Plant Archives Volume 20 No. 2, 2020 pp. 7056-7062 **e-ISSN:2581-6063 (online), ISSN:0972-5210**

ANALYSIS OF CHLOROPLAST GENETIC DIVERSITY IN THE ROSACEAE FAMILY USING A NEW PHYLOGENETIC APPROACH

Ali Imad Mohammad Moner

Genetic Engineering and Biotechnology Institute for Post Graduate Study, University of Baghdad, Iraq.

Abstract

Relationships among plants in the Rosaceae plant family have been little studied at the chloroplast genome level. Here we evaluate 174 species belonging to this family according to their full chloroplast sequences in the biggest study of this type so far. Three strategies have been utilized, analysis of full genome sequences for all 174 species, grouping them by main genus with full chloroplast sequence and grouping them by main genus but removing all identical sequences (strip sequences) from the chloroplast sequence. Results show that most software fails to produce a phylogenetic tree except for the Fast Tree and Neighbor Joining methods. While all software successfully produced a phylogenetic tree when the second and third strategies were used.

Key words: chloroplast genetic diversity, Rosaceae family, phylogenetic approach.

Introduction

The Rosaceae is one of the most economically significant plant families in the plant kingdom. It has more than 2.5 thousand species. It is importance for economic fruits like apple, almond, pears, cherries, raspberries and strawberries and timber and medicinal uses, with value around $45 billion according to the FAO 2005 (Hummer & Janick, 2009).

Several researchers have worked on the classification and the relationship among the family member using morphological characterization and molecular marker (Potter *et al.,* 2007). As sequencing technology has become more affordable, many projects have been developed to assemble or reassemble the whole genome of those important species (Zhang *et al.,* 2019). The chloroplast is an organelle with its own genome and it is inherited from the mother plant, therefore it has become a very important tool to study and determine the origin of species and their evolutionary relationships (Moner, Furtado and Henry, 2018).

The position of the Rosaceae among angiosperm families still not well understood but over the last decades, molecular technique have great helped to clarify this (Potter *et al.,* 2007). The study of evolutionary relationships among the genera in this family using the whole chloroplast genome, have been limited, except for a recent study which focused on 15 selected regions from the chloroplast (Sun *et al.,* 2018). This is despite of the availability of hundreds of chloroplast sequences that belong to this family deposited in the NCBI database.

Although, there is abundance of these genomes sequence, finding the tool that can be used to deal with this data to produce a phylogenic tree is challenging. Therefore, there is an urgent need to develop a tool that can process these huge data flies with acceptable resources and affordable timeline. There are several modern tools which have developed to do so, but still can not meet the research requirements and new procedures are required, especially with increases in the output of NGS. The aim of this research was to evaluate the genetic diversity of the chloroplasts genomes of Rosaaceae plant family as a model to find the impact of strip the identical sequences on the tree topology using different software.

Material and methods

Available annotated full chloroplast sequences which belong to the Rosaceae plant family were downloaded from the NCBI database in GenBank format table 1. All sequences obtained were aligned using MAFFT tools with the following settings; (Auto, 1PAM/K = 2 scoring matrix, 1.53 open gap penalty and 0.123 offset value) (Katoh & Standley, 2013). All chloroplast gene sequences were extracted from these sequences and concatenated

Table 1: The scientific names, accession numbers of the chloroplast used in this study.

Organism Name	accession number	Organism Name	accession number	Organism Name	accession number	Organism Name	accession number	Organism Name	accession number
Alchemilla	NC_	Rosa	NC_	Cotoneaster	NC_	Malus	NC_	Prunus	NC_
argyrophylla	049037.1	rugosa	044094.1	wilsonii	046834.1	trilobata	035671.1	persica	014697.1
Alchemilla	NC_	Rubus	NC_	Crataegus	NC_	Malus	NC_	Prunus	NC_
pedata	049038.1	amabilis	047211.1	kansuensis	039374.1	tschonoskii	035672.1	pseudocerasus	030599.1
Bencomia	NC_	Rubus	NC_	Cydonia	NC_	Malus x	NC_	Prunus	NC_
exstipulata	039924.1	boninensis	046015.1	oblonga	045415.1	atrosanguinea	045409.1	rufa	048528.1
Fragaria	NC_	Rubus	NC_	Dasiphora	NC_	Malus	NC_	Prunus	NC_
chiloensis	019601.1	coreanus	042715.1	fruticosa	036423.1	yunnanensis	039624.1	salicina	047442.1
Fragaria	NC_	Rubus	NC_	Dichotomanthes	NC_	Osteomeles	NC_	Prunus	NC_
iinumae	024258.1	crataegifolius	039704.1	tristaniicarpa	045335.1	schwerinae	045420.1	serotina	036133.1
Fragaria	NC_	Rubus	NC_	Docynia	NC_	Phippsiomeles	NC_	Prunus	NC_
mandshurica	018767.1	hybrid cultivar	042716.1	delavayi	045424.1	matudae	045421.1	speciosa	043921.1
Fragaria	NC_	Rubus	NC_	Eriobotrya	NC_	Phippsiomeles	NC_	Prunus	NC_
nipponica	035500.1	takesimensis	037991.1	henryi	045345.1	mexicana	045422.1	takesimensis	039379.1
Fragaria	NC_	Rubus	NC_	Eriobotrya	NC_	Photinia	NC_	Prunus	NC_
orientalis	035501.1	trifidus	046585.1	japonica	034639.1	beckii	045353.1	tenella	044965.1
Fragaria	NC_	Sanguisorba	NC_	Eriobotrya	NC_	Photinia	NC_	Prunus	NC_
pentaphylla	034347.1	filiformis	044693.1	laoshanica	049114.1	blinii	045412.1	tomentosa	036394.1
Fragaria vesca	NC_	Sanguisorba	NC_	Eriobotrya	NC_	Photinia	NC_	Prunus	NC_
subsp. vesca	015206.1	officinalis	044694.1	malipoensis	045346.1	integrifolia	045344.1	triloba	046742.1
Fragaria	NC_	Sanguisorba	NC_	Eriobotrya	NC_	Photinia	NC_	Prunus	NC_
virginiana	019602.1	sitchensis	044691.1	obovata	045347.1	lanuginosa	045354.1	yedoensis	026980.1
Fragaria	NC_	Sanguisorba	NC_	Eriobotrya	NC_	Photinia	NC_	Prunus	NC_
viridis	048474.1	tenuifolia	042223.1	salwinensis	045348.1	lochengensis	045352.1	zippeliana	043926.1
Fragaria x	NC_	Sanguisorba tenui-	NC_	Eriobotrya	NC_	Photinia	NC_	Pyracantha	NC_
ananassa	035961.1	folia var. alba	044692.1	seguinii	045349.1	prionophylla	045355.1	fortuneana	042192.1
Geum	NC_	Amelanchier	NC_	Fragaria vesca	NC_	Photinia	NC_	Pyrus	NC_
rupestre	037392.1	alnifolia	045314.1	subsp. bracteata	018766.1	serratifolia	045331.1	communis	045336.1
Gillenia	NC_	Amelanchier	NC_	Gillenia	NC_	Photinia	NC_	Pyrus	NC_
trifoliata	045311.1	arborea	045313.1	stipulata	045321.1	sorbifolia	045416.1	pashia	034909.1
Pentactina	NC_	Amelanchier	NC_	Hesperomeles	NC_	Photinia	NC_	Pyrus	NC_
rupicola	016921.1	asiatica	045312.1	cuneata	045326.1	villosa	045411.1	pyrifolia	015996.1
Potaninia	NC_	Amelanchier	NC_	Hesperomeles	NC_	Pourthiaea	NC_	Pyrus	NC_
mongolica	046739.1	bartramiana	045315.1	ferruginea	045328.1	amphidoxa	045414.1	spinosa	023130.1

Table 1 Continued...

Table 1 Continued...

Potentilla centigrana	NC_ 041209.1	Amelanchier cusickii	NC_ 045316.1	Hesperomeles goudotiana	NC_ 045327.1	Pourthiaea arguta	NC_ 045413.1	Pyrus ussuriensis	NC_ 041461.1
Potentilla freyniana	NC_ 041210.1	Amelanchier humilis	NC_ 045317.1	Hesperomeles pernettyoides	NC_ 045329.1	Pourthiaea tomentosa	NC_ 045417.1	Rhaphiolepis ferruginea	NC_ 045332.1
Potentilla hebiichigo	NC_ 041199.1	Amelanchier interior	NC_ 045318.1	Kageneckia angustifolia	NC_ 045322.1	Prunus armeniaca	NC_ 043901.1	Rhaphiolepis impressivena	NC_ 045350.1
Potentilla indica	NC_ 041178.1	Amelanchier pallida	NC_ 045319.1	Kageneckia lanceolata	NC_ 045323.1	Prunus avium	MK 622380.1	Rhaphiolepis indica	NC_ 045330.1
Potentilla stolonifera	NC_ 044418.1	Amelanchier sanguinea	NC_ 045320.1	Kageneckia oblonga	NC_ 045324.1	Prunus camp-anulata PCS11	NC_ 044123.1	Rhaphiolepis lanceolata	NC_ 045333.1
Prinsepia utilis	NC_ 021455.1	Aronia arbutifolia	NC_ 045391.1	Malacomeles denticulata	NC_ 045325.1	Prunus cerasoides	NC_ 035891.1	Rhaphiolepis major	NC_ 045351.1
Rosa banksiae	NC_ 042194.1	Chaenomeles cathayensis	NC_ 045392.1	Malus angustifolia	NC_ 045410.1	Prunus davidiana	NC_ 039735.1	Rhaphiolepis salicifolia	NC_ 045342.1
Rosa berberifolia	NC_ 045126.1	Chaenomeles japonica	NC_ 035566.1	Malus baccata	NC_ 045389.1	Prunus dulcis	NC_ 034696.1	Rhaphiolepis umbellata	NC_ 045334.1
Rosa canina	NC_ 047295.1	Chaenomeles sinensis	NC_ 045337.1	Malus coronaria	NC_ 045308.1	Prunus humilis	NC_ 035880.1	Rosa hybrid cultivar	NC_ 044126.1
Rosa chinensis	CM 009590.1	Cotoneaster acuminatus	NC_ 045340.1	Malus doumeri	NC_ 045343.1	Prunus kansuensis	NC_ 023956.1	Sorbus aria	NC_ 045418.1
Rosa chinensis var. spontanea	NC_ 038102.1	Cotoneaster buxifolius	NC_ 045356.1	Malus florentina	NC_ 035625.1	Prunus leveilleana	NC_ 049028.1	Sorbus chamaemespilus	NC_ 045419.1
Rosa laevigata	NC_ 046824.1	Cotoneaster frigidus	NC_ 045341.1	Malus hupehensis	NC_ 040170.1	Prunus matuurae	NC_ 045230.1	Sorbus setschwanensis	NC_ 046777.1
Rosa laevigata var. leiocarpa	NC_ 047418.1	Cotoneaster horizontalis	NC_ 045357.1	Malus ioensis	NC_ 045393.1	Prunus maximowiczii	NC_ 026981.1	Sorbus torminalis	NC_ 033975.1
Rosa lucieae	NC_ 040997.1	Cotoneaster microphyllus	NC_ 045339.1	Malus micromalus	NC_ 036368.1	Prunus mira	NC_ 040125.1	Sorbus ulleungensis	NC_ 037022.1
Rosa maxi-mowicziana	NC_ 040960.1	Cotoneaster rubens	NC_ 045359.1	Malus prattii	NC_ 043902.1	Prunus mongolica	NC_ 037849.1	Torminalis clusii	NC_ 045423.1
Rosa multiflora	NC_ 039989.1	Cotoneaster schantungensis	NC_ 045840.1	Malus prunifolia	NC_ 031163.1	Prunus mume	NC_ 023798.1	Vauquelinia australis	NC_ 045309.1
Rosa praelucens	NC_ 037492.1	Cotoneaster silvestrii	NC_ 045358.1	Malus sieversii	NC_ 045390.1	Prunus padus	NC_ 026982.1	Vauquelinia pauciflora	NC_ 045310.1
Rosa roxburghii	NC_ 032038.1	Cotoneaster taylorii	NC_ 045338.1	Malus toringoides	NC_ 049113.1	Prunus pedunculata	NC_ 037850.1		

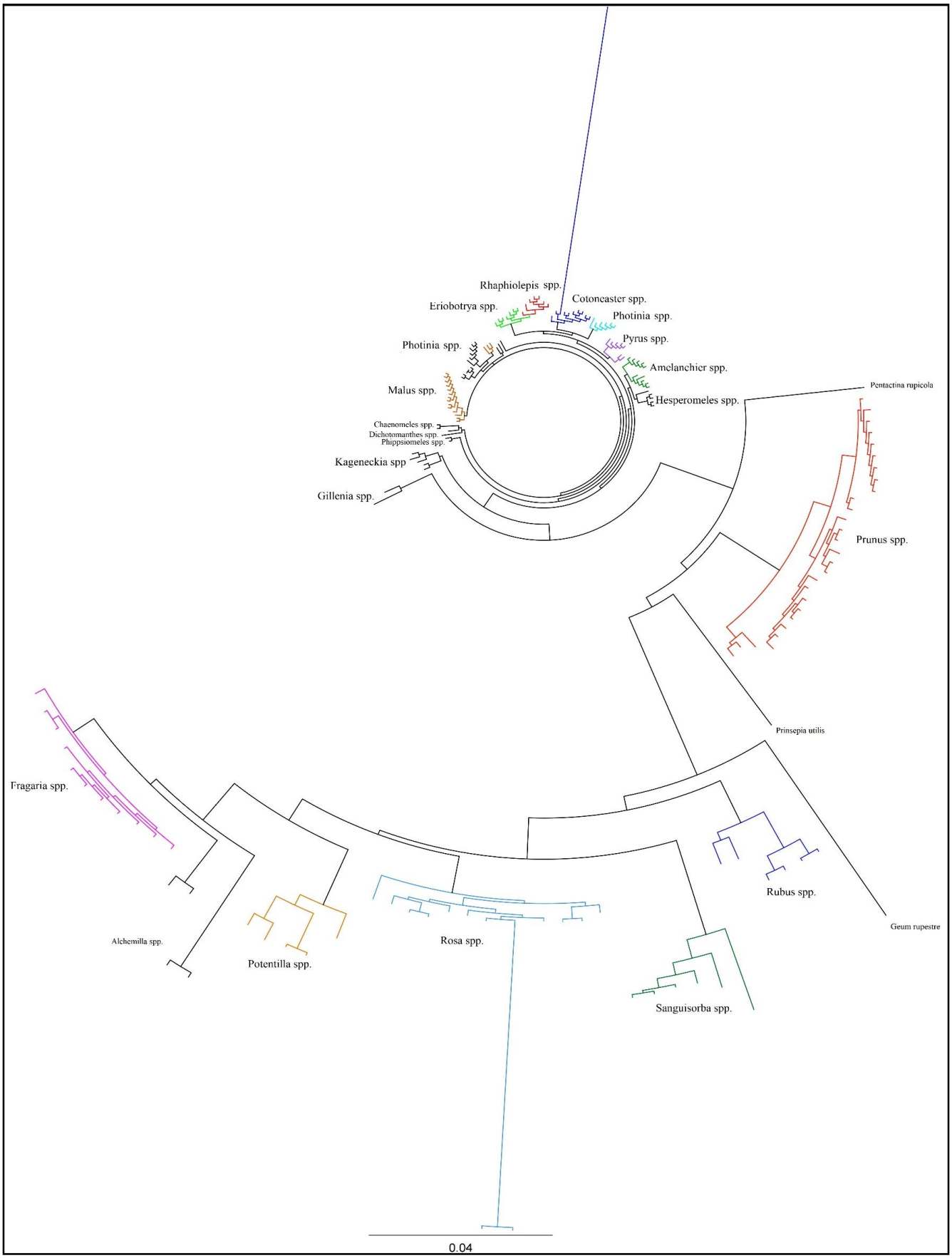

Fig. 1: The phylogenetic tree of the Rosaceae family based on full chloroplast sequence using Fast Tree software.

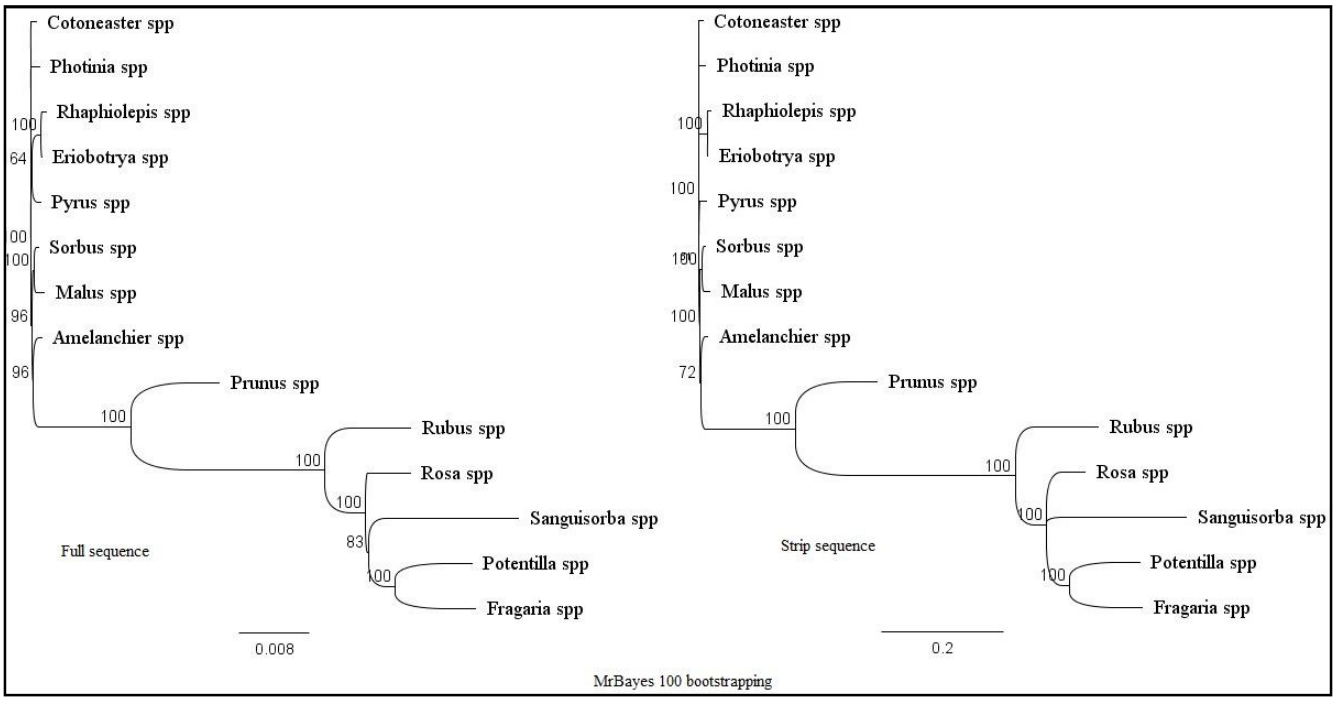

Fig. 2: Phylogenetic of the main 14 genus in Rosaceae family using MrBayes software and 100 bootstrapping with both full and strip sequences.

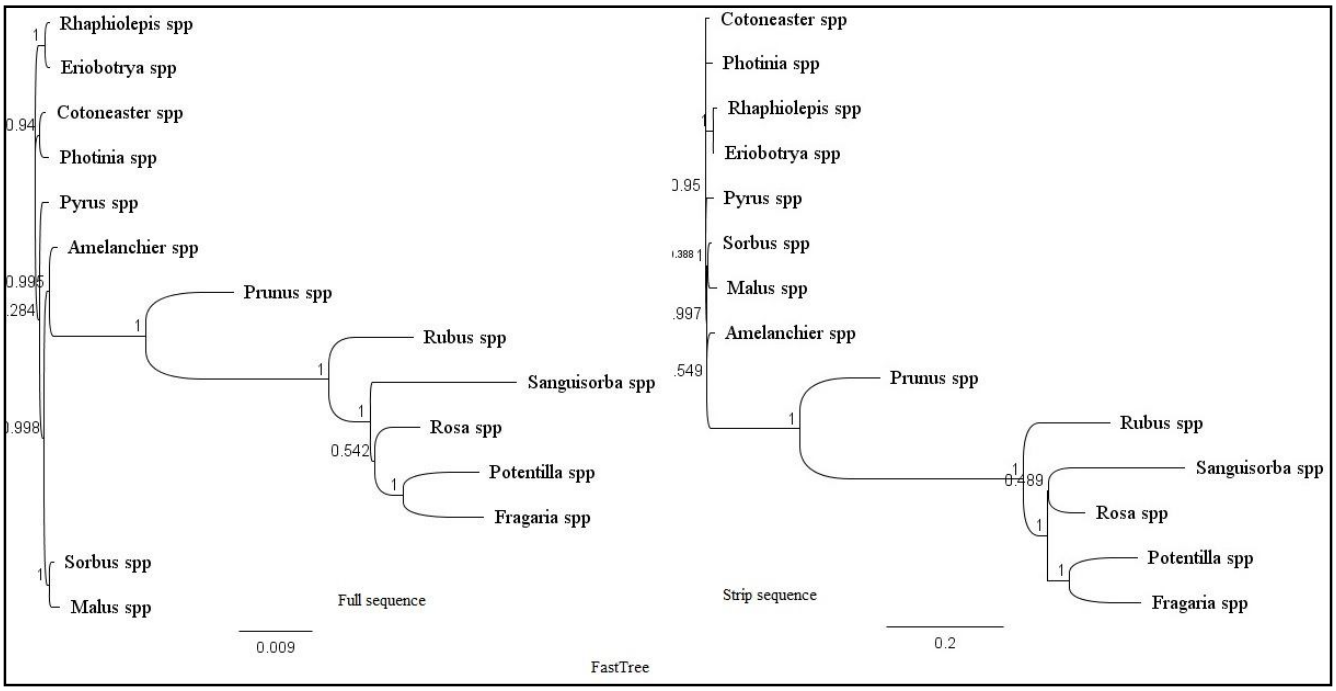

Fig. 3: Phylogenetic of the main 14 genus in Rosaceae family using FastTree software with both full and strip sequences.

according to their accession name and number and aligned to represent just the genes sequences (Brozynska *et al.,* 2017; Moner, Furtado, Chivers, *et al.,* 2018).

In order to preform phylogenetic analysis six different software packages were used, namely (Moner, Furtado, & Henry, 2018)B : PAUP, Maximum Parsimony, Heuristic search and 100 bootstrapping (Swofford, 2003), FastTree, GTR, optimize the Gamma 20 likelihood (Price *et al.,* 2009), PhyMl, GTR and 100 bootstrapping (Guindon *et al.,* 2010), RaxMl, maximum likelihood, GTR, Gamma and 100 bootstrapping (Stamatakis, 2014), MrBayes, Bayesian, GTR, Gamma and 100 bootstrapping (Huelsenbeck & Ronquist, 2001) Neighbor Joining, Tamura-Nei, 100 bootstrapping (Saitou & Nei, 1987).

The full sequence alignment file was utilized to preform the phylogenic analysis, then to simplify the

analysis, the species were grouped by genus then, the aligned consensus sequences were used to represent the genus. Thereafter two strategies were utilized, firstly construction of the phylogeny by full aligned sequence of the grouped species according to main genus, and secondly using the same sequences but after removing all identical bases to minimize the input data. Aligned sequences were striped of all 100% identical bases (uninformative sequence) using Geneious (Kearse *et al.,* 2012) to reduce the amount of computer resources that were need to analyze the data. Each phylogeny package was used twice one with full sequence length and the other with striped consensus, with the same settings as described previously.

A local workstation with two CPU 2.9GHz, 16 core, 40 MB cash5100 and 96 Gb of RAM was used to preform all the above analysis.

Results and Discussion

Alignment of the chloroplast sequences of 174 species belonging to 37 genera from the Rosaceae plant family was achieved successfully using the MAFFT aligner tool. Chloroplast sequences range from 128.788 to 160.937 Kbp. The large variation among those species generated gaps which extended the alignment consensus to 283.205 Kbp. This file was utilized in the phylogenetic analysis. The size of this file was bigger than the capability of most of the phylogeny software which led to failure and did not produce a tree. Only FastTree Fig. 1 and Neighbor Joining were finished successfully (Moner, Furtado & Henry, 2018 B).

Species were grouped by main genus and aligned. The consensus sequences for the main genus was used to align them to produce a reduced file size effectively to 154.094 Kbp which shorten the consensus about 129 Kbp. In addition to decreasing the number of accessions to that of the 14 main genus groups. This size could be handled with some of the software like FastTree RAxML and PhyMl and trees could constructed. (Guindon *et al.,* 2010; Price *et al.,* 2009; Stamatakis, 2014).

Concatenated gene (Brozynska *et al.,* 2017) sequences of all species were generated and grouped by genus and aligned to produce a 154.100 Kbp length sequence for 14 genera. This effectively allowed the analysis to finish and produce trees for all packages. This significantly reduced the computational resources that are needed to process these kinds of data. In addition, for greater reduction of data and reduced analysis time, identical bases (which are not informative) were removed and gave an even shorter alignment which could be processed faster and with less resources. All software packages completed analysis and trees were generated. The topology of these trees varied from identical in Bayesian analysis to slightly different in one branch as in PAUP MP and PhyMl and two branches as in RAxMl, but many differences in FastTree. Fig. 2 and 3. These results can help researchers working with huge data sets to achieve their results easier and faster with high reliability especially when computational resources are limited.

References

Brozynska, M., D. Copetti, A. Furtado, R.A. Wing, D. Crayn, G. Fox, R. Ishikawa and R.J. Henry (2017). Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice. *Plant Biotechnology Journal*, **15(6):** 765-774. https://doi.org/10.1111/pbi.12674.

Guindon, S., J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk and O. Gascuel (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, **59(3):** 307-321. https://doi.org/10.1093/sysbio/syq010.

Huelsenbeck, J.P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17(8):** 754-755. https://doi.org/10.1093/bioinformatics/17.8.754.

Hummer, K.E. and J. Janick (2009). Genetics and Genomics of Rosaceae. *Genetics and Genomics of Rosaceae*. https://doi.org/10.1007/978-0-387-77491-6.

Katoh, K. and D.M. Standley (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30(4):** 772-780. https://doi.org/10.1093/molbev/mst010.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz and C. Duran (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28(12):** 1647-1649.

Moner, A.M., A. Furtado, I. Chivers, G. Fox, D. Crayn and R.J. Henry (2018). Diversity and evolution of rice progenitors in Australia. *Ecology and Evolution*, **8(8):** https://doi.org/10.1002/ece3.3989.

Moner, A.M., A. Furtado and R.J. Henry (2018). Chloroplast phylogeography of AA genome rice species. *Molecular Phylogenetics and Evolution*, **127:** https://doi.org/10.1016/j.ympev.2018.05.002.

Potter, D., T. Eriksson, R.C. Evans, S. Oh, J.E.E. Smedmark, D.R. Morgan, M. Kerr, K.R. Robertson, M. Arsenault, T.A. Dickinson and C.S. Campbell (2007). Phylogeny and classification of Rosaceae. In *Plant Systematics and Evolution,* **266(1-2):** https://doi.org/10.1007/s00606-007-0539-9.

Price, M.N., P.S. Dehal and A.P. Arkin (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26(7):** 1641-1650. https://doi.org/10.1093/molbev/msp077.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, ***4*(4):** 406-425.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30(9):** 1312-1313. https://doi.org/10.1093/bioinformatics/btu033.

Sun, J., S. Shi, J. Li, J. Yu, L. Wang, X. Yang, L. Guo and S. Zhou (2018). Phylogeny of Maleae (Rosaceae) based on multiple chloroplast regions: Implications to genera circumscription. *Bio. Med. Research International*, **2018:** 6-9. https://doi.org/10.1155/2018/7627191.

Swofford, D.L. (2003). PAUP*: Phylogenetic Analysis Using Parsimony. Sunderland, MA. *Sinauer Associates, Version*, **4:** b10.

Zhang, L., J. Hu, X. Han, J. Li, Y. Gao, C.M. Richards, C. Zhang, Y. Tian, G. Liu, H. Gul, D. Wang, Y. Tian, C. Yang, M. Meng, G. Yuan, G. Kang, Y. Wu, K. Wang, H. Zhang and P. Cong (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications*, **10(1):** 1-13. https://doi.org/10.1038/s41467-019-09518-x.