterial endophyte communities for the leaf, root and stem tissues to be 2.221, 6.603 and 1.491 respectively. In a similar study on culturable endophytic bacteria in the same *A. vera* plants (unpublished work), the dominance of *Pseudomonas* and *Bacillus* genera was similar, with equal proportion of four species each in root, stem and leaf tissues. It is evident that NGS technology captured effectively the metagenomics of microbiota in plant tissues and this can improve our understanding of the microbial–plant host interactions.

1. Introduction

The diversity of microorganisms on earth remains poorly understood although an estimated 1.5 million species of bacteria and fungi have vital functions as decomposers, symbionts, and pathogens in ecosystems [1]. To date, only 5% of the estimated number of bacterial species has been documented. In recent years, metagenomics studies have improved our understanding of the diversity of microbes in various habitats. This includes microbes associated with plants, which thrive below ground in the rhizosphere, above in the phyllosphere [2] and within the plant tissues as endophytes [3,4]. These microbes can have beneficial, neutral, or detrimental effects on the plant. The association between microbiota and plants is important, as it leads to understanding microbes and "What are they doing and how do they respond to environmental changes and interact with each other?" These could further contribute to understanding the significant roles of plant microbiota in supporting plant growth and improved crop yield.

Genomic analyses of individual strains or metagenomics studies of whole microbial communities may provide insight into the composition or diversity and physiological potential of endophytes associated with plants. For example, the study of microbial (endophyte) diversity in plants tissues, reveals both culturable and unculturable endophytes that may be beneficial microbes, subsequently gearing towards their isolation and characterisation. It is also possible to further evaluate evolutionary trend of the associated microbes and how they are related with one another. We may also be able to evaluate the statement of their close endophyte–host association and co-evolution in relation to their ability to produce similar compounds to that of their host [3].

In recent studies, endophytes have been shown to have an important role in promoting plant growth and yield, suppress pathogens, aid in removing contaminants, solubilize phosphate or contribute to nitrogen assimilation for plants [5,6]. Over the past decade, our understanding of microbial diversity and function in complex environments has increased significantly, primarily as a result of the introduction of next generation sequencing (NGS) [7]. Both PCR based analysis of 16S rRNA gene and shotgun metagenomics studies have been used recently to characterise soils [8], oceans [9], the atmosphere, as well as the human microbiome [10]. Prior to the introduction of NGS, these characterisations were done at extremely high cost [11]. The use of the Illumina platform to generate data sets of unprecedented size [12,13] further revealed more information of various microbiomes but at much lower cost.

Next generation sequencing of hypervariable regions from small-subunit ribosomal RNA genes (16S rRNA) is useful for analyses of microbial communities in several habitats [14]. The use of high-throughput short-read sequencing of the 16S rRNA amplicon for the profiling of

* Corresponding author.
E-mail address: adeline.ting@monash.edu (A.S.Y. Ting).

http://dx.doi.org/10.1016/j.gdata.2015.09.004

microbial communities has become an increasingly attractive option by researchers as the amplicon consists of the conserved region interspersed by variable regions that facilitate sequencing and phylogenetic classification. The 454 GS-20 pyrosequencing by Roche in 2006 was the first high-throughput sequencing technology successfully applied for biodiversity analysis, and further improvement of the technology offers read lengths of up to 1000 bp [15]. The Illumina technology is highly effective in performing comparatively high sequencing depth despite having short read lengths and reduced per base costs [9,13,16]. Therefore, this technology has been used for amplicon sequencing of bacterial and fungal marker genes to characterise microbial communities in the phyllosphere and rhizosphere [17]. In this study, Illumina technology was applied for the 16S rRNA sequencing targeting the V3–V4 regions (amplicon of 150–400 bp) using primers designed against the surrounding conserved regions [18]. The bioinformatics tools provided by mothur pipeline were explored to process raw data reads and to analyse the microbiota communities. Previous work done using Illumina platform has suggested the effectiveness of this fragment size to be sufficient for resolving microbial community differences [19].

This study focussed on endophyte communities from the *Aloe vera* plant. In our separate study (unpublished work), we have characterised some of the culturable endophyte isolates from *A. vera* with beneficial bioactive compounds. *Aloe* plants are known for its nutritional and therapeutic values. The leaf exudates are used to a great extent in traditional medicine [20]. Other uses include treating wounds and burns, also diabetes and elevated blood lipids in humans. These effects are believed to be attributed to compounds such as polysaccharides, mannans, flavonoids, anthraquinones, lectins and other phytochemical compounds that are isolated from the plant. Unmasking the overall endophytic bacteria communities may help in identifying and describing the microbial plant colonisation by both the culturable and unculturable species and their link to the bioactive compounds produced. Hence, we employed the NGS technology to unveil the culturable and unculturable endophytic bacteria in *A. vera*, and to elucidate the microbial plant colonisation pattern and evaluate its microbial diversity.

2. Results

The primary analysis of the reads through base calling directly on the MiSeq sequencing reporter (MSR), revealed raw reads statistics and sequence quality assessment as in Table 1. The fastq reads obtained per sample were in paired-end reads labelled as (L001_R1_001.fastq and L001_R2_001.fastq). The project (PRJNA288893) was registered with the GenBank, with BioSample accession numbers SAMN03839381 (root), SAMN03975610 (stem), and SAMN03975611 (leaf). The highest reads were obtained from the root tissues (2,528,030 reads) followed by the leaf (1,372,180 reads) and stem tissues (1,298,892 reads). The GC content followed the same order (52.01, 50.86 and 50.01), respectively. This trend may not be unusual since the root is closer to the soil microbial communities than other tissues. The higher reads in leaf tissues compared to the stem, may probably be the result of the relatively larger size of leaf tissues than the stem tissues, which might harbour more microbial communities.

2.1. Sequence processing

In this study, the sequences were processed using mothur, a software package with less computational demands [21]. Analysis of the raw data indicated that the reads covered V3 region successfully (size ranged ~ 200 bp). Forward and reverse reads were merged and >99% were overlapped at V3 region using the mothur pipeline (Refer supplementary Figs. S1, S2 and Table 2). The merged sequences were further processed. According to Huse et al. [22], accumulation of errors within a rather small subset of 454 reads may occur hence it was necessary to remove reads with ambiguous base calls (Ns), unusual or unexpected length, low quality scores or those that cannot be aligned to the gene of interest (assumed to be unspecific PCR products) [22,23]. Reads were trimmed based on quality scores, singletons (sequence reads that occur only once) are removed from the datasets to further reduce the error rate [9].

The mothur "seqNoise algorithm" incorporated with UCHIME further removed chimeric sequences originated during PCR (5–45% of PCR product) [24,25]. UCHIME was reported to perform best in a comparative study where a reference database was used [26]. Critical analyses of different denoising tools demonstrated that parameters have to be chosen very carefully so as not to introduce bias by read modification during the generation of representative consensus reads. Hence, mothur which combined the above analyses such as OTU clustering, taxonomy assignment and multiple sample comparison, has been considered to be more appropriate or the UPARSE pipeline [26,13,27] for OTU estimation. The resulting merged sequences and processing details are shown in Tables 2 and 3 and supplementary Table S1.

Table 2
Sequence processed details: merged sequence.

Sample reference	Before merge process Number of sequence (total sequence length in bp)	After merge process Number of sequence (total sequence length in bp)
Root	2,528,030 (361,652,861)	1,264,015 (220,836,340)
Stem	1,298,892 (191,468,046)	649,446 (124,733,765)
Leaf	1,372,180 (200,256,446)	686,090 (127,684,140)
Total		2,599,551 (473,254,245)

Sequence input (forward and reverse sequences), quality encoding (Illumine 1.8+) and Alignment method (needleman).

2.2. Characterisation of community composition

The relative abundance of bacterial communities as obtained in the three tissues evaluated is shown in Fig. 1. Of the three tissues analysed, *Proteobacteria* sub-phylum is predominant followed by *Firmicutes*, *Actinobacteria* and *Bacteroidetes*. It was noted that the stem tissue has more of the genus—*Pseudomonas* and unclassified *Pseudomonadaceae* than the root and leaf tissues. On the contrary, leaf tissues have more of genus—*Propionibacterium*, *Serratia* and *Brevibacterium* than the root and stem tissues (refer supplementary Tables S2 and S3). In all, the root tissues have the highest richness of the four bacteria groupings.

Computational analyses of the α-diversity estimated the richness and diversity of the three samples at OTU cutoffs of 1 distance units (by using the number of observed OTUs). Chao1 estimated minimum number of OTUs, and inverse Simpson diversity index indicates the richness of the communities (refer supplementary Fig. S4). Chao1 curves continue to climb with sampling; however, the inverse Simpson diversity indices are relatively stable. The summary table of the diversity gave us the insight to the sampling coverage of the communities which is well above 99%. Also, there are significant differences of the

Table 1
Raw reads statistics and sequence quality assessment of 16S rRNA sequence from *A. vera* tissues.

Sample reference	Sample label	Sequence type	Sequence format	Read type	Read size (bp)	Total number of reads	Total sequence length (nt)	GC%
Root	22	Illumina MiSeq	Fastq	Paired-end	35–151	2,528,030	361,652,861	52.01
Stem	23					1,298,892	191,468,046	50.01
Leaf	24					1,372,180	200,256,446	50.86

Raw data from MiSeq sequencing reporter (MSR).

Table 3
Sequence processed details.

Sequence details	Number of sequence	Percentage
Merge sequence	2,599,551	
Removed redundancy sequence	152,919	100
Contaminant removal		
Chimeric	1779	1.16
Chloroplast	34,816	22.77
Mitochondria	487	0.32
Eukaryote	42	0.03
Unknown	3	0
Cleaned sequence	115,792	75.72

Table 4
Diversity and richness of the communities in plant tissue samples.

Group	Method	Number of sequences	Coverage %	Observed OTUs	Inverse Simpson diversity index
Leaf	Average	430,023	99.992	175.0	2.221
Root	Average	430,023	99.994	211.0	6.603
Stem	Average	430,023	99.991	147.9	1.491
Leaf	STDEV	0	0	0	0.000071
Root	STDEV	0	0	0	0
Stem	STDEV	0	0	0.063119	0.000024

ANOVA statistical analysis showed that there are significant differences of the observed OTUs and inverse Simpson diversity index between the tissues. Coverage also reflected over 99% sampling of the communities in the tissues.

observed OTUs and diversity or richness of the communities (Table 4) in the tissues as computed by mothur pipeline.

3. Discussion

Our study revealed for the first time the possibility of Illumina sequencing protocol to evaluate microbiota present in plant tissues–bacterial endophytes. The sequencing can be improved with good choice of primer pair to amplify a longer stretch of the 16S rRNA gene. Our empirical results highlight the utility of this platform for precise and high resolution microbiota profiling (>90% at species level) of endophytic communities, or perhaps extended to other resources/samples. The improvement to the various analyses tools was equally important to minimise the biasness introduced by the host DNA (chloroplast) and chimaera which was removed without affecting the overall quality of the reads. Mothur pipeline relatively provides us a good opportunity to effectively process the read sequence on a single platform. This minimised the probable loss of quality of reads. The use of novel shotgun 16S rRNA gene by NGS has also revealed the overall richness and diversity of microbiota communities in plant tissues to encompass both the culturable and unculturable endophytic bacteria. The α-diversity analysis indicated the richness and inverse Simpson diversity index of the bacterial endophyte communities for the leaf, root and stem tissues to be 2.221, 6.603 and 1.491 respectively. It further elucidates the microbial colonisation of plant tissues as revealed by the Venn diagram which illustrates the distribution of the bacterial communities across the tissues and the total shared richness (Fig. 2).

The colonisation pattern as illustrated by the Venn diagram of the OTU distribution indicated that 41% of microbes found in the root tissues were also present in all the three tissues, such as *Pseudomonas*, *Bacilli*, *Klebsiella* and unclassified families of *Pseudomonadaceae*, *Enterobacteriaceae* and *Bacillaceae* (refer supplementary Table S2). It was interesting to note that the *Klebsiella* genus was not captured in the stem and leaf tissues from culturable isolation method (unpublished), strengthening the use of NGS in this study. It was also noted that 9% of the microbes were present only in both root and leaf tissues, and 5% of microbes identified were present only in both root and stem tissues. Nevertheless, the stem tissues accommodated some genera (6%) such as *Gluconacetobacter* and *Anoxybacillus* which were not detected in both the root and leaf tissues. This further collaborated the facts that there are many routes by which these microbes enter into the plant tissues [28,29]. Conclusively, four prominent phyla were identified to colonise *A. vera* plant tissues; *Proteobacteria*, *Firmicutes*, *Actinobacteria* and *Bacteriodetes*, which have been shown to produce beneficial bioactive compounds (unpublished).

4. Materials and methods

4.1. Sample collection and DNA extraction

The *A. vera* plants were collected from Sungai Buloh horticulture nursery area (3.235283 N, 101.568342 E), Selangor. For a relatively wide coverage, five different plants were dug at different points in the same location, neatly transferred into sterile biosafety bag and brought to Laboratory for immediate analyses. Each plant was washed under running tap water to remove soil particles and allowed to drain. The leaves, stems and roots were detached with sterile knife and washed with sterile distilled water plus a few drops of Tween-20 and left for 10–15 min to drain. These were then cut into 4–5 pieces (2–3 cm in size). Surface sterilization was performed according to the methods described by Azevedo et al. [30] with modifications to the duration for sterilization and ethanol concentrations. Briefly, tissues were immersed separately in 90% ethanol (5 min), followed by sodium hypochlorite (3%) solution (2 min), and into 75% ethanol (3 min). The disinfected leaves, stems and roots were rinsed three times in sterile distilled

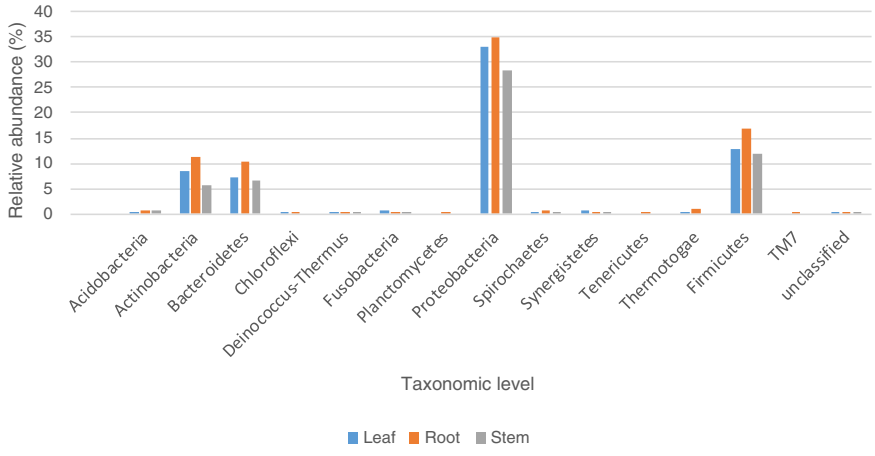

Fig. 1. Bacterial taxonomic composition histogram. The average composition of bacteria communities obtained from surface sterilized tissues of *A. vera* using culture-independent method (MiSeq Illumina platform) was analysed and compared. The nomenclatures of the phylotypes are based on the SILVA rRNA database (http://www.mothur.org/wiki/Silva_reference_files).

Fig. 2. Venn diagram describing the OTU distribution across tissue samples.

water and drained in laminar flow hood. To validate the effectiveness of the surface sterilization procedure, the surface-sterilized tissues (control) and the last rinsing water were inoculated onto nutrient agar plates and any bacteria growth in the control agar plates within 24 h of incubation (30 °C ± 2 °C) indicates ineffective surface-sterilization and samples discarded. After surface sterilization, the tissues were gently homogenized separately with sterile 12.5 mM potassium phosphate buffer (pH 7.1) using sterile test-tube and glass rod to release the microbes in the tissues. The resulting homogenates were centrifuge at 8000 ×*g* for 3 min and the supernatants were collected separately in triplicate per tissue. Bacteria genomic DNA was extracted from the supernatants using GF-1 bacterial DNA extraction kit by Vivantis. The extracted genomic DNA was quantified and checked for purity at A260/280 nm (1.9–2.0) (Nanodrop, Thermo Fisher Scientific, U.S.A.) and stored at −20 °C.

4.2. Illumina Library preparation

The microbial genomic DNA from root, stem and leaf tissue samples was normalized to concentration ≤ 10 ng/μL. PCR amplification was carried out to amplify V3–V4 conserved regions of 16S rRNA gene sequences in triplicate [31,22] using the 16S rRNA gene primers (forward primer 5′-CCTACGGGNGGCWGCAG-3′ and reverse 5′-GACT ACHVGGGTATCTA-3′) [32]. The PCR library preparation was carried out using KAPA HiFi HotStart Ready-mix PCR Kit (KAPA BIOSYSTEMS® U.S.A.) and Nextera® XT index kit to add multiplexing indices (dual-index barcodes). Briefly, each 25 μL of PCR reaction contains 10 ng/μL (6 μL) of genomic DNA template, 12.5 μL 2× Mastermix KAPA HiFidelity DNA polymerase (1 U), 1.5 μL each primer (10 μM) and nuclease free water. PCR reactions were carried out with initial denaturation step at 95 °C for 3 min followed by 24 cycles of 98 °C for 20 s, 55 °C for 15 s, and 72 °C for 10 s and ended with an extension step at 72 °C for 1 min. The PCR products were confirmed by 2% agarose gel electrophoresis and recovered using QIAquick gel extraction kit (Qiagen, Mississauga, Ontario, Canada). The quality and quantity of the PCR amplicon were analysed by TECAN infinite M200 Multi-Detection Microplate Reader, Chemopharm. The PCR amplicons were then tagged with sequencing adapters using Nextera® XT index kit to add multiplexing indices (dual-index barcodes). The libraries (3 samples per tissue) were normalized and pooled prior to sequencing. These samples were then loaded onto MiSeq reagent cartridge (MiSeq Kit V_2 300 cycles) for sequencing on the MiSeq system where automated cluster generation and paired-end sequencing with dual index reads were performed [33].

4.3. Initial processing of sequencing datasets and sequence quality assessment

Preliminary analysis of the image and base calling were done on the MiSeq instrument. MiSeq Sequencing Reporter (MSR) was used for de-multiplexing of data and removal of reads that failed Illumina's purity/chastity filter (PF = 0), and reads obtained in FASTQ format [34,35].

4.4. Sequence processing

The raw data forward and reverse reads were merged using mothur pipeline alignment method. These were then filtered and trimmed by removing trailing bases with quality scores lower or equal to 2, maximum number of N allowed = 4, maximum number of homopolymer allowed = 8 and contaminant removed. All processing were done using mothur pipeline software (http://www.mothur.org/wiki/Download_mothur).

4.5. Characterisation of community composition

Operational taxonomic units (OTUs) were assigned to the reconstructed read sequences obtained from the root, stem and leaf samples using SILVA rRNA database (http://www.mothur.org/wiki/Silva_reference_files), for the SILVA database we used Silva bacterial reference release 102 (http://www.mothur.org/w/images/9/98/Silva.bacteria.zip). Hence, for the assignment of the OTU we used "splitting by classification" method of the mothur pipeline (http://www.mothur.org/wiki/Cluster.split#method).

5. Statistical analysis

One-way ANOVA was used to analyse all data obtained. The analysis was carried out using the Statistical Package for Social Science (SPSS) version 16.0 and means are compared using Tukey's Studentized Range Test ($HSD_{(0.05)}$) and p values < 0.05 are consider statistically different.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.09.004.

Funding

This work was supported by School of Science, MONASH University Malaysia, Higher Degree Research Scholarship (HDR).

Conflict of interest

The authors report no conflict of interest and are responsible for the content and writing of the manuscript.

Acknowledgements

The authors acknowledged the Tropical Medicine and Biology Multidisciplinary Platform of School of Science, MONASH University Malaysia for the use of MiSeq and Codon Genomics Malaysia for the assistance on bioinformatics.

References

1. D.L. Hawksworth, Fungal diversity and its implications for genetic resource collections. Stud. Mycol. 50 (2004) 9–18.
2. D. Bulgarelli, K. Schlaeppi, S. Spaepen, E.V.L. van Themaat, P. Schulze-Lefert, Structure and functions of the bacterial microbiota of plants. Annu. Rev. Plant Biol. 64 (2013) 807–838.
3. G. Strobel, B. Daisy, Bioprospecting for microbial endophytes and their natural products. Microbiol. Mol. Biol. Rev. 67 (2003) 491–502.
4. G.A. Strobel, Endophytes as sources of bioactive products. Microbes Infect. 5 (2003) 535–544.
5. C. Chen, E. Bauske, G. Musson, R. Rodriguezkabana, J. Kloepper, Biological control of *Fusarium wilt* on cotton by use of endophytic bacteria. Biol. Control 5 (1995) 83–91.
6. J. Hallmann, G. Berg, B. Schulz, Isolation procedures for endophytic microorganisms. in: B.E. Schulz, C.C. Boyle, T. Sieber (Eds.),Microbial Root Endophytes, Vol. 9, Springer, Berlin Heidelberg 2006, pp. 299–319.
7. C.A. Lozupone, R. Knight, Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. 104 (2007) 11436–11440.
8. N. Fierer, R.B. Jackson, The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 626–631.
9. J.G. Caporaso, C.L. Lauber, W.A. Walters, et al., Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. 108 (2011) 4516–4522.
10. J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, R. Knight, Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Methods 7 (2010) 813–819.
11. D.B. Rusch, A.L. Halpern, G. Sutton, et al., The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5 (2007), e77.
12. V. Lazarevic, K. Whiteson, S. Huse, et al., Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J. Microbiol. Methods 79 (2009) 266–271.
13. J.G. Caporaso, C.L. Lauber, W.A. Walters, et al., Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 6 (2012) 1621–1624.
14. S.M. Huse, L. Dethlefsen, J.A. Huber, D.M. Welch, D.A. Relman, M.L. Sogin, Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 4 (2008), e1000255.
15. J. Liu, J. Luo, H. Ye, Y. Sun, Z. Lu, X. Zeng, Production, characterization and antioxidant activities in vitro of exopolysaccharides from endophytic bacterium *Paenibacillus polymyxa* EJS-3. Carbohydr. Polym. 78 (2009) 275–281.
16. N.J. Loman, R.V. Misra, T.J. Dallman, C. Constantinidou, S.E. Gharbia, J. Wain, M.J. Pallen, Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30 (2012) 434–439.
17. E. Yergeau, S. Sanschagrin, C. Maynard, M. St-Arnaud, C.W. Greer, Microbial expression profiles in the rhizosphere of willows depend on soil contamination. ISME J. 8 (2014) 344–358.
18. A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, F.O. Glöckner, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41 (2013), e1.
19. Z. Liu, C. Lozupone, M. Hamady, F.D. Bushman, R. Knight, Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35 (2007), e120.
20. B.E. Van Wyk, B. Van Oudtshoom, N. Gericke, Medicinal Plants of South Africa. Briza Publications, Pretoria, 1997.
21. J.J. Kozich, S.L. Westcott, N.T. Baxter, S.K. Highlander, P.D. Schloss, 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 (2013) 5112–5120.
22. S.M. Huse, J.A. Huber, H.G. Morrison, M.L. Sogin, D.M. Welch, Genome Biol. 8 (2008) R143.
23. S.M. Huse, D.M. Welch, H.G. Morrison, M.L. Sogin, Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. 12 (2010) 1889–1898.
24. B.J. Haas, D. Gevers, A.M. Earl, et al., Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21 (2011) 494–504.
25. R.C. Edgar, B.J. Haas, J.C. Clemente, C. Quince, R. Knight, UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27 (2011) 2194–2200.
26. P.D. Schloss, D. Gevers, S.L. Westcott, Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6 (2011), e27310.
27. R.C. Edgar, UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10 (2013) 996–998.
28. J.-S. Huang, Ultrastructure of bacterial penetration in plants. Annu. Rev. Phytopathol. 24 (1986) 141–157.
29. A. Quadt-Hallmann, J. Kloepper, N. Benhamou, Bacterial endophytes in cotton: mechanisms of entering the plant. Can. J. Microbiol. 43 (1997) 577–582.
30. J.L. Azevedo Jr., J.O. Pereira, W.L. Araújo, Endophytic microorganisms: a review on insect control and recent advances on tropical plants. Electron. J. Biotechnol. 3 (1) (2000) 15–16.
31. G. Muyzer, E.C. de Waal, A.G. Uitterlinden, Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. 59 (1993) 695–700.
32. D.P. Herlemann, M. Labrenz, K. Jürgens, S. Bertilsson, J.J. Waniek, A.F. Andersson, Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J. 5 (2011) 1571–1579.
33. Illumina, 16S Metagenomic Sequencing Library Preparation Guide. http://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.ilmn 2013.
34. C.S. Miller, B.J. Baker, B.C. Thomas, S.W. Singer, J.F. Banfield, EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 12 (2011) R44.
35. S.H. Ong, V.U. Kukkillaya, A. Wilm, et al., Species identification and profiling of complex microbial communities using shotgun Illumina sequencing of 16S rRNA amplicon sequences. PLoS One 8 (2013), e60811.