Proceedings of 15th Agricultural Research Symposium (2016) 111-115

Development of a Bioinformatics Pipeline and Identification of Plant Pathogens using Next Generation Sequencing (NGS) Data

Y.B. ALAHAPPERUMA[1], K. VIVEHANANTHAN[1], N. NETHTHIKUMARA[2], S. GNANESHAN[3], V.H.W. DISSANAYAKE[2] and R.W.P.M. RAJAPAKSHA[1]

[1]*Department of Biotechnology, Faculty of Agriculture and Plantation Management, Wayamba University of Sri Lanka, Makandura, Gonawila (NWP), 60170, Sri Lanka*

[2]*Human Genetics Unit, Faculty of Medicine, University of Colombo, Colombo 08, 00800, Sri Lanka*

[3]*Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada*

ABSTRACT

Plant pathogens affect significant yield losses in each and every agricultural cropping system with a high abundance of plant diseases. Accurate identification of plant pathogens at species or strain level is essential to develop a disease management and surveillance system. With immerging technology, metagenomics using Next-generation sequencing (NGS), an accurate diagnosis of the microbial pathogens can be made in less time than with conventional methods and other molecular methods. In the present study initially, a bioinformatics pipeline was developed and it was used to analyze the previously characterized next generation sequencing data of pathogens of imported seed potato. To analyze NGS data, the following bioinformatics pipeline was developed; 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 MarkDuplicates tool and finally generation of the phylogenetic tree and taxonomic profile with MEGAN 5. In parallel, cashew pathogen DNA was extracted from the infected cashew plant parts and amplified with 16S *rRNA* primers. The output obtained through bioinformatics pipeline identified the presence of *Xanthomonas* spp. and *Agrobacterium* spp. in imported seed potato, which are the important pathogens in plant quarantine and other plant diseases. Further, the extracted DNA of cashew pathogens was good in quality and it was amplified with the 16S *rRNA* primers. Analysis of the NGS data of cashew pathogens through this pipeline which is ongoing will further confirm the effectiveness of this pipeline which could be used as a tool to analyze any type of pathogens from plants in Sri Lanka.

KEYWORDS: 16S rRNA, Bioinformatics pipeline, Metagenomics, Next Generation Sequencing (NGS), Plant pathogens

INTRODUCTION

Microbial plant pathogens have the capability to infect a wide range of plant species causing quantitative and qualitative losses in crops cultivated in different agriculture systems. Overall losses caused by crop diseases have been estimated to range from 9% to 14.2% of potential yield (Orke *et al.*, 1994). Rapid and precise detection of plant pathogens at species or strain level is necessary to develop a disease management and surveillance system. Developing direct detection assays is challenging because pathogens can occur as multiple species complexes or at very low concentration in natural environments (Tsui *et al.*, 2011).

Conventional methods of disease diagnosis involve the study of symptomatology, isolation of the pathogen in suitable culture media and determining by using their physiological, morphological and biochemical characteristics, but these are very time consuming and require extensive taxonomical expertise (Narayanasamy, 2014). These limitations can be reduced by utilization of molecular technologies such as real-time PCR, microarray techniques and immunological assays (enzyme-linked immunosorbent assay-ELISA). Development of monoclonal antibody technology has greatly enhanced the sensitivity and specificity of detection of the immunoassays (ELISA). But prior sequence data and knowledge of the target pathogen is needed for microarray techniques and real-time PCR to leave out uncharacterized pathogens (Monteiro *et al.*, 2015).

Next-generation sequencing technologies (NGS) is a solution for such practical limitations and also has the capability of exploring uncharacterized plant disease systems. Metagenomics studies take the benefit of NGS for a large-scale study of microbial populations by exploring the whole nucleotide sequence content of a sample (Cuadros-Orellana *et al.*, 2013). Recently metagenomics appeared as a novel tool for studying pathogenic microbe–plant interactions (Faure *et al.*, 2011; Knief, 2014) holding great potential to categorize the entire pathogen range in uncharacterized plant disease systems. Metagenomic approach is used to identify plant pathogens from infected plants and any other

pathogen that can present in nucleotide sequence of NGS data (Monteiro *et al.*, 2015).

Analysis of plant pathogen sequences obtained by NGS through a bioinformatics pipeline allows us to detect all pathogens in a single run with the reliable and accurate way in a short time period. This will enable precise identification of the causal pathogens along with the categorization of the main incident diseases. This can also uncover previously unknown/ undetermined pathogens or unculturable species (Monteiro *et al.*, 2015).

In the light of the increasing need to control the emergence and spread of diseases in Sri Lankan large scale plantations, this tool helps to develop improved disease management strategies. The objective of this study is to develop a pipeline to analyze the NGS data of plant pathogens as currently there is no such bioinformatics pipeline used in Sri Lanka. Further, this study focused on isolation of pathogens present in cashew plant parts. Analysis of NGS data of cashew pathogen using the newly developed bioinformatics pipeline will be confirmed in the future.

MATERIALS AND METHODS

Experimental Location

The study was carried out at the Department of Biotechnology, Faculty of Agriculture and Plantation Management, Wayamba University of Sri Lanka with the collaborations of Human Genetics Unit, Faculty of Medicine, University of Colombo and Department of Laboratory Medicine and Pathobiology, University of Toronto, Canada.

Sample Collection

Infected Cashew samples (leaf, inflorescence, roots) were collected from Makandura premises and Kamandaluwa cashew research station. Samples were taken into autoclaved polyethylene bags with the aseptic condition and stored at 4 °C until taken for DNA extraction.

DNA Extraction

Infected Cashew samples were surface sterilized with 70% ethanol and cultured in autoclaved LB media with 1% glucose followed by aeration for overnight at 120 rpm. Extraction of DNA was carried out according to the method mentioned in Rajapaksha *et al.*, 2011.

PCR Amplification of DNA with 16S rRNA Primer

The extracted microbial DNA samples were subjected to PCR amplification with universal primers of 16S *rRNA* gene and subjected to 1.0% agarose gel electrophoresis to check the DNA quality and to confirm the presence of 16S *rRNA* fragment.

Analysis of Next Generation Sequencing Data

Previously sequenced fastq data of the 16S *rRNA* gene region (400 bp) was used (Unpublished data of Rajapaksha *et al.*). The data had been generated on an IonTorrent PGM platform at Credence Genomics. The sequenced DNA was extracted from pathogens of seed potato imported from the USA exporter by the National Plant Quarantine Service, Sri Lanka.

Development of Workflow

Open source bioinformatics tools that were already developed for different bioinformatics task were used to develop the workflow for the pipeline. BASH scripting language was used to pipe as described in Figure 1.

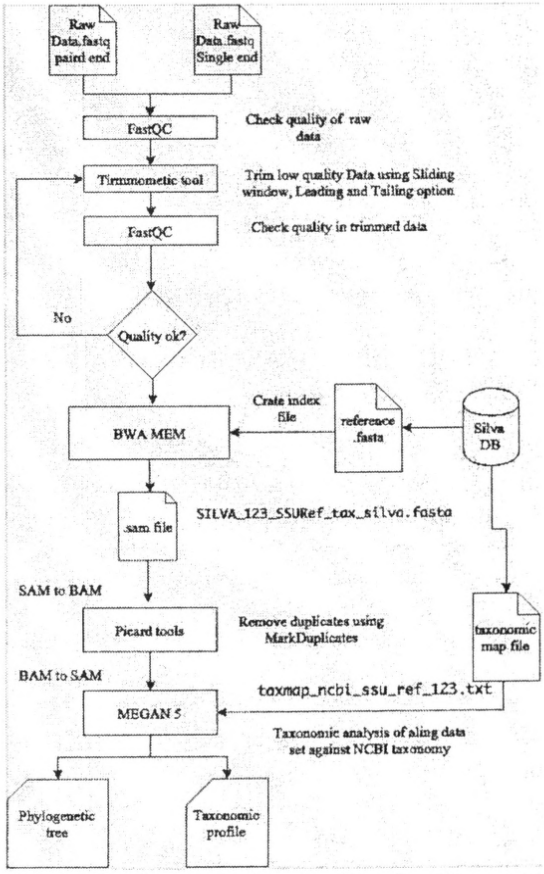

Figure 1. Workflow for pipeline development

Computational Bioinformatics Analysis

A functional laptop with 8 GB RAM was used for running the computational pipeline with Ubuntu 14.04 (Debian based operating system).

According to the workflow (Figure 1), raw fastq files were subjected to the pipeline for analyzing plant pathogens. Trimming of low quality data was carried out based on FastQC report with Trimmometic 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_123.1_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=100.0) and its other default parameters.

RESULTS AND DISCUSSION

DNA Extraction

Initially, the research was focused on isolating microbial DNA from the samples collected from infected tissues of cashew plants (Figure 2). The optimized DNA extraction protocol that was described by Rajapaksha *et al.*, (2011) gave sufficient amounts of DNA (about 10-50 µg/µL).

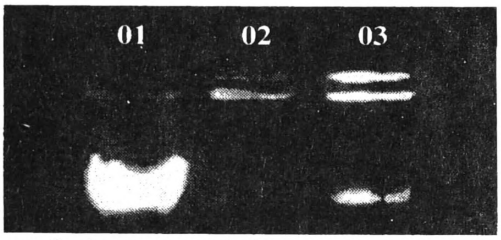

Figure 2. Agarose gel electrophoresis of DNA samples extracted from different infected cashew plant parts. *Lane 1- cashew leaf part, Lane 2- cashew inflorescence part, Lane 3- cashew root zone part*

PCR Amplification of 16S rRNA Gene

PCR amplification with the 16S *rRNA* universal primers confirmed the presence of expected band of 1484 bp of the 16S *rRNA* region. All three samples used for this study were amplified with the 16S *rRNA* primer indicating that extracted DNA was of good quality. The extracted DNA samples will be used for the metagenomics analysis in near future (Figure 3).

Development of Workflow for Pipeline

The computational pipeline for analysis of NGS data described in this study was successful. First, some simple quality control checks had to be 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). 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 sequence, which give idea of PCR duplications.

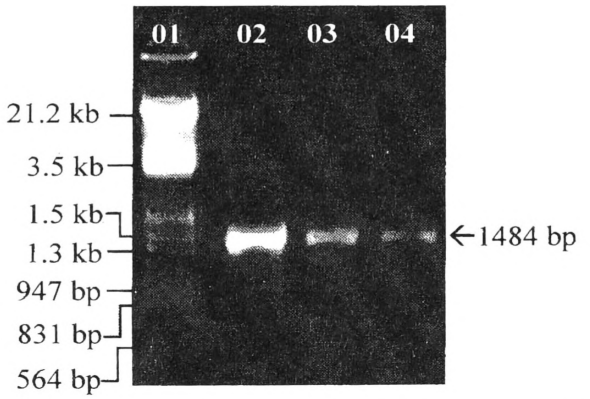

Figure 3. Agarose gel electrophoresis of PCR products using 16S rDNA universal primers. *Lane 1- Lambda Hind III + EcoRI ladder, Lane 2- sample 01, Lane 3- sample 02, Lane 4- sample 03*

The occurrence of poor quality or technical sequences such as adapters and PCR primers in next-generation sequencing (NGS) data can result in drawbacks for analyses. So, trimming and filtering of raw sequences carried out with the Trimmomatic tool (Bolger *et al.*, 2014) that 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

Next, reference-based alignment carried out for trimmed data with the BWA-MEM algorithm (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 compare to others. Moreover it produced significant for

short reads with long reference like 16S analysis. 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 (http://www.picard.sf.net) MarkDuplicates option used to remove PCR duplicate reads of output that was obtained from the alignment algorithm to avoid duplicate reads in phylogenetic analysis.

Then MEGAN 5 (MEtaGenome Analyzer; Huson *et al.,* 2011) was used as it compute 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).

Analyzing Plant Pathogen using NGS Data

From the phylogenetic results of NGS data obtained from seed potato samples, confirmed the presence of *Xanthomonas* spp. and *Agrobacterium* spp. in high abundance (Figure 4). *Xanthomonas* spp. is important bacterial pathogen found in rice and it is also a quarantine important pathogen. *Agrobacterium* spp. causes crown gall disease in plants (Bull *et al.,* 2010) and well known for its ability to transfer DNA between itself and the plant. To confirm the accuracy of the phylogenetic analysis results, it is necessary to trim out plant and virus sequences to screen only for plant pathogens. To enable this, it is necessary to incorporate a new fasta reference file and map file to the pipeline that includes the reference sequences of plant pathogens for alignment and mapping with MAGEN 5.

CONCLUSIONS

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 like *Xanthomonas* spp. and *Agrobacterium* spp. were identified through this analysis isolated from the imported seed potato samples. Further improvement of the pipeline can be undertaken to identify only for plant pathogens by in-cooperating reference database of plant pathogens to be developed the pipeline in future.

ACKNOWLEDGEMENTS

The authors offer their sincere thanks to academic and non-academic staff members of Department of Biotechnology for their support. Authors wish to extend their sincere thanks to Human Genetic Unit, Colombo University for their assistance with the development of the bioinformatics pipeline.

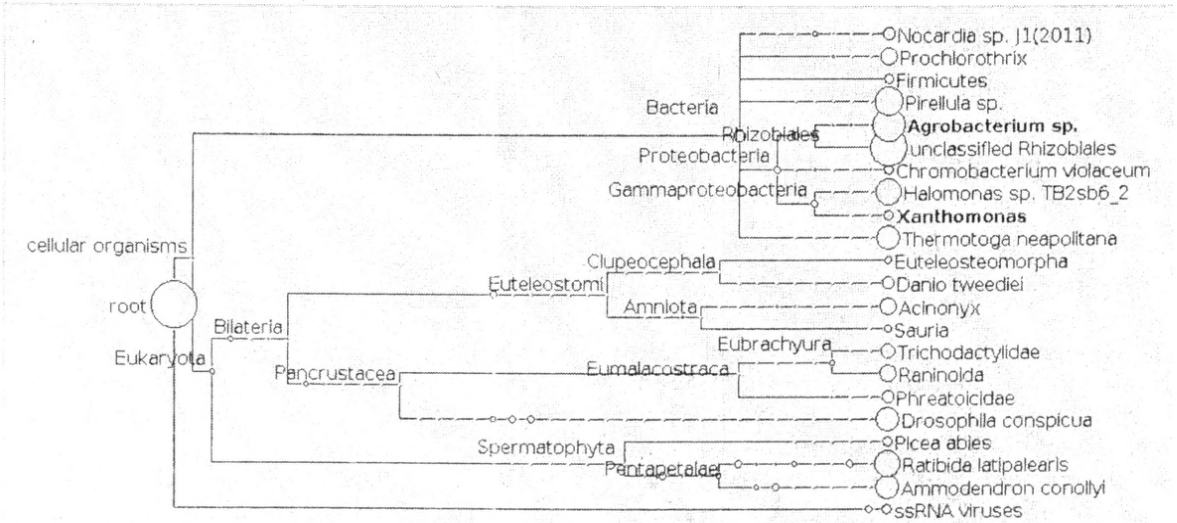

Figure 4. Phylogenetic tree for seed potato sample Wyb3 with given Lowest Common Ancestor (LCA) parameters.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* **215** (3), 403-410.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed 08 June 2016).

Bolger, A.M., Lohse, M. and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30** (15), 2114–2120.

Bull, C.T., Coutinho, T.A., Denny T.P., Firrao G., Fischer-Le Saux, M., Saddler, G.S., Scortichini, M., Stead, D.E. and Takikawa Y. (2010). Comprehensive list of names of plant pathogenic bacteria, 1980-2007. *Journal of Plant Pathology*, **92** (3), 551-592.

Cuadros-Orellana S., Leite L. R., Smith A., Medeiros J. D., Badotti F., Fonseca P. L. C., Vaz, A. B. M., Oliveira, G. and Góes-Neto, A. (2013). Assessment of fungal diversity in the environment using metagenomics: a decade in review. *Fungal Genome Biology*, **3**, 110-123.

Faure D., Tannières M., Mondy S. and Dessaux, Y. (2011). Recent contributions of metagenomics to studies on quorum-sensing and plant-pathogen interactions, In: Metagenomics: Current Innovations and Future Trends, Ed. Marco D., Norfolk, Caister Academic Press, 253–263.

Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17** (3), 377–386.

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N. and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN 4. *Genome Research*, **21** (9), 1552-1560.

Knief, C. (2014). Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in Plant Science*, **5**, 216-239.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9** (4), 357–359.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25** (14), 1754–1760.

Monteiro, F., Romeiras, M. M., Figueiredo, A., Sebastiana, M., Baldé, A., Catarino, L. and Batista, D. (2015). Tracking cashew economically important diseases in the West African region using metagenomics. *Frontiers in Plant Science*, **6**, 482-488.

Narayanasamy, P. (2014) Microbial Plant Pathogens-Detection and Disease Diagnosis: Fungal Pathogens, Vol 1. Netherlands, Springer, 274-278.

Orke E.C., Dehne H.W., Schonbeck F. and Weber A. (1994). Crop Production and Crop Protection: Estimated Losses in Major Food and Cash Crops. Elsevier, Amsterdam.

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 Research*, 41 (Database issue), D590–D596.

Rajapaksha, R.W.P.M., Vivehananthan K. and Perera, A.N.K. (2011). Denaturation Gradient Gel Electrophoresis Analysis of Microbial Populations in Contaminated Environments for Bioremediation. In: proceedings of 11th Agriculture Research Symposium, 12-13 August, 2011. Makandura, Wayamba University of Sri Lanka, 41-45.

Tsui, C. K. M., Woodhall, J., Chen, W., Lévesque, C. A., Lau, A., Schoen, C. D., Baschien, C., Najafzadeh M. J and de Hoog, G. S. (2011). Molecular techniques for pathogen identification and fungus detection in the environment. *IMA Fungus: The Global Mycological Journal*, **2** (2), 177–189.