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

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 In a similar study on culturable endophytic bacteria in the same A. vera plants (unpublished work), the dominance of Pseudomonas and Bacillus genera was similar, with equal proportion of four species each in root, stem and leaf tissues.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.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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
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 reads 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 their 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 SAMN03893981 (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.

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-phyllum 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

| Sample reference | Sample label | Sequence type | Sequence format | Read type | Read size (bp) | Total number of reads | Total sequence length (nt) | GCX |
|------------------|--------------|---------------|----------------|-----------|---------------|-----------------------|---------------------------|-----|
| 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                      |     |

Table 1

Raw reads statistics and sequence quality assessment of 16S rRNA sequence from A. vera 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 choice of primer pair to amplify a longer stretch of the 16S rRNA gene, thus, 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; *Proteobacteria, Firmicutes, Actinobacteria* and *Bacteroidetes*, 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
of the PCR amplicon were analysed by TECAN (Qiagen, Mississauga, Ontario, Canada). The quality and quantity gel electrophoresis and recovered using QIAquick gel extraction kit step at 72 °C for 1 min. The PCR products were con

20 s, 55 °C for 15 s, and 72 °C for 10 s and ended with an extension denaturation step at 95 °C for 3 min followed by 24 cycles of 98 °C for

and nuclease free water. PCR reactions were carried out with initial KAPA HiFidelity DNA polymerase (1 U), 1.5 μL of genomic DNA template, 12.5 μM 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 used for DNA extraction. Bacteria genomic DNA was extracted and checked for purity at A260/280 nm (1.9 ≤ A260/280 ≤ 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 5ʹ-CCTACGGGNGGCWGCAG-3ʹ and reverse 5ʹ-GACTACHVGGGTATCTA-3ʹ) [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 V2, 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 collection. Stud. Mycol. 50 (2004) 9–18.

2. D. Bulgarelli, K. Schlaepfer, S. Spaen, 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. A.A. Strobel, Endophytes as sources of bioactive products. Microbes Infect. 5 (2003) 555–544.

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, C. Maynard, M. St-Arnaud, C.W. Greer, Microbial expression profiles in the rhizosphere of willows depend on soil contamination. ISME J. 8 (2014) 544–558.

8. A. Kindworth, E. Pruesse, T. Schweer, J. Peplis, C. Quast, M. Horn, F.O. Glockner, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41 (2013), e1.