Asian Jr. of Microbiol. Biotech. Env. Sc. Vol. 25, No. (4) : 2023 : 818-826

ISSN-0972-3005
DOI No.: http://doi.org/10.53550/AJMBES.2023.v25i04.040

ESTABLISHMENT OF A BIOINFORMATICS PIPELINE FOR THE DETECTION OF PATHOGENIC BACTERIA

RAJAPAKSHA R.W.P.M. [1]*, VIVEHANANTHAN K.[2] AND ATTANAYAKA D.P.S.T.G.[1]

[1]Department of Biotechnology, Faculty of Agriculture & Plantation Management, Wayamba University of Sri Lanka, Makandura, Gonawila (NWP), Sri Lanka
[2]Department of Basic Sciences, Faculty of Health Sciences, The Open University of Sri Lanka

(Received 20 August, 2023; Accepted 12 October, 2023)

Key words: *Bioinformatics, 16s Metagenomics, Microbiome, Pipeline Development, Seed Potato*

Abstract– Next-generation sequencing-based methods using partial 16S rRNA gene amplicons are extensively applied today in studies of the plant metagenomes. NGS sequencing creates huge sets of raw data making analysis a challenging task. Lack of computational and bioinformatics knowledge and tools for analyzing high throughput data to interpret correct biological variations is a major problem. In addition, downstream analysis of NGS data with the available bioinformatics platforms create various challenges in inferring microbial composition. The available commercial software are expensive and individual open-source tools are usually operate stand alone as they are not combined for a user-friendly workflow. Therefore, beginners in bioinformatics might find analysis procedures are complicated, expensive, and time-consuming with the associated learning. In the present study, a bioinformatics pipeline is developed to analyze the 16S rDNA amplicons of plant metagenome. Microbial DNA was extracted from imported seed potato tubers. Extracted DNA was sequenced using Ion Torrent Next Generation Sequencing technology by amplifying 400 bp V1-V2 region of 16S rRNA gene for the detection of bacterial pathogens. The pipeline was built by stringing together many command line tools; Quality checking of raw fastq data using FastQC, trimming of low-quality data with Trimmomatric, alignment of trimmed data using the BWA-MEM algorithm, removal of duplicate reads with Picard Mark Duplicates tool and finally generation of the phylogenetic tree and taxonomic profile with MEGAN 5. The developed pipeline was used to analyze the 16S rRNA next generation sequences and the reliability of the results has been checked with the use of mock communities for validation. The pipeline often can be executed on laptop sized machines to obtain the output in a couple of hours enabling easy access for the researchers.

INTRODUCTION

Emerging, re-emerging and endemic plant pathogens continue to challenge our ability to safeguard plant health worldwide (Miller *et al.*, 2009). Plant diseases cause massive losses in agriculture (Lacombe, 2010), nearly 20% yield loss in food and cash crops worldwide (Doornbos, Van Loon and Bakker, 2011). Pathogens are important yield limiting factors, requiring the need for advanced disease detection and prevention measures to minimize pathogen damage to plants (Dong, 2021). Novel bioinformatic tools have been opened up pathways for the low-cost rapid identification of pathogens and prevention of diseases (Dong, 2021).

The new sequencing technologies provide a big impact in plant metagenomics and revolutionize diagnostics, epidemiology, and infection control (Studholme *et al*., 2011). Metagenomics has the capability of exploring uncharacterized plant microbiome and disease systems through Next-generation sequencing technologies. This offered the opportunity of cultivation-independent assessment of microbial communities and therefore revealed a multitude of thus far unknown bacteria (Graspeuntner, 2018). Advancements in NGS technologies regarding throughput, read length and accuracy had a major impact on microbiome research by significantly improving 16S rRNA amplicon sequencing. Metagenomics has the capability of exploring uncharacterized plant microbiome and disease systems through Next-generation sequencing technologies. Applications of these high-throughput sequencing methods that are relevant to phytopathology, with the associated

computational and bioinformatics tools could overcome the challenges for microbe detection.

As rapid improvements in sequencing platforms and new data analysis pipelines are introduced, it is essential to evaluate their capabilities in specific applications (Allali *et al.*, 2017). However, Methods and software requirements for analyzing this data to interpret correct biological meaning are not experiencing the same growth rate (Naranpanawa *et al.*, 2020). Metagenomics take the benefit of NGS for the analysis of microbial populations by exploring the whole nucleotide sequences of a DNA sample (Cuadros Orellana, 2013) and bioinformatics analysis holding great potential to categorize the entire microbial range in uncharacterized plant disease systems.

Sequencing of 16S rRNA amplicons is now a well-established and robust method (Dixit, 2021). Many databases and tools available for classification and taxonomic assignment of the 16S rRNA gene make it challenging to select the best-suited method for a particular dataset. Even though sequencing technologies have advanced rapidly in a short span of time, methods and software used for assembly and analyses of sequence data have not seen the same degree of improvement. While most of these tools are still being revised for better algorithmic approaches and efficiency (Naranpanawa, 2020). In this study a pipeline was developed to analyze the NGS data of plant pathogens due to its importance to detect plant pathogens as well as to screen quarantine pathogens invading to the country with the plant commodities. Although many bioinformatics analysis pipelines are available online, researchers have to face the difficulties in uploading large sequence files together with the slow processes.. Therefore, inbuilt pipelines are advantageous over other methods to detect the bacteria efficiently. Bioinformatics pipelines will enable precise identification of the causal pathogens along with the categorization of the main incident diseases. This can also uncover previously unknown or undetermined pathogens or unculturable species (Monteiro *et al.*, 2015). This study is focused on developing a bioinformatics pipeline to analyze the 16S rDNA amplicons to detect bacteria. It is currently implemented in the Biotechnology laboratory of Wayamba University of Sri Lanka for analyzing plant metagenomes and specially for testing pathogenic bacteria associated with the imported seed potato consignments to Sri Lanka. This is a user-friendly, validated assembly pipeline is using free bioinformatics software and tools. This can be executed in lapto sized machines and the interested researchers can follow the given workflow to install the pipeline in their personal laptops and desktop computers for their microbiome analysis purposes.

MATERIALS AND METHODS

Plant Material Collection

The seed potato tubers randomly sampled from the imported consignments at the entry ports of Sri Lanka were collected from the pathology division of the National Plant Quarantine Service- Katunayake, Sri Lanka. The seed potato tubers within the consignments were not found to be infected with any quarantine-important disease by visual observation. The collected seed tubers were separately stored in a cold room under 4 ^{0}C at the National Plant Quarantine Service.

Microbial DNA Extraction from the Imported Potatoes

The seed tubers were surface sterilized with 10% NaOCl and then washed with distilled water prior to DNA extraction. Ten grams (10 g) of each seed potato tuber (especially tissues from the stem end and eyes, including both the peel and inside tissues) was crushed and transferred to a conical flask with 15 ml of sterile liquid LB medium. The mixture was incubated at room temperature while shaking for 12 hrs. under 120 rpm. The obtained turbid liquid culture was filtered in a muslin cloth to remove potato tissues and the filtrate was centrifuged for 10 min. at 12 400 g under 4 ^{o}C to obtain the microbial pellet. The supernatant was removed and the pellet was washed twice with 500 µl of wash buffer (50 mM Tris-HCl, 5 mM EDTA, pH 8.0) and centrifuged again at 12,400 × g for 10 min. The supernatant was removed and an aliquot of 500 µl of lysis buffer (100 mM Tris-HCl, 100 mM EDTA, 1.5 M NaCl, pH 8.0) was added to the pellet. The suspension was centrifuged for 15 min. at 12 400 g under 4^{o}C. The pellet was removed and the supernatant was transferred to another Eppendorf tube. 75 µl of NaOAC and 500 µl of ice-cold isopropanol were added to the supernatant along the wall of the tube. It was centrifuged for 15 min. at 12 400 g at 4 ^{o}C. The supernatant was removed and the pellet was washed with 70% ethanol and re-centrifuged at 12 400 g for 10 min. at 4^{o}C. Ethanol was completely removed by air drying. The DNA was re-suspended

in 50 µl of sterile de-ionized water.

DNA sequencing by Ion Torrent Next Generation Sequencing

Ion PGM™ HiQ™ OT2 Kit (Cat. No. A27739) was used for library preparation targeting the V1-V2 region of 16s rRNA gene. Library preparation was carried out by Credence Genomics Pvt. Ltd. Colombo Sri Lanka .

The Ion PGM HiQ Sequencing Kit which includes reagents and materials for sequencing upto 400 bp inserts was used together with Ion PGM HiQ template preparation kit for sequencing V1-V2 region of the 16SrRNA gene. The Ion 318 Chip v2 was used on the Ion PGM System. DNA sequencing was performed on an Ion Torrent sequencing machine at Credence Genomics Pvt. Ltd. Colombo, Sri Lanka. Seventeen DNA samples were sequenced for bacterial identification. Hypervariable region of bacterial 16S rRNA gene was amplified using oligonucleotides, prepared into single-ended libraries of 400 bp size followed by Next Generation Sequencing on Ion Torrent PGM. Sequences of seventeen samples were recovered from ion torrent personal genome machine and aligned with the 16S rDNA of representative fragments. The sequences generated in this study were deposited in the National Center for Biotechnology Information SAMN35449846, SAMN35449893, SAMN35449894, SAMN35450266, SAMN35451683, SAMN35453206, SAMN35453207, SAMN35453263, SAMN35453300, SAMN35453302, SAMN35453303, SAMN35453304, SAMN35453305, SAMN35453436, SAMN35453444, SAMN35453446, SAMN35453447

Development of Workflow

Open-source bioinformatics tools have been used to develop the workflow for the pipeline for the present study. BASH scripting language used to pipe was described in **Figure 1**.

Computational Bioinformatics Analysis

A functional laptop with 8 GB RAM was used for running the computational pipeline using Ubuntu 14.04 operating system (Debian based operating system). Raw fastq files were subjected to the pipeline for analyzing plant pathogens (Figure 1). Trimming of low quality data was carried out based on FastQC report with Trimmometric 0.35 tool, with LEADING: 10, TAILING: 10 and SLIDING WINDOW 4:15 (Window size: Average Quality) and the trimmed data was aligned with BWA index, that was created with Silva_ 123. I SSURef tax trunc.fasta using BWA-MEM. Allinged SAM file was subjected to Picard tool 2.1.0 for removal of PCR duplicates and other duplicates reads using MarkDuplicate option. Finally, phylogenetic tree and taxonomic profile were created with MEGAN 5.11.3 with its option import BLAST results, by using Synonymous Map file as taxmap_ncbi_ssu_ref_123.1.txt and LCA parameters (minScore=10.0, maxExpected=0.1, minSupport=10 and lcaPercent=l 00.0) and its other default parameters.

NGS Data Analysis procedure using developed Pipeline

Testing of Raw Data Quality

- Open terminal
- Type "metapipe" on terminal
- Select data type (single end=1 / Paired end=2) enter selection one for single end reads and enter selection 2 for paired end reads

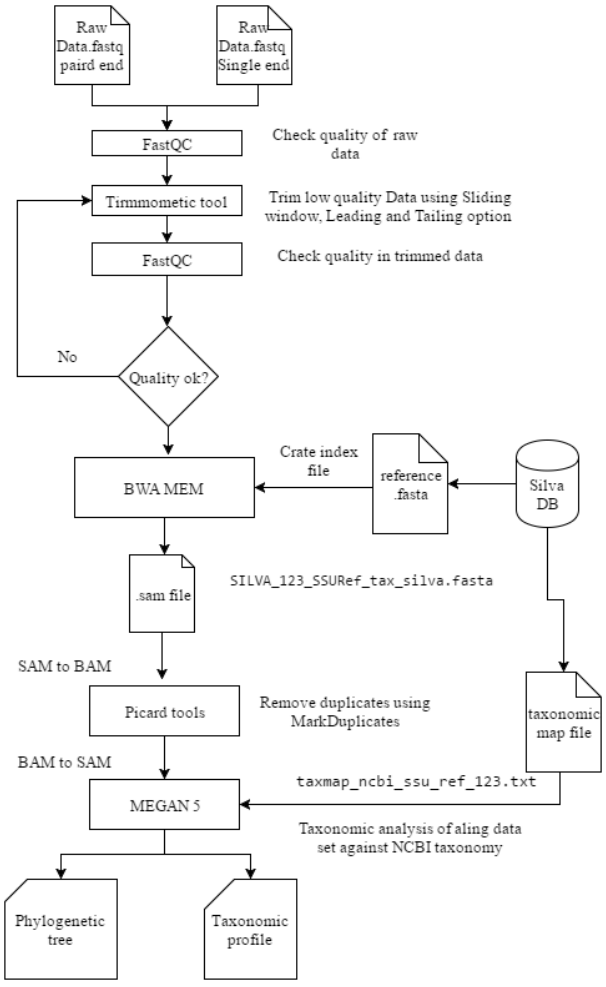

Fig. 1. Workflow for pipeline development

- Enter the file location or drag the FASTQ file to this location
- FASTQC Report will be automatically opened
- Go to the terminal again. It asks : Do you want to run the trimming process? (1=YES, 2=NO) enter selection one to run trimming process.
- It asks: Maximum core that you can allocate for process? Type 3 (to run this pipeline we were using a Core I5 computer having 4 cores. Therefore we wanted to allocate 3 cores for analysis process)
- It asks: Maximum memory you can allocate? Type 3 (our computer was having a 4GB RAM, therefore, we allocated 3GB's for analysis process)
- Enter quality value or leading? Type 20 and press enter (quality value 20 means : one base pair is wrong for 100 base pairs)
- Enter quality value for trailing? Type 20 and press enter
- Enter window size? Type 4 and press enter
- Enter required quality? Type 20 and press enter
- Minimum length of reads? Type 0 and press enter for single end reads (this option is for analysis of paired end reads)
- Trimmometric process finishes at this point. We can repeat this step by increasing the quality values

Testing the quality of trimmed DATA

- Run FASTQC for trimmed data? Type 1 and press enter to run FASTQC
- FASTQC Report opens
- Go to the terminal again
- It asks: Do you want to run the Trimmometric again? Type 2 and press enter (1=YES, 2=NO)
- BWA Starts (BWA is the name of the algorithm for aligning. Aligning starts at this point)
- After few minutes/hours we can see the message on the screen "BWA Done"
- Open terminal , type "picard" and press enter
- Select the file location or drag and drop the align.sam file generated in BWA (eg. If the name of the file we analysed was wyb2, a file named wyb2_align.sam is automatically generated within new wyb2 file. Drag and drop this fie here

MEGAN 5

- At this step MEGAN 5 starts
- Go to: File # Import rom Blast # Open the final file (eg. Wyb2.finalSAM) #NEXT
- Load the synonyms mapping file (This is a file named taxmap_ncbi_ ssu_ref_123.1.txt) select # open # Apply
- Get the phylogenetic trees and taxonomy tables of the analyzed data set.

RESULTS

Bioinformatic processing of the resulting millions of DNA sequences can be challenging, and a standardized protocol would aid in reproducible analyses. The 16S rRNA gene sequencing pipeline developed in this study is an automated 16S rRNA gene sequences processing tool. The pipeline was designed by combining pre-existing tools into a computational pipeline. This pipeline automates the processing of raw 16S rRNA gene sequencing data to create readable tables, graphs, and figures to make the collected data more readily accessible.

After validation, this computational pipeline was used for analysis of Next Generation Sequence data of potato DNA which revealed the endophytic pathogenic and non pathogenic bacteria associated with the tubers, which is described in this study. Currently, This pipeline is implemented in the Biotechnology laboratory of wayamba university of sri lanka and is being used for quarantine pathogen screening and 16s Next Generation Sequence analysis purposes.

DNA extraction from Potato tubers and amplification of V1-V2 region

In this study, microbial DNA extracted from the seed potato tubers was used for next generation sequencing and analysis (Figure 2). The extracted DNA resulted successful amplification with 16s rDNA V1-V2 region used for Next Generation Sequencing (Figure 3).

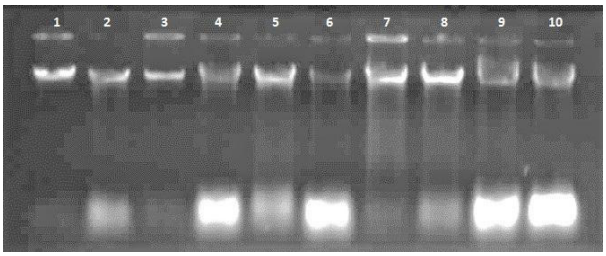

Fig. 2. DNA extracted from seed potato samples. Lane 1-10: DNA samples

Analysis of NGS data using Computational pipeline

FastQC

FQC is software that facilitates quality control of

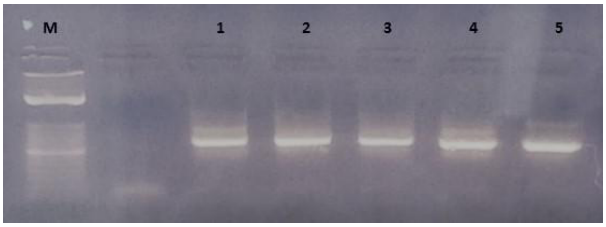

Fig. 3. Amplification of V1-V2 region of 16S rDNA from Microbial DNA extracted from seed potato M: 100bp Marker; 1,2,3,4: DNA samples; 5: Positive control

FASTQ files by carrying out a QC protocol using FastQC, parsing results, and aggregating quality metrics into an interactive dashboard designed to richly summarize individual sequencing runs (Brown *et al.,* 2017). First, some simple quality control checks were performed with FastQC to ensure that the raw data stands of good quality before analyzing the raw sequence. FastQC (Andrews, 2010) is a computational tool that provides a quick impression of raw sequence data coming from any sequencing platform (various platforms exist such as Solexa, 454 Roche, Illumina and Ion Torrent). The FastQC analysis execute quality control (QC) on FASTQ files by aggregating QC data like per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, adapter content and Kmer content (Supplementary Figure 1). This enables to detect problems in sequences of raw NGS data and gives quick impression of quality distribution of NGS data. Additionally, FastQC access GC content, over-abundance of adapters and over represented sequences, which gives an idea of sequence quality and PCR duplications (Supplementary Figure 1). The pipeline is able to report a wide range of information related to the quality profile and produces a per-sample summary as a PDF file containing all the main FastQC plots.

Trimmomatic Tool

The presence of poor quality or technical sequences such as adapters in next-generation sequencing (NGS) data can easily result in suboptimal downstream analyses (Bolger *et al.,* 2014). Therefore, Trimmomatic tool was used for trimming and filtering of raw sequences generated in this study. this includes a variety of processing steps for read trimming and filtering, but the main algorithmic innovations are related to identification of adapter sequences and quality filtering (Bolger, Lohse and Usadel, 2014). Trimmomatic tool contains a variety of processing steps to generate quality data from raw NGS data. Normally, adapters like short read fragment with low quality are removed from Illumina like NGS platform, but in practice, this process is not perfectly effective for analysis work. Deletion of technical sequences (adapters, PCR primers) and quality filtering using both Palindrome mode and Sliding Window quality filtering of the Trimmomatic tool were used to improve quality of raw sequence data further.

BWA-MEM algorithm

Next, reference-based alignment carried out for trimmed data with the BWA-MEM algorithm (Heng Li and Durbin, 2009) produced SAM output by aligning with reference index that created with BWA algorithem at fast and memory-efficient way. Other possible algorithms were also tested during workflow development, for their performance using NCBI Blastn (Altschul *et al.,* 1990) and Bowtie2 (Langmead and Salzberg, 2012). However, BWA-MEM was found to be an efficient algorithm as it can often be executed on laptop sized machines in a couple of hours compared to others. Moreover, it produced significant for short reads with long reference like 16S analysis.

Alignment

Reference-based alignment works well if the metagenomic dataset contains sequences which closely match the reference genomes for microbes. Different databases are available such as Silva rRNA database, Greengens, and NCBI 16S rRNA project. For the present study, Silva rRNA gene database project (Quast *et al.,* 2013) was selected as it contains upto date, quality-controlled databases of aligned ribosomal RNA (rRNA) gene sequences for bacteria, achaea, and eukaryotes. Taxonomic mapping files of the particular database are also needed to generate a taxonomical distribution of the pathogens present in the data.

Picard tool

Picard tool (http://www.picard.sf.net) Mark Duplicates option used to remove PCR duplicate reads of output that was obtained from the alignment algorithm to avoid duplicate reads in phylogenetic analysis.

MEGAN5

Then MEGAN 5 (MEtaGenome Analyzer; Huson *et al.*, 2011) was used as it computes and discover the taxonomic content of the dataset, using NCBI taxonomy which summarized and gave taxonomical classification of available pathogens. It also delivered graphical and statistical output for comparing different data sets with output generated from different blast programs. MEGAN 5 uses a simple algorithm that assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hits in the comparison (Huson *et al.*, 2007).

NGS analysis using the developed pipeline

Clone library sequences were aligned and trimmed using Molecular Evolutionary Genetic Analysis (MEGA5; Tamura *et al.*, 2011). Next generation sequencing 16s amplicons were assessed for quality by the developed pipeline by using a quality trimming option . Amplicons with tags that did not have 100% homology to the sample designator and those less than 15 bp long were removed from the analysis by quality trimming. Trimmed clone library and amplicon sequences were analyzed for duplicates using the developed NGS pipeline.

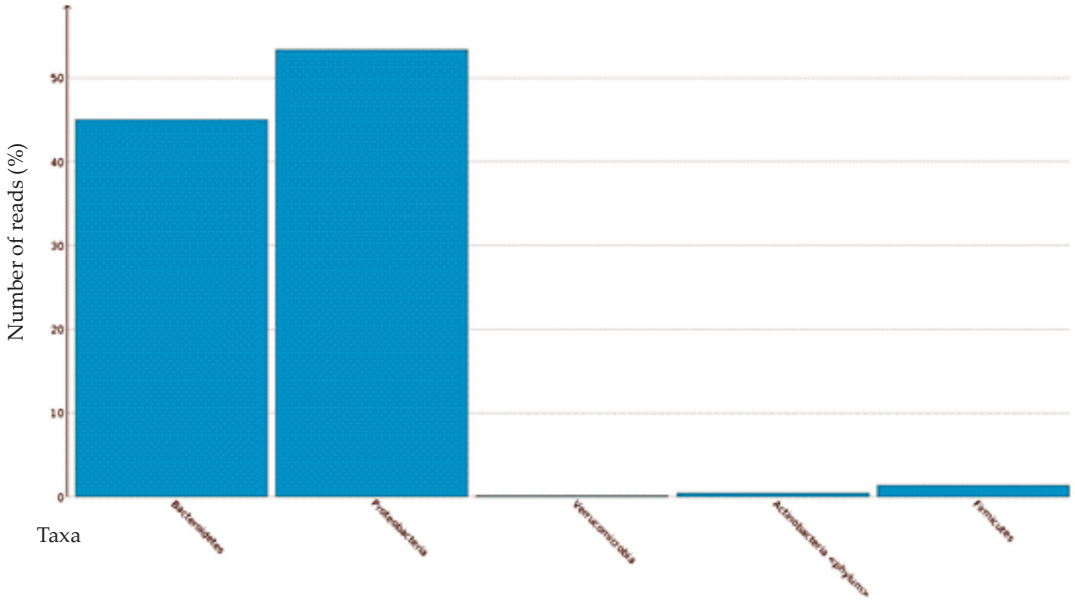

Fig. 4. Overall relative abundance of microbial populations by Phylum

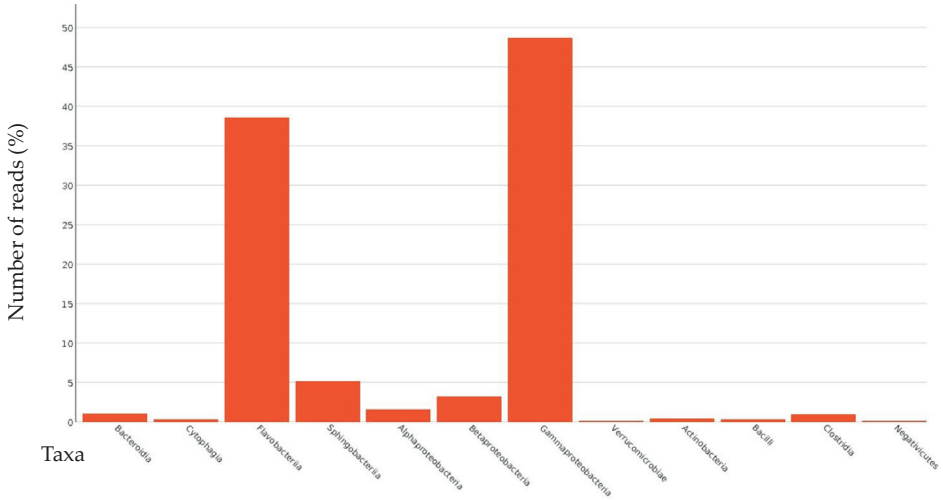

Fig. 6. Overall relative abundance of microbial populations by Class

Sequences for each sample were aligned using SILVA Ribosomal Database version. 1.1, clustered using the complete-linkage clustering method, and operational taxonomic units (OTUs) were determined by multiple pairwise distances using a cutoff of 97% similarity (3% divergence). The OTUs (representing pooled clone library and amplicon sequence data) for each sample were phylogenetically classied. Any OTU sequence that fell below the required identity at any taxonomic level was grouped with other sequences at the next highest level, so that for each sequence the "most certain" taxonomy is reported. Classication of OTUs was visualized and MEGA5 was used to build a maximum likelihood phylogenetic tree of all samples. The sequence analysis results are presented at Phylum (Figure 4), Class (Figure 5), Order (Figure 6), family (Figure 7), genus (Figure 8), species levels (Figure 9) and phylogenetic trees were drawn (Fig.10) to identify the available microorganisms.

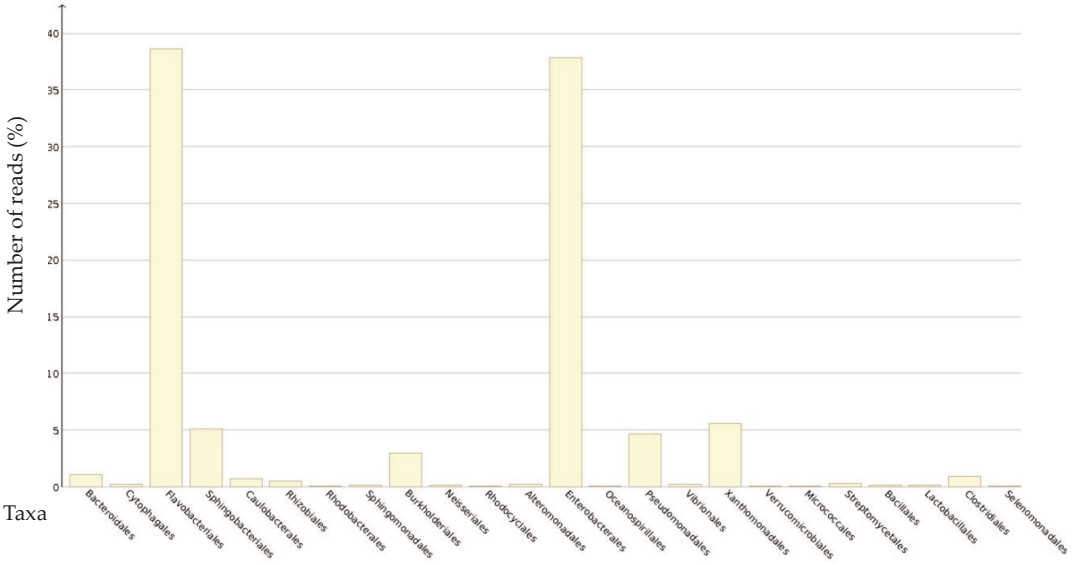

Fig. 7. Overall relative abundance of microbial populations by Order

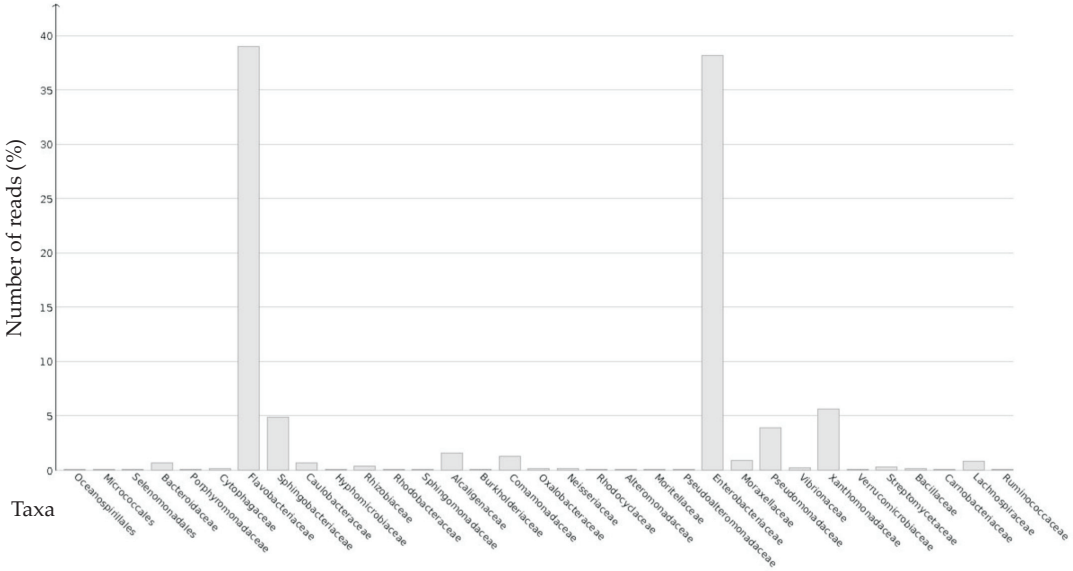

Fig. 8. Overall relative abundance of microbial populations by Family

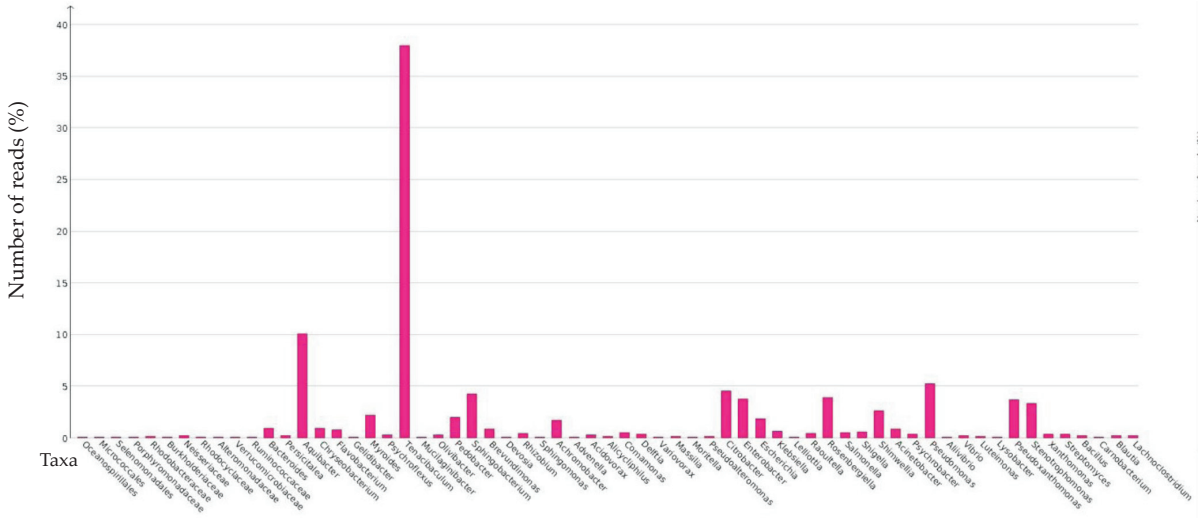

Fig. 9. Overall relative abundance of microbial populations by Genus

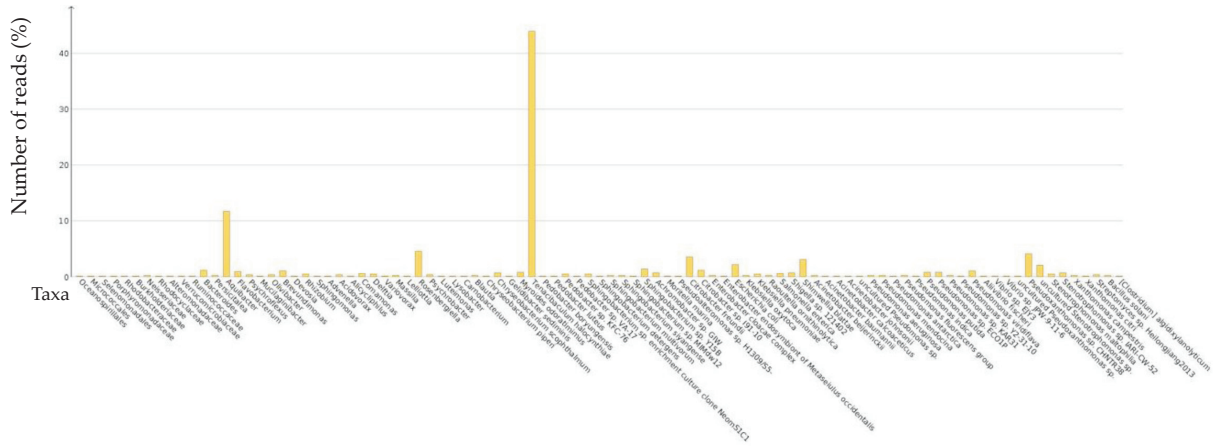

Fig. 10. Overall relative abundance of microbial populations by Species

CONCLUSION

The workflow for analyses of plant pathogens using Next Generation Sequencing data (NGS) of 16S rRNA amplicon sequences was successfully designed and implemented. Further development of the pipeline by piping open-source bioinformatics tools was also successfully done for detecting any kind of pathogen presence in NGS data. Important plant pathogens were identified through this analysis isolated from the imported seed potato samples. Further improvement of the pipeline can be undertaken by designing a graphical interphase and incorporating more plant pathogen related databases into the developed pipeline.

ACKNOWLEDGEMENTS

This study was supported by the National Science Foundation of Sri Lanka (NSF/SCH/2015/01). Authors wish to extend their sincere gratitude to Human Genetics Unit, University of Colombo for their assistance with the development of the bioinformatics pipeline. We gratefully acknowledge Mr. Y.B. Alahapperuma, Prof. Saravanamuttu Gnaneshan and Ms. Dinithi Rajapaksha for their valuable support.

REFERENCES

Allali, I. 2017. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis

Fig. 11. Part of Phylogenetic classification

of the gut microbiome. *BMC Microbiology*. BioMed Central Ltd., 17(1) : 1–16. doi: 10.1186/S12866-017-1101-8/FIGURES/7.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3): 403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger, A. M., Lohse, M. and Usadel, B. 2014. Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data. 30(15) : 2114–2120. doi: 10.1093/bioinformatics/btu170.

Brown, J., Pirrung, M. and Mccue, L. A. (no date) 'Data and text mining FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool'. doi: 10.1093/bioinformatics/btx373.

Cuadros-Orellana, S., Leite, L.R., Smith, A., Medeiros, J.D., Badotti, F., Fonseca, P.L., Vaz, A.B., Oliveira, G. and Góes-Neto, A. 2013. Assessment of Fungal Diversity in the Environment using Metagenomics: A Decade in Review. *Fungal Genomics and Biology.* 3(2).

Dixit, K. 2021. Benchmarking of 16S rRNA gene databases using known strain sequences. *Bioinformation*. Biomedical Informatics Publishing Group, 17(3), p. 377. doi: 10.6026/97320630017377.

Dong, A. Y. 2021. Bioinformatic tools support decision-making in plant disease management. *Trends in Plant Science. Elsevier Current Trends.* 26(9): 953–967. doi: 10.1016/J.TPLANTS.2021.05.001.

Doornbos, R. F., Van Loon, L. C. and Bakker, P. A. H. M. 2011. Impact of root exudates and plant defense signaling on bacterial communities in the rhizosphere. A review. *Agronomy for Sustainable Development 2011 32:1*. Springer, 32(1): 227–243. doi: 10.1007/S13593-011-0028-Y.

Heng, Li and Richard Durbin 2009. Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*. 25(14): 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17(3): 377-86. doi: 10.1101/gr.5969107. PMID: 17255551; PMCID: PMC1800929

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., Schuster, S.C. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21(9) : 1552-60. doi: 10.1101/gr.120618.111. Epub 2011 Jun 20. PMID: 21690186; PMCID: PMC3166839

Graspeuntner, S. 2018. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract', *Scientific Reports 2018 8:1*. Nature Publishing Group, 8(1), pp. 1–7. doi: 10.1038/s41598-018-27757-8.

Lacombe, S. 2010. Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. *Nature Biotechnology.* 28(4): 365–369. doi: 10.1038/nbt.1613.

Langmead, B. and Salzberg, S. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9: 357–359. https://doi.org/10.1038/nmeth.1923

Miller, S. A., Beed, F. D. and Harmon, C. L. 2009. Plant Disease Diagnostic Capabilities and Networks. *https://doi.org/10.1146/annurev-phyto-080508-081743. Annual Reviews*. 47: 15–38. doi: 10.1146/ANNUREV-PHYTO-080508-081743.

Naranpanawa, D. N. U. 2020. Raw transcriptomics data to gene specific SSRs: a validated free bioinformatics workflow for biologists. doi: 10.1038/s41598-020-75270-8.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41(Database issue):D590-6. doi: 10.1093/nar/gks1219. Epub 2012 Nov 28. PMID: 23193283; PMCID: PMC3531112.

Studholme, D. J., Glover, R. H. and Boonham, N. 2011. Application of High-Throughput DNA Sequencing in Phytopathology. *https://doi.org/10.1146/annurev-phyto-072910-095408*. Annual Reviews, 49: 87–105. doi: 10.1146/ANNUREV-PHYTO-072910-095408.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28(10): 2731-9. doi: 10.1093/molbev/msr121. PMID: 21546353; PMCID: PMC3203626.