

DAAD



# BIOINFORMATICS WORKSHOP

PROGRAM BOOK

BIOINFORMATICS  
IN LIFE SCIENCE

MEDICAL  
AGRICULTURE  
ENVIRONMENT  
ENERGY  
FOOD

BIOINFORMATICS WORKSHOP  
Tahun 2015-2016

PROGRAM BOOK

DAAD



# DNA Barcoding Analysis of *matK* Gene of Some *Syzygium* Species

Trina E. Tallei\*, Presticilla D. Irawan, Beivy J. Kolondam  
Department of Biology, Faculty of Mathematics and Natural Sciences,  
Sam Ratulangi University  
Corresponding author: trina\_tallei@unsrat.ac.id

## Abstract

*Syzygium* is a genus of flowering plants that belongs to the myrtle family (Myrtaceae). Accurate identification of *Syzygium* species using traditional methods can take a long time due to lack of knowledge about plants and/or lack of flowers and fruit characters required for identification. Reliability or accuracy of identification relies on expert determination, which is not always readily available. Moreover, *Syzygium* genera exhibit many overlapping characters, making them difficult to identify. In this research, the utility of *matK* gene as DNA barcoding in *Syzygium* was assessed. The assessment included barcoding gap, intraspecific and interspecific divergence, as well as phylogenetic tree reconstruction. The evolutionary history was inferred by using UPGMA method. Estimation of evolutionary divergence between sequence was analysed using Kimura's two-parameter. Evolutionary analyses were conducted in MEGA7.0.18 The result showed an overlapping distribution of interspecific divergence and intraspecific distance for *matK* region. The phylogeny indicated that some of *Syzygium cumini* are embeded within other *Syzygium* species. Therefore, *matK* gene is not suitable to be used as barcode region for identification of *Syzygium* species.

Keywords: *Syzygium*, *Syzygium cumini*, DNA barcode, DNA barcoding gap, intraspecific, interspecific

*Trina Talley*

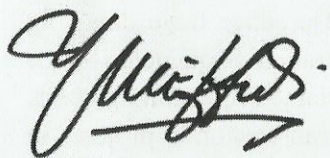
## FOREWORD

Biological data comes in many forms and often in large volumes. Analytical, statistical, mathematical, and computational tools had contribute significantly to helping in the process of decoding the secrets of biology. For understanding those data, bioinformatics has become an increasingly important part of how biological problem in life science researches have solved. Growth of bioinformatics in the last decade has accelerated due to path breaking advancement in biology and new technologies that produce huge data (big data) like high throughput sequencing (Next Generation Sequencing-NGS) which has been recently used for performing full genome sequencing projects including the human genome projects, large-scale transcriptomic projects in human, plants and animals. The data mining and analysis of such large data and extract of knowledge from this data is being made possible only with the help of new software tools and computational intensive techniques. Thus, it is necessary for researcher to learn and use all new technological developments which are taking place in bioinformatics to solve complex biological problems leading to advancement i.e. in medical science, agricultural improvement systems, drugs discovery, environmental and ecological problems, etc. Moreover, bioinformatics offers an alternative approach for supporting inadequate wet lab researches to provide better data of those researches.

This bioinformatics workshop is designed as a scientific meeting for providing a forum for researchers who has concern in development of life science research using bioinformatics. From these three days intensive workshop we hope that transfer of bioinformatics knowledges particularly in understanding of RNA bioinformatics, NGS analysis, protein annotation, protein and molecules docking will help all participants in this bioinformatics workshop to have a new perspective in approaching life science researches.

On behalf of the Organizing Committee, it is our pleasure to welcome you to Bioinformatics Workshop which supported by DAAD (Deutscher Akademischer Austausch Dienst) and IGN (Indonesian- German Network). We hope you enjoy and gain new knowledges and new perspective in Bioinformatics.

#### Workshop Chairman



Dr. rer. nat. Yunus Effendi, M.Sc

## TABLE OF CONTENTS

FOREWORD.....	1
TABLE OF CONTENTS .....	3
COMMITTEE.....	5
SCHEDULE.....	6
ABSTRACT PLENARY PRESENTATION (PPS) AND POSTER PRESENTATION (POP).....	8
Monster in RNA Land: The Evolution of "Strange" Transcripts, Prof. Dr. Peter F. Stadler (PPS-01).....	9
Using RNAseq 'trash bin' data to explore archaeal splicing processes, Sarah Berkemer, M.Sc. (PPS-02).....	10
Evolutionary Analysis of the Protein Domain Distribution, Dr. rer. nat. Arli Aditya Parikesit (PPS-03).....	11
Bioinformatics: Cheap and Robust Method to Explore Biomaterial from Indonesia Biodiversity, Dr. Widodo, MSc (PPS-04).....	13
DNA Sequence Database on Computer Cluster using Hadoop, Dr. Ir. Ade Jamal (PPS-05).....	14
DNA Barcoding Analysis of <i>matK</i> Gene of Some <i>Syzygium</i> Species, Trina E. Tallei, Presticilla D. Irawan, Beivy J. Kolondam (POP-01).....	15
Biosistematics of <i>Annona muricata</i> , <i>Annona squamosa</i> and <i>Annona reticulata</i> with a numerical approach, Hamidah, Santosa, Soegihardjo, Rina S.K (POP-02) .....	16
Diversity study of Diatom from Indonesian Tropical Ocean, Yanti Rachmayanti, Sutomo, Elva Stiawan, Intan Safarina, Rika Felicita, Zeily Nurachman (POP-03) .....	17
Analysis of 18S rRNA Universal Primer for Molecular Identification of Vera Durian, Waheni Rizki Aprilia, Hermin Pancasakti Kusumaningrum, Anto Budiharjo (POP-04) .....	18
On the Convolution Models for Background Correction of Bead Arrays, Rohmatul Fajriyah (POP-05) .....	19
DNA Barcoding of <i>Cytochrome Oxidase 1 (CO1)</i> in Pleco (Loricariidae, <i>Pterygoplichthys</i> sp.) from Ciliwung river of South Jakarta region, Rosnaeni, Dewi Elfidasari, Melta Rini Fahmi (POP-06).....	20
Identification of suckermouth armored catfish (Loricariidae) based on morphometric and meristic characteristic from Ciliwung river watershed Jakarta region, Fatihah Dinul Qoyyimah, Dewi Elfidasari, Melta Rini Fahmi (POP-07).....	21

Shed light on Plant-Pathogen Interaction:A Study Case of long non coding RNAs (lncRNAs) during <i>Fusarium oxysporum</i> f. sp. <i>cubense1</i> ( <i>Foc1</i> ) infection in Banana, Husna Nugrahapraja, Fenny M. Dwivany (POP-08) .....	22
DAAD ALUMNI PARTICIPANT .....	23
PARTICIPANTS.....	27

## COMMITTEE

### Steering Committee

- Dr. Ir. Ahmad Lubis, M.Sc.  
(Rector of Al Azhar Indonesia University)
- Dr. Ade Jamal  
(Vice Rector II of Al Azhar Indonesia University)
- Dr. ArySyahriar, DIC  
(Dean of Science and Technology Faculty, Al Azhar Indonesia University)
- Dr. Nita Noriko  
(Head of Biology Study Program, Al Azhar Indonesia University)

### Organizing Committee

- Chairman:  
Dr.rer.natYunus Effendi, MSc (DAAD Alumni-IGN)
- Secretary:  
Dr.agr. TaufiqWisnuPriambodo (DAAD Alumni-IGN)
- Treasurer:  
RiriSafitri, S.Si, MTI and RirisL Puspitasari, S.Si, M.Si
- Division of Seminar:  
Ir. WinangsariPradani, MT andDenny Hermawan, ST, M.Kom.
- Division of Accomodation :  
Analekta Tiara P, M.Si and AriefPambudi, S.Si, M.Si

### Observer

Prof. Dr. rer.nat. Wolfgang Nellen – (IGN), Universitaet Kassel Germany

## PPS-05

**DNA Sequence Database on Computer Cluster using Hadoop****Dr. Ir. Ade Jamal**Informatics Engineering Study Program, Faculty of Science and Technology  
Al Azhar Indonesia UniversityEmail: [adja@uai.ac.id](mailto:adja@uai.ac.id)

In 2011, Center for Bioinformatics of University of Al Azhar Indonesia is established to enhance research collaborations between Department of Informatics and Biology. In the same year the center took part in consortium for research in vaccine of hepatitis B lead by state owned medicine company PT Bio Farma. The joint research which accompanied by a number of researchers from prominent research institution in bioscience and biomolecular in Indonesia is aimed to develop the second generation hepatitis vaccine. Lesson learned from the joint research is that Indonesia has no central database for molecular biology which collects and disseminates biological information from bioscience researcher in the country. This fact has been already National interest because Indonesia as a big country and big nation potentially has huge data of molecular biology since we have one of the largest biodiversity natural laboratories in our tropical forest and under the sea, in the biggest palm plantation, the variety of tropical diseases and still more to name it. Nevertheless, after so many years Indonesian biomolecular researchers have no nationwide molecular biology database for collecting and disseminating their work before they publish their result in international database such the well known as Genbank. There are some reasons for this problem, i.e. technical reason and non-technical one. In this article we will discuss one of the technical reasons, namely excessive computational time required to process the very large data, in this case DNA sequences on an affordable computer system. An effort to speed up the computational time has been worked out by uploading the DNA sequence data on Hadoop Distributed File System on low-end cluster. A so called MapReduce computation model is invoked for keyword searching algorithm in conjunction with Hadoop Distributed File System as both technologies are main component of Hadoop framework. Recently, we extended searching algorithm by considering global sequence alignment. The result has shown that using Hadoop Framework technology which is inspired by Google search engine technology it is very possible to handle big data of DNA and also protein sequences using affordable computer system.

Keyword: DNA Sequence Database, Global Sequence Alignment, Big Data

## POP-01

**DNA Barcoding Analysis of *matK* Gene of Some *Syzygium* Species**Trina E. Tallei<sup>1,2</sup>, Presticilla D. Irawan<sup>1</sup>, Beivy J. Kolondam<sup>1</sup><sup>1</sup>Department of Biology, Faculty of Mathematics and Natural Sciences,  
Sam Ratulangi University<sup>2</sup>Corresponding author: [trina\\_tallei@unsrat.ac.id](mailto:trina_tallei@unsrat.ac.id)

*Syzygium* is a genus of flowering plants that belongs to the myrtle family (Myrtaceae). Accurate identification of *Syzygium* species using traditional methods can take a long time due to lack of knowledge about plants and/or lack of flowers and fruit characters required for identification. Reliability or accuracy of identification relies on expert determination, which is not always readily available. Moreover, *Syzygium* genera exhibit many overlapping characters, making them difficult to identify. In this research, the utility of *matK* gene as DNA barcoding in *Syzygium* was assessed. The assessment included barcoding gap, intraspecific and interspecific divergence, as well as phylogenetic tree reconstruction. The evolutionary history was inferred by using UPGMA method. Estimation of evolutionary divergence between sequence was analysed using Kimura's two-parameter. Evolutionary analyses were conducted in MEGA 7.0.18. The result showed an overlapping distribution of interspecific divergence and intraspecific distance for *matK* region. The phylogeny indicated that some of *Syzygium cumini* are embedded within other *Syzygium* genera. Therefore, *matK* gene is not suitable to be used as barcode region for identification of *Syzygium* species.

Keywords: *Syzygium*, *Syzygium cumini*, DNA barcode, DNA barcoding gap, intraspecific, interspecific

Trina E. Tallei\*, Presticilla D. Irawan, Beivy J. Kolondam

Department of Biology, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University

Corresponding author: trina\_tallei@unsrat.ac.id

BIOINFORMATICS WORKSHOP 2016: Developing knowledge and skill in bioinformatics for Young Indonesian Scientists in improving research quality in life science and sustainable exploration of biodiversity in Indonesia. Al Azhar University Jakarta, 13 – 15 September 2016

## ABSTRACT

*Syzygium* is a genus of flowering plants that belongs to the myrtle family (Myrtaceae). Accurate identification of *Syzygium* species using traditional methods can take a long time due to lack of knowledge about plants and/or lack of flowers and fruit characters required for identification. Reliability or accuracy of identification relies on expert determination, which is not always readily available. Moreover, *Syzygium* genera exhibit many overlapping characters, making them difficult to identify. In this research, the utility of *matK* gene as DNA barcoding in *Syzygium* was assessed. The assessment included barcoding gap, intraspecific and interspecific divergence, as well as phylogenetic tree reconstruction. The evolutionary history was inferred by using UPGMA method. Estimation of evolutionary divergence between sequence was analysed using Kimura's two-parameter. Evolutionary analyses were conducted in MEGA7.0.18. The result showed an overlapping distribution of interspecific divergence and intraspecific distance for *matK* region. The phylogeny indicated that some of *Syzygium cumini* are embedded within other *Syzygium* species. Therefore, *matK* gene is not suitable to be used as barcode region for identification of *Syzygium* species.

Keywords: *Syzygium*, *Syzygium cumini*, DNA barcode, DNA barcoding gap, intraspecific, interspecific

## INTRODUCTION

*Syzygium* is a genus comprising of many species. According to the Plant List, at least there are 1.374 species in the genus *Syzygium*. However, only 1.123 species names (81.7%) are accepted. There are 240 (17.5%) synonym of species names. This number is relatively high. This is due to the difficulty of identification of *Syzygium* species based on morphological characters. Some species showed overlapping characters, especially their flowers. The confidence level of taxa placement for *Syzygium* is 98.0%. Morphological character based-identification sometimes cannot be relied upon, because it can be influenced by the environment (Khan et al. 2005). This can lead to difficulty in the identification process. DNA barcoding provides a way out to this problem, for it is able to identify specimens using a very short fragment of gene sequences obtained from a small amount of tissue (Weitschek et al. 2014; Kairupan et al. 2015; Rembet et al. 2016).

MaturaseK (*matK*) and Rubisco (*rbcL*) genes are two standard genes used in plant DNA barcoding recommended by Consortium for Barcode of Life (CboL) (Lawodi et al. 2013). Both genes have important roles in the genetic distribution of species as well as in the reconstruction of plant phylogenetic. Compared to *rbcL*, *matK* has three times higher of nucleotide substitution and six times higher of amino acid substitution. The *matK* is a 1500 bp plastid gene located between *trnK* intron (Hilu and Liang 2007; Goyal and Sen 2015; Tallei and Kolondam 2015). In order to assess the reliability of DNA barcode in *Syzygium* species differentiation, we employed *matK* gene.

## Materials and Methods

### Sampling Material, DNA Extraction, Amplification, and Sequencing

Young leaf of *Syzygium cumini* was obtained from Batuputih Village, North Sulawesi. Genomic DNA was extracted from approximately 40 mg young leaf using AxyPrep™ Multisource Genomic DNA Miniprep Kit with a slight modification. Amplification of *matK* Region was performed using primer sets MatK-1RKIM-f 5'-ACCCAGTCCATCTGGAAATCTTGGTTC-3' and MatK-3FKIM-r 5'-CGTACAGTACTTTTGT GTTTACGAG-3' designed by K. J. Kim from School of Life Sciences and Biotechnology, Korea University, Seoul, Korea. DNA was amplified using 5X Firepol PCR Master Mix Read-to-Load (Solis Biodyne). Total volume for amplification was 40 mL, consisted of 2 mL (0.6 µg) DNA sample and 1.5 mL of each primer (10 mM). DNA amplification was carried out in PCR TPersonal (Biometra). The process began with pre-denaturation at 95°C for 2 min, followed by 35 cycles of 30s denaturation at 95°C, 30s of annealing at 50°C, and 50s of elongation at 72°C, with a final extension for 2 min at 72°C. A single clear cut band fragment was purified and sequenced bi-directionally at First BASE Laboratory Malaysia.

### Data Analysis

The resulting sequences were processed using Geneious V6.1.6. The sequences were pairwise aligned using global alignment with free end gaps and to identify regions of 93% similarity. The region of similarity was subsequently extracted, producing a consensus DNA sequences which were used for further analysis. A BLAST search was performed to identify the closest allied of *Syzygium cumini*. Other closest *Syzygium* allies were retrieved from GenBank. All closest *matK* sequences were subjected to final alignment using Multalin V.5.4.1 (<http://multalin.toulouse.inra.fr/multalin/>). Sequences were trimmed at both ends of the alignment to avoid many missing data at the end of the sequences, leaving 677 final characters of each sequence. Phylogenetic tree for nucleotide sequence was constructed using MEGA7.0.18.

**Acknowledgment:** This research was financed by Sam Ratulangi University through University Competitive Research Grant Scheme (Riset Unggulan Universitas) Fiscal year 2006

## Results

Figure 1. Evolutionary relationship of *Syzygium* species. The evolutionary history was inferred using the UPGMA method. The evolutionary distances were computed using the Kimura's 2-parameter method and are in the units of the number of base substitutions per site. Evolutionary analysis was conducted in MEGA7.0.18.

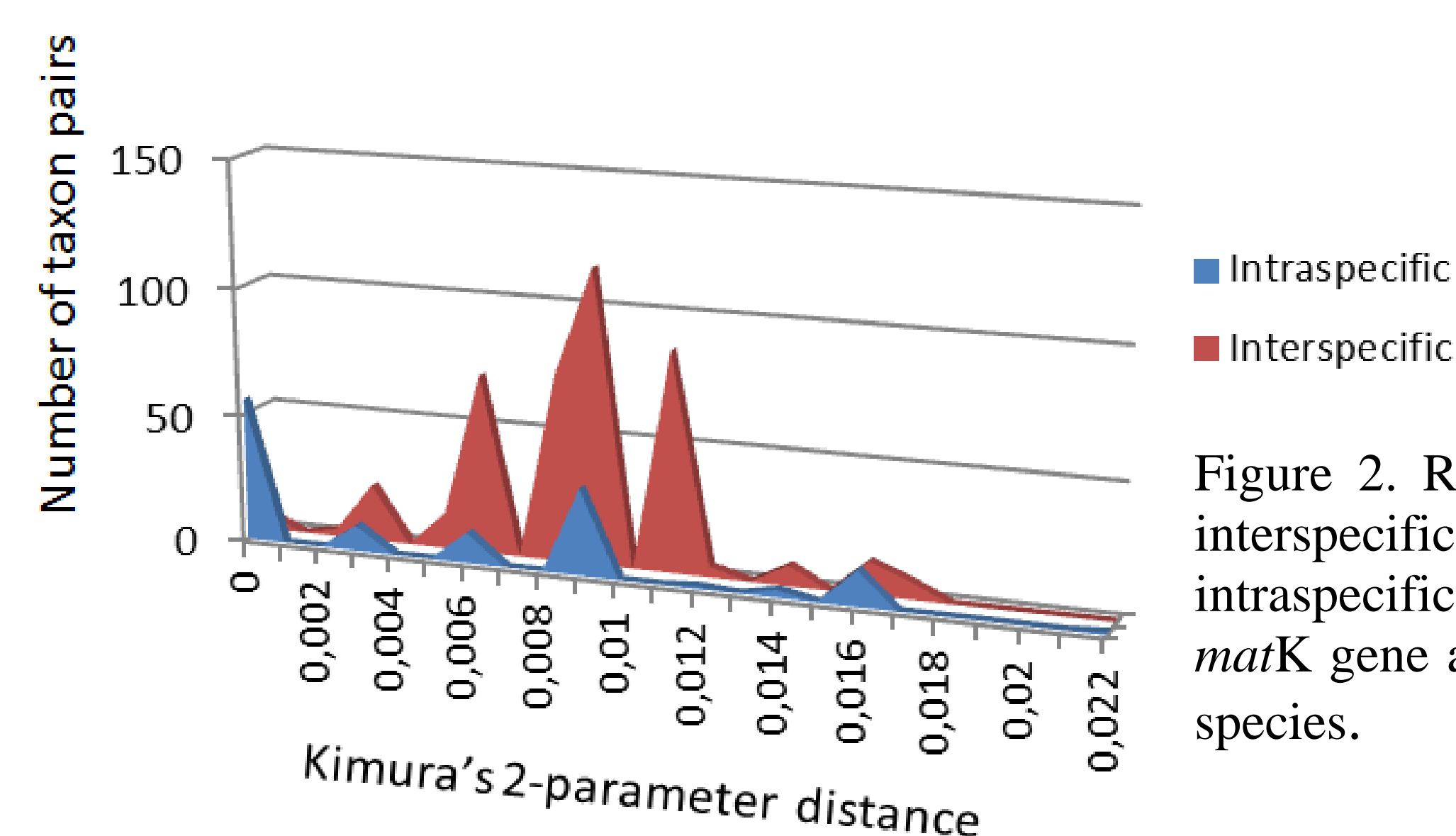
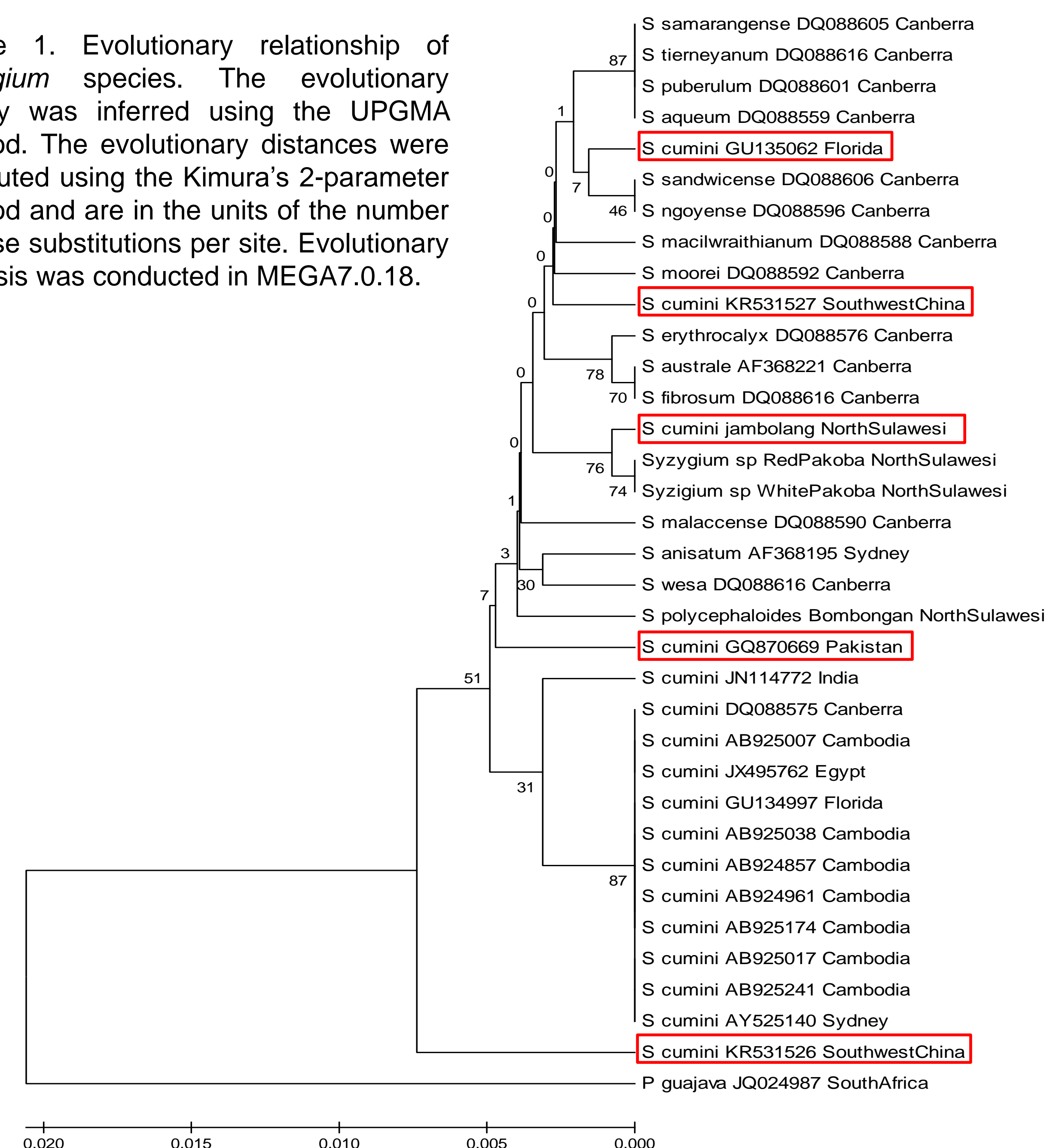


Figure 2. Relative distribution of interspecific divergence (red) and intraspecific distances (blue) in the *matK* gene among some *Syzygium* species.

## Discussion

According to the evolutionary relationship (Figure 1), some of *Syzygium cumini* are embedded in other closely related *Syzygium* species. Figure 2 shows the overlap between intraspecific distance and interspecific divergence. Ideally, DNA barcodes must exhibit a "barcoding gap" between interspecific divergence and intraspecific distance. This means *matK* gene alone can not be used for DNA barcode in some *Syzygium* species. The existence of barcoding gap provides assurance that the genetic distance can be used to determine the species name for an unknown specimen, for example by looking at the name of taxa closest to the specimen. The absence of barcoding gap could mean there was an error in interpreting the data, for example taxa might be the same species or classified as a subspecies despite having different morphologies. Molecular taxonomy-based approach may suggest the potential for discovery of new species. But the existence of the new species should be supported by well-integrated approach to the study of morphology.

## Conclusion

This study infers that *matK* gene alone is not suitable to be used as barcode region for *Syzygium* species.

## References

- Goyal PA, Sen A. 2015. Maturase K gene in plant DNA barcoding and phylogenetics. *Plant DNA Barcoding and Phylogenetics* 5: 79-90.
- Hilu KH, Liang H. 1997. The *matK* gene: sequence variation and application in plant systematics. *Am J Bot* 84(6): 830-839.
- Kairupan CF, Koneeri R, Tallei TE. 2015. Variasi Genetik *Troides helena* (Lepidoptera: Papilionidae) Berdasarkan Gen COI (Cytochrome C Oxidase I). *Journal MIPA UNSRAT ONLINE* 4(2):141-147
- Khan IA, Awan FS, Ahmad A, Fu YB, Iqbal A. 2005. Genetic diversity of Pakistan wheat germplasm as revealed by RAPD markers. *Genet Resour Crop Evol* 52(3): 239-244.
- Lawodi EN, Tallei TE, Mantri FR, Kolondam BJ. 2013. Variasi genetik tanaman tomat dari beberapa tempat di Sulawesi berdasarkan gen *matK*. *Pharmakon Jurnal Ilmiah Farmasi* 2(4): 114-121
- Rembet R, Pelealu JJ, Kolondam BJ, Tallei TE. 2016. Analisis sekuens gen *matK* *Sansevieria trifasciata* var. *Laurentii* dan var. *Hahnii*. *Pharmakon Jurnal Ilmiah Farmasi* 5(2): 99-106.
- Tallei TE, Kolondam BJ. 2015. DNA barcoding of Sangihe nutmeg (*Myristica fragrans*) using *matK* gene. *HAYATI Journal of Biosciences* 22(1): 41-47
- The Plant List .2016. Version 1. Published on the Internet <http://www.theplantlist.org/browse/A/ Myrtaceae/Syzygium/> (accessed 21st August 2016).
- Weitschek E, Fison G, Felici G. 2014. Supervised DNA Barcodes species classification: analysis, comparisons and results. *BioData Mining* 7(4):1-18