Diversity Analysis of Endophytic Bacterial Microflora in *Emilia sonchifolia* (Linn.) DC on Illumina Mi Seq Platforms

Sithara K. Urumbil and M. Anil Kumar*

1Department of Botany, Little Flower College, Guruvayoor - 680 103, Kerala, India. 2Cell Culture Lab, Department of Botany, Union Christian College, Aluva, Ernakulam, Pin-683 102, Kerala, India.

Abstract
Bacterial endophytes inhabiting medicinal plants are less explored, but are diverse and play crucial roles in regulating growth and development of the host. Metagenomics using Illumina MiSeq platform facilitate whole community level characterization. The present study reports the diversity of bacterial endophytic microflora from the medicinal plant *Emilia sonchifolia* (Linn.) DC. Metagenomic analysis of medicinal plants leads to the identification of novel organisms or genes which will help the correlative elucidation of plant-microbe interactions. Effective sequences were amplified from 16S rRNA gene V3-V4 variable region. OTU analysis at different taxonomic level clearly catalogues two Phyla viz. Proteobacteria and Firmicutes which belonged to Gammaproteobacteria and Bacilli. In these classes five orders such as Enterobacteriales, Pseudomonadales, Xanthomonadales, Bacillales and Betaproteobacteriales were detected. Among these orders five families were identified in which the most predominant was Enterobacteriaceae and Pseudomonadaceae while the other three families viz. Xanthomonadaceae, Planococcaceae and Burkholderiaceae were less represented. At genus level very less number of bacteria were identified while a bulk majority remained unclassified. Of the seven identified genus the most prominent one was *Pseudomonas* followed by Stenotrophomonas, Cronobacter, Lysinibacillus, Pantoea, Kluvyera and Pseudorhodoferax. At species level only two were identified vz. *Pseudomonas otitidis* and *P. geniculata*. Alpha diversity analysis using various statistical indices like Simpson and Shannon explains the diversity of microbiome. Next generation sequencing survey of DNA sample extracted from host plant through metagenomic data screening identified different endophytic bacteria which are difficult to grow in culture conditions.

Keywords: Illumina MiSeq, metagenomics, *E. sonchifolia*, medicinal plants, endophytic bacteria

*Correspondence: drmakumar@gmail.com

(Received: January 03, 2020; accepted: February 26, 2020)
INTRODUCTION

In the past decades, the microbiota associated with plants gained researcher’s attention due to their wide diversity and beneficial applications of plant-microbe interactions. They may be epiphytic or endophytic and their presence can be detected in the rhizospheric or phyllospheric regions (Strobel and Daisy, 2003). Endophytes reside inside the plant tissues asymptotically and their association can be symbiotic or mutualistic. The endophytic communities of various medicinal plants like Cannabis (McKernan et al., 2018), Aloe vera (Akinsanya et al., 2015), Plectranthus (El-Deeb et al., 2013), Catharanthus, Mentha, and Ocimum (Anjum and Chandra, 2015) were isolated and identified. From these earlier reports it was observed that the interaction between endophytes and host plants play some roles in their medicinal properties.

The understanding of plant-microbe interactions is expanding due to the advancements in next generation sequencing (Lozupone and Knight, 2007). New sequencing platforms like Illumina MiSeq enables tracking of large microbial communities at a faster rate and lower cost (Ram et al., 2011). In the analysis of complex microbial communities like endophytic populations metagenomic sequencing acts as an advanced technique when compared to rDNA sequencing. In this context, Illumina sequencing provides information with fewer errors on biodiversity analysis of microbes (von Mering et al., 2007).

Emilia sonchifolia (Linn.) DC. is a herbaceous medicinal plant used for the treatment of various inflammatory disorders (Shylesh et al., 2000, Chopra et al., 1986, Essien et al., 2009). The multifaceted applications of this medicinal herb has been analysed by many researchers, but the information on the endophytic microbiome and its metataxonomy remain unclear. The present investigation aims to analyse the endophytic bacterial diversity of E. sonchifolia using metagenome analysis through Illumina sequencing platform.

MATERIALS AND METHODS

Plant sample collection and surface sterilization

The whole plant of E. Sonchifolia during its flowering season (July-October) was collected from fourteen districts of Kerala (Kasargode 12° 30′ N75° 00 E to Thiruvananthapuram 8° 29′ N76° 59′ E), India, pooled and used for DNA extraction after surface sterilization. E. sonchifolia is a herbaceous medicinal plant which is one among the members of ‘Dasapushpa’ (ten flowers of sacred value) in Ayurvedic medicine. It belongs to the family Asteraceae of dicots and possess anti-cancerous, antiinflammatory and analgesic properties (Essien et al., 2009; Shylesh et al., 2000). The whole plant has been used for medicinal purposes and hence the biodiversity analysis was carried out using the entire plant after surface sterilization. Healthy and mature flowering plant was washed thoroughly in running tap water followed by dipping in sterile double distilled water for ten minutes and surface sterilized with 0.1% mercuric chloride for one minute. The surface sterilized material was then rinsed in sterile distilled water and dipped in 70% ethanol for 60 seconds. Further immersion of the plant tissue in distilled water was required to remove the traces of ethanol from the tissue. The effectiveness of the surface sterilization procedure was validated by culturing the final wash into nutrient agar plates. Any bacterial growth on the control plates indicates inadequate surface sterilization.

DNA Isolation and PCR Amplification

DNA was extracted using Purelink genomic DNA extraction and purification kit (Invitrogen, Life Technologies, USA) following the manufacturer’s instruction. After electrophoresis dsDNA concentration was checked by Qubit®4.0 fluorometer. The V3-V4 hyper variable regions from 16S rRNA gene of the purified DNA were amplified using the universal forward primer-5’ CCTACGGRRBGCASCAGKVRVGAAT 3’ and reverse primer-5’GGACTACNVGGGTWTCTAA GCC3’.

The amplified DNA was further quantified with Qubit®4.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). 30-50ng DNA sample along with Meta Vx™ Library preparation kit was used to prepare the amplicons. The PCR amplicons were tagged with adapters for creating indexed libraries. Validation of DNA libraries were done by Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) and quantified by Qubit®4.0 Fluorometer. DNA libraries were multiplexed and loaded on an Illumina MiSeq instrument (Oomics Gen Life Sciences Pvt. Ltd.) according to manufactures
instructions (Illumina, 2013, Illumina, San Diego, CA, USA). Sequencing was performed using a 2x250 paired-end (PE) configuration. Image analysis and base calling were performed by the MiSeq control software (MCS) with in the MiSeq instrument. The original image data were analysed using bcl2Fastq (V2.17.1.14) and the result was stored in FASTQ format (Ong et al., 2013).

**Sequence Data Analysis**

The forward and reverse reads were joined and separated on the basis of barcodes. After merging and separation the primers, adapters, barcodes and undetermined bases were removed. This reduces sequencing errors and the analysis software used was Cutadapt [V (1.9.1) V search (1.9.6), Qiime (1.9.1)]. Quality filtering on joined sequences was performed and the sequences without any ambiguous bases, length<200bp, mean quality score≥20 were selected for downstream analysis. As a final step for the preparation of clean data for further analysis chimera sequences were identified by comparing with reference database (RDP Gold database) using UCHIME algorithm and they were removed.

**OTU analysis**

Sequences were grouped into different Operational Taxonomic Units (OTU) based on 97% sequence identity using the clustering Programme V SEARCH (1.9.6) against the Silva 119 database. Taxonomic category at confidence threshold of 0.8 was grouped into different OTU with the help of Ribosomal database programme classifier.

**Alpha diversity analysis**

Alpha diversity analysis was performed to check the microflora biodiversity. The different species in the microbial community and its abundance were calculated through a series of statistical indices like ACE (the number of species), Shannon (Diversity index alpha for the estimation of microbial diversity), Simpson (Quantify biological diversity) and Goods coverage (library coverage of each sample). The analysis software used for the alpha diversity indices was Qiime (1.9.1).

**RESULTS AND DISCUSSION**

*E. sonchifoila*, a major medicinal plant of ‘Dasapushpa’ category of Indian traditional medicine, was used for the present analysis. Relative abundance, composition and diversity of endophytic microflora were analysed on Illumina base sequencing platform.

**Preliminary Sequencing Data Statistics**

Raw read statistics and sequence quality assessments were collected from MiSeq sequencing reporter generated through base calling. Preliminary sequencing data statistics were presented in Table 1. The sequence data has been deposited at NCBI under Sequence Read Archive (SRA) database with accession no. PRJNA542222.

**Sequencing data quality optimization**

The sequenced data was further processed by analysis software Cutadapt [V (1.9.1)] and Qiime (1.9.1) for the removal of unclassified bacterial reads and the elimination of low complexity reads. Sequence processed was detailed in Table-2, and sequence reads were generated after carry-on overlap splicing through the merging of forward and reverse reads. Alpha diversity analysis require error free sequences and hence it is necessary to remove reads with undesirable length, low quality score, and ambiguous base calls (Ns) (Huse et al., 2008). Reads were trimmed on the

---

**Table 1. Preliminary data statistics**

<table>
<thead>
<tr>
<th>Sample Ref No.</th>
<th>Sequence Format</th>
<th>Sequence Type</th>
<th>Read type</th>
<th>Read size</th>
<th>Total no. of reads</th>
<th>Total sequence length</th>
<th>Q 20</th>
<th>Q30</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_01</td>
<td>Fastq</td>
<td>Illumina MiSeq</td>
<td>Paired end</td>
<td>250bp</td>
<td>155860</td>
<td>38965000</td>
<td>91.09</td>
<td>87.03</td>
<td>54.77</td>
</tr>
</tbody>
</table>

**Table 2. Sequencing data Quality optimization**

<table>
<thead>
<tr>
<th>Sample</th>
<th>#PE_reads</th>
<th>#Nochimera</th>
<th>Avg Len(bp)</th>
<th>GC(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_01</td>
<td>77930</td>
<td>63453</td>
<td>463.45</td>
<td>54.48</td>
</tr>
</tbody>
</table>
above mentioned criteria to reduce the error rate. Out of the 155860 high quality reads obtained, 77930 reads were generated after these filtering processes with average length of 463.45bp (Graph I).

**OTU and Taxonomic Composition Analysis**

The accurate and high resolution microbiota profiling of endophytic communities can be elucidated through Illumina sequencing platform. All the sequences after filtering and error reduction were classified to obtain information on species and genus. The grouping of sequences relayed on the 97% identity threshold for data statistics and analysis using software Qiime (1.9.1). The number of OTUs of sample was six which is comparatively low with other published metagenomic sequence analysis indicating the host specificity of the endophytic population. The OTU distribution pattern indicated that majority of endophytic bacteria are included under *Enterobacteriaceae*.

The statistics of the number of species on different taxonomic levels revealed the vast diversity of microbes. The sequence data disclosed phylum *Proteobacteria* as the predominant taxa followed by *Firmicutes* and *Bacteroidetes*. It was interesting to note that the major phylum detected was *Proteobacteria* but when tried to isolate endophytes on culturable methods the major group was *Bacilli* from phylum *Firmicutes* (Urumbil and Anilkumar, 2019).

Few studies have been conducted in medicinal plants for analyzing the biodiversity of bacterial endophytes using Illumia Mi Seq. Illumina based sequencing were employed for the screening of endophytic bacterial diversity in the medicinal plant tree Peony and they reported Proteobacteria, *Firmicutes*, *Bacteroidetes*, Actinobacteria and acidobacteria as dominat groups of bacterial endophytes(Yang et al., 2017). *Panax notoginseng* a medicinal plant with antihypertensive, antithrombic and neuroprotective bioactivities were screened for the diversity analysis of bacterial endophytes associated with the plant using the QUIIME Pineline and indicated the presence of Proteobacteria, Actinobacteria, *Bacteroidetes* and *Firmicutes* as major communities of bacterial endophytes in this plant (Dong et al., 2018). Previous works on endophytes by Sessitsch et al., (2012) reported

<table>
<thead>
<tr>
<th>sample</th>
<th>ace</th>
<th>chao1</th>
<th>shannon</th>
<th>Simpson</th>
<th>goods_coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-01</td>
<td>14</td>
<td>14</td>
<td>2.221</td>
<td>0.69</td>
<td>1</td>
</tr>
</tbody>
</table>

Graph I. Effective sequence length distribution

---

Table 3. Alpha diversity through a series of statistical indices
the dominance of phylum Proteobacteria in the endophytic community. In accordance with the results of Romero et al., (2014), in the present analysis also threw light on the occurrence of Gammaproteobacteria comprising Enterobacterials and Pseudomonadales as key stone taxon.

The endophytic bacteria detected from the medicinal plant E.sonchifolia were classified under the orders Enterobacteriales, Psuedomonadales, Xanthomonadales, Bacillales and Betaproteobacteriales. This clearly explained the diversity of endophytic microbiota and similar results were obtained in Poplar trees by Moore et al., (2006). The most dominant genus reported in the present study is Pseudomonas. The other dominant genera were Stenotrophomonas, Cronobactor, Lysinibacillus and Pantoea. Liu et al. (2016 ) reported the dominance of the phylum actinobacteria, proteobacteria and firmicutes in the diversity study of endophytic bacterial populations from Ferula songorica. Several members of Enterobacteriales, Psuedomonadales, Bacillales and other orders of bacteria identified in the present study were also reported by various researchers with special mention to the endophyte-host interactive growth promotion (Brigido et al., 2019). These endophytes were found to be effective in preventing the development of disease causing pathogenic fungus in Agave tequilana (Martinez-Rodriguez et al., 2014) and they check the development of wilt disease in Pine Trees (Proença et al., 2017). Role of endophytes in helping host plants for existing in adverse environmental conditions were analysed and the Bacillaceae and Enterobacteraceae were detected as keystone taxa from plants growing in extreme environmental conditions.

A wide variety of bioactive compounds were produced by endophytic bacteria that included the class Gamaproteobacteria and Bacilli points to the significance of studying these categories of bacteria as endophytes (Pimentel et al.,2011 and Gouda et al., 2016). Studies on the medicinal plant Latana camara showed the presence of Bacillus as a prominent member in the endophytic bacterial diversity (Janardhan et al., 2012). Zam et al., (2019) recently reported
the occurrence of the genus *Bacillus* as major endophytic bacteria from 8-10 medicinal plants. The endophytic bacterial classes present in this study were found to have different levels of interactions with the host plant and they can have wide range of applications. The complete genome sequencing analysis of *Enterobacter* sp. spotted a number of genes which act as key factors for the dualistic life cycle as soil bacterium and endophyte (Andrés-Barra et al., 2017). These results support our study on the occurrence of *Enterobacteriales* as endophytes was common and even their genome analysis proves this factor.

The microbial composition of the sample at different taxonomic level (Phylum, Class, Order, Families, and Genus) was plotted in stalked bar plots (Fig. 1). Phylogenetic tree of the genus constructed and infers approximately maximum likelihood from alignments of the major OTU sequences (Fig 2).

**Alpha diversity analysis**

Alpha diversity is mainly used to reflect the species diversity in a single sample through a series of statistical indices like ACE, Chao I, Shannon, Simpson and Good’s coverage using the software Qiime (1.9.1) (Table 3). The rarefaction curve is a useful tool for the species composition characterization in a sample and it predicts the abundance of species in it. It determines whether

![Graph 2. Rarefaction curve: The X axis is the valid sequences extracted, and Y axis is the number of OTUs. The number of OTUs increases with the increase of extracted sequence count until reaching a plateau, which indicates the number of detected OTUs will not increase with the amount of extracted sequences and reflects the reasonable sequence depth.](image)
the sample size is sufficient to estimate the species abundance in biodiversity and community surveys. The alpha diversity and rarefaction analysis (Graph 2) focus on the OTU at 97% similarity and points that bacterial communities were not completely revealed and it was diverse with higher taxonomic richness.

Microflora biodiversity analysis along with metagenome studies on Illumina MiSeq platform enables the proper identification of bacterial population. According to Shi et al., (2014) endophytes are symbiotic microorganisms and their genetic diversity within the host depends on various other parameters like host genotype, environmental conditions, age of host etc, and was studied in detail on cotton plats by Adams and Kloepper (2002). They also mentioned that the endophytic bacterial diversity studied so far relayed on the culturable communities which is almost <1% of the bacterial endophytic species present. Identity and cellular interactions of these endophytic microbes can be deciphered through metagenomic analysis both in the case of culturable and unculturable bacterial communities (Dinsdale et al., 2008). So this particular study helps for the further identification of some specific endophytic candidates which can contribute to the medicinal properties of this particular host plant E. sonchifolia.

CONCLUSION
Multidisciplinary research approaches were required for elucidating the beneficial aspects of plant microbe interactions. Metagenomic data analysis is an inevitable complementary information illustrating complex network of factors controlling endophytic colonization and its association with the host. The present study revealed an average read length of 463.45bp from the V3-V4 hypervariable region of 16S rRNA sequences. OTU analysis at different taxonomic level clearly catalogues two Phyla viz., *Proteobacteria* and *Firmicutes* which belonged to *Gammaproteobacteria* and *Bacilli*. In these classes five orders such as *Enterobacteriales*, *Pseudomonadales*, *Xanthomonadales*, *Bacillales* and *Betaproteobacteriales* were detected. Among these orders five families were identified in which the most predominant was *Enterobacteriaceae* and *Pseudomonadaceae* while the other three families viz. *Xanthomonadaceae*, *Planococcaceae* and *Burkholderiaceae* were less represented. At genus level very less number of bacteria were identified while a bulk majority remained unclassified. Of the seven identified genus the most prominent one was *Pseudomonas* followed by *Stenotrophomonas*, *Cronobacter*, *Lysinibacillus*, *Pantoea*, *Kluyvera* and *Pseudorhodoferax*. At species level only two were identified viz. *Pseudomonas otitidis* and *P. geniculate*. Comprehensive knowledge of both culturable and non-culturable endophytes from medicinal plants can unveil the presence of novel compounds and the genes associated with them. In future this can be used as a platform for the development of new drugs. Whole Metagenome data analysis of these endophytes is in progress in our lab and can help in the identification of genes with wide range of applications in biotransformation process.

ACKNOWLEDGEMENTS
The first author gratefully acknowledges the financial support and leave under FDP by the University Grants Commission (UGC), New Delhi, Government of India.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

FUNDING
None.

AUTHORS’ CONTRIBUTION
Both authors have made a substantial direct and intellectual contribution to the work and approved for publication.

DATA AVAILABILITY
All data sets generated during this study are included in the manuscript.

ETHICS STATEMENT
This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES
1. Adams PD and Kloepper JW. Effect of host genotype on indigenous bacterial endophytes of Cotton (*Gossypium*

![Image](https://via.placeholder.com/150)


8. El-Deeb B, Fayez K and Gherbawy Y. Isolation and characterization of endophytic bacteria from Agave tequilana vara, S1517-83822014000400025


27. Shylesh BS and Padikkala J. *in vitro* cytotoxic and antitumour property of *E. sonchifolia* (L.) in Mice, *J.Ethnopharmacol.*, 2000; **73**: 495-500. https://doi.org/10.1016/S0378-8741(00)00317-2


