TREEasy: An automated workflow to infer gene trees, species trees, and phylogenetic networks from multilocus data

doi: info:doi/10.1111/1755-0998.13149

DR. YAFEI MAO (Orcid ID : 0000-0002-9648-4278)
MR. SIQING HOU (Orcid ID : 0000-0003-4062-5868)

Article type : Resource Article

Title:

TREEasy: an automated workflow to infer gene trees, species trees, and phylogenetic networks from multilocus data

Authors: Yafei Mao[1+*], Siqing Hou[2], Junfeng Shi[3] and Evan P. Economo[1]

Affiliations:

[1]Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[2]Cognitive Neurorobotics Research Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[3] Shanghai Key Laboratory of Stomatology & Shanghai Research Institute of Stomatology, Shanghai Ninth People's Hospital, College of Stomatology, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

*Correspondence to: Yafei Mao: yafmao@uw.edu

+Present address: Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA

Yafei Mao: yafei.mao@oist.jp
Siqing Hou: siqing.hou2@oist.jp
Junfeng Shi: strangephone@126.com
Evan P. Economo: economo@oist.jp

Abstract

Multilocus genomic datasets can be used to infer a rich set of information about the evolutionary history of a lineage, including gene trees, species trees, and phylogenetic networks. However, user-friendly tools to run such integrated analyses are lacking, and workflows often require tedious reformatting and handling time to shepherd data through a series of individual programs. Here, we present a tool written in Python—TREEasy—that performs automated sequence alignment (with MAFFT), gene tree inference (with IQ-Tree), species inference from concatenated data (with IQ-Tree and RaxML-NG), species tree inference from gene trees (with ASTRAL, MP-EST, and STELLS2), and phylogenetic network inference (with SNaQ and PhyloNet). The tool only requires FASTA files and nine parameters as inputs. The tool can be run as command line or through a Graphical User Interface (GUI). As examples, we reproduced a recent analysis of staghorn coral evolution, and performed a new analysis on the evolution of the "WGD clade" of yeast. The latter revealed novel patterns that were not identified by previous analyses. TREEasy represents a reliable and simple tool to accelerate research in systematic biology (https://github.com/MaoYafei/TREEasy).

KEYWORDS: Species tree; Phylogenetic network; Gene trees; Introgression; Phylogenetic inference; Pipeline/Workflow

Introduction

The inference of evolutionary history from molecular data is a core goal of modern evolutionary biology (Barraclough & Nee, 2001; Soltis & Soltis, 2018). With the increasing availability of large-scale multilocus datasets and advances in computational power, phylogenetic methods have diversified in the past two decades (Delsuc, Brinkmann, & Philippe, 2005; Liu, Xi, Wu, Davis, & Edwards, 2015). Instead of inferring a bifurcating tree with a single locus as the focus of analysis, biologists regularly infer populations of trees representing the histories of different loci (Edwards, Liu, & Pearl, 2007; Gadagkar, Rosenberg, & Kumar, 2005). From these, species tree methods can be used which take into account the fact that gene trees can be discordant with species trees even under a bifurcating evolutionary history (Kubatko & Degnan, 2007; Lambert, Reeder, & Wiens, 2015; Page & Charleston, 1997; Shen, Salichos, & Rokas, 2016; Tonini, Moore, Stern, Shcheglovitova, & Ortí, 2015). In addition, introgression is relatively common occurrence across the tree of life (Berner & Salzburger, 2015; Bravo et al., 2019; Morrison, 2014; Xu, 2000), thus, evolutionary histories are not always bifurcating (Bravo et al., 2019; Degnan & Rosenberg, 2009; Gadagkar et al., 2005; Page & Charleston, 1997). Phylogenetic network methods can also be used to infer evolutionary histories that include reticulation (Bastide, Solis-Lemus, Kriebel, William Sparks, & Ane, 2018; Huson & Bryant, 2005).

In total, these methods increasingly reflect the complexity of evolution, and for each of these analyses types, multiple programs are available to the researcher. For example, methods allowing inference of species trees and phylogenetic networks include NJst (Liu & Yu, 2011), MP-EST (Liu, Yu, & Edwards, 2010), ASTRAL (Mirarab et al., 2014), STELLS2 (Pei & Wu, 2017), Guenomu (de Oliveira Martins & Posada, 2017), SNaQ (Solís-Lemus, Bastide, & Ané, 2017) and PhyloNet (Wen, Yu, Zhu, & Nakhleh, 2018). However, each method requires gene tree input and control files in different formats. In particular, ASTRAL requires an unrooted gene tree list, whereas MP-EST requires a rooted gene tree list. In addition, Guenomu, a Bayesian hierarchical model, requires posterior distributions of gene trees. SNaQ runs in Julia language, whereas PhyloNet runs in a command line with a special control file. Thus, idiosyncratic preliminary work is needed to prepare inputs to run these tools.

The need for a researcher to figure out how to reformat files and run multiple individual programs is a common issue in modern phylogenomic analysis, reducing efficiency and making it less likely that a broad range of methods and programs are used. To address this, several other pipelines have been developed to integrate workflows for phylogenomic-related analyses. STRAW was developed as a Web-based server requiring a gene tree list as input to infer species tree with STAR, MP-EST, and NJst (Shaw, Ruan, Glenn, & Liu, 2013), but it only runs for rooted gene trees and cannot directly take sequences as input and cannot infer phylogenetic networks. In addition, PhyloToL (Cerón-Romero et al., 2019), HybPhyloMaker (Fér & Schmickl, 2018) and ezTree (Wu, 2018) are designed to reconstruct species trees from raw reads, as well, ParGenes (Morel, Kozlov, & Stamatakis, 2018) and NGPhylogeny.fr (Lemoine et al., 2019) are used to infer species tree form aligned sequences and unaligned sequences, respectively.

Several of the tools listed already automate various steps of the phylogenomics workflow, but despite their utility for many purposes, there is no single platform available to integrate sequence alignment, gene tree reconstruction, species tree, and phylogenetic network inferences into a single run. PhyloTol (Cerón-Romero et al., 2019) in particular has an overlapping workflow to the one presented here, with a suite of tools for gene family assessment and gene tree estimation, but is limited to concatenation-based approaches for species tree inference and does not include phylogenetic network inference. The latter steps have been shown to be important for enhancing inference of evolutionary history in many groups (Edwards, Liu, & Pearl, 2007). To address this gap, we present a multi-thread open-source tool, named TREEasy, to shepherd data through a series of programs to infer gene trees, species trees, and phylogenetic networks from molecular sequences. In addition to reconstructing species trees with supermatrix and multispecies coalescent methods, TREEasy can infer phylogenetic networks from two different programs from unaligned sequences.

TREEasy architecture

TREEasy is written in Python integrating sequence alignment, gene tree reconstruction, species tree inference, and phylogenetic network inference.

Component software choice and justification

Although there are sometimes many options of different software available for different phylogenomic analyses, when possible we generally chose popular workhorse programs given their wide familiarity in the field and relatively understood behavior. We also considered computational efficiency and diversity of methods and research groups in selecting among different options. For sequence alignment, we chose the program MAFFT (Nakamura, Yamada, Tomii, & Katoh, 2018). There are numerous programs available for maximum-likelihood reconstruction of gene trees, but we chose two of the best performing and most heavily used, IQ-Tree (Nguyen, Schmidt, von Haeseler, & Minh, 2014) and RaxML-NG (Kozlov, Darriba, Flouri, Morel, & Stamatakis, 2019). Considering of the diversity of software developers and performance of software (Pei & Wu, 2017), for species-tree reconstruction from gene trees, we used MP-EST (Liu et al., 2010), ASTRAL (Mirarab et al., 2014), and STELLS2 (Pei & Wu, 2017), which have the same overall aim but differ in implementation. Finally, we chose PhyloNet and SNaQ, to infer phylogenetic networks, two reliable and the best performing programs suitable for these data (Solís-Lemus et al., 2017; Wen et al., 2018).

Installation and subroutines

BioPython must be installed and a few executable dependencies are needed: MAFFT (Nakamura et al., 2018), Translatorx (Abascal, Zardoya, & Telford, 2010), AMAS (Borowiec, 2016), IQ-TREE (Nguyen et al., 2014), RAxML-NG (Kozlov et al., 2019), ASTRAL (Mirarab et al., 2014), MP-EST (Liu et al., 2010), STELLS2 (Pei & Wu, 2017), PhyloNet (Wen et al., 2018) and SNaQ (Solís-Lemus et al., 2017). Molecular sequences in FASTA format (SNP, microsatellites, protein-coding sequences, etc.) are mandatory inputs. In addition, to identify multiple individuals in a species and identify mismatches between species names in different gene files, two text files including species numbers and gene names respectively are needed. There are 3 subroutines in TREEasy as follows (Figure 1).

(1) Gene tree reconstruction

Molecular sequences are aligned using MAFFT with localpair model and then gene trees are reconstructed with Maximum likelihood (ML) method in IQ-TREE with model selection. This process runs as parallel processing with the threading module in python.

(2) Species tree inference

Firstly, alignments of multi-loci are concatenated to build a concatenated species tree using IQ-TREE and RAxML-NG. Then, the tool selects gene trees of which each node's bootstrap value is greater than B (B is a preset parameter from 0 to 100) in order to avoid uncertainty of gene tree reconstruction. Next, un-rooted selected gene trees generated by IQ-TREE are put together as input to infer a species tree using ASTRAL. Meanwhile, the un-rooted gene trees are rooted with a preset parameter R (species name(s)) and then the rooted gene trees are used to infer species trees using STELLS2 and MP-EST.

(3) Phylogenetic network inference

A species tree generated by ASTRAL and the un-rooted gene trees are used to infer a phylogenetic network using SNaQ. Then, the rooted gene trees are used to infer a phylogenetic network using PhyloNet.

The tool can be run as command line or through Graphical User Interface (GUI) for users. The GUI interface can be seen in Figure 2.

Datasets and analyses

Evaluation with simulated data

We used simulated data from the published study (Solís-Lemus et al., 2017) to evaluate following aspects of TREEasy (Figure 3): running time and memory usage with (1) different processors (Figure 1A); (2) with different gene numbers (Figure 1B); (3) with different taxon numbers (Figure 1C).

First, with 6 taxa and 300 genes, we found that the running time decreased with the increase of processor number and the speed with 12 processors was 6.5 times faster compared to with 1 processor (Figure 1A). Yet, the maximum memory usage did not show a significant change. Second, with 6 taxa and 4 processors, we found that the running time and maximum memory usage increased with increase of gene number (Figure 1B). Third, with 300 genes and 12 processors, we found that the running time excluding the run of PhyloNet increased from 6 taxa to 15 taxa, but the maximum memory usage was increased dramatically from 10 taxa to 15 taxa (Figure 1C). It is worth noting that PhyloNet running was extremely slow with > 10 taxa (> 7 days) and thus we excluded the running of PhyloNet in this analysis.

Empirical data validation

As a second test, we used TREEasy to reproduce a previous analysis of *Acropora* genome evolution that inferred reticulation events among five coral species (Mao, Economo, & Satoh, 2018). The *Acropora* data included 4,945 single-copy orthologs among five *Acropora* species. The whole process took ~13 hours with maximum memory usage: 3,992 Mb, running on 8 processors.

We found that the concatenated species tree has the same topology as the other species trees inferred from gene trees with ASTRAL, MP-EST, and STELLS2. Then, the inferred phylogenetic network topologies with SNaQ and PhyloNet are identical (Figure 4). Both of these results are coincident with our previous study (Mao et al., 2018).

Inferring species trees and phylogenetic networks of the "WGD clade" of yeast

As a third test, we applied TREEasy to data from a recent study that did not perform all the analyses presented here. The previous study investigated the evolutionary relationships of subphylum *Saccharomycotina* based on hundreds of yeast genomes (Shen et al., 2018). In particular, there is a clade ("WGD clade") including common and important yeasts such as the baker's yeast (Gonçalves et al., 2016; Ludlow et al., 2016), and there are two "non-robust internodes" in this clade. In addition, introgression has been reported in yeast (Leducq et al., 2016; Marcet-Houben & Gabaldón, 2015). Therefore, in order to conduct a preliminary investigation into whether the "non-robust internodes" were caused by introgression, we first retained sequences for 40 species from the "WGD clade" and an outgroup species (*Neurospora crassa*) with no missing data from two datasets (2408OG dataset and 1292BUSCO dataset). All horizontal gene transfer (HGT) genes were removed in these two datasets. 320 genes and 777 genes were extracted from the 1292BUSCO and the 2408OG datasets respectively and we applied TREEasy on these two datasets. Then, we found that the species trees inferred from different methods or datasets were not identical. Moreover, most of incongruences between the inferred species trees were located on the "non-robust internodes" (Figure 5).

Next, we extracted the two sub-clades including "non-robust internodes" ("*Saccharomyces*" clade and "*Kazachstania*" clade) with two species as outgroups (*Yueomyces sinensis* and *Tetrapisispora blattae*) and run these two clades on TREEasy. After filtering the gene trees with bootstrap values smaller than 30 (B parameter), in the 1292BUSCO dataset, we found 279 gene trees and 115 gene trees in "*Saccharomyces*" clade and "*Kazachstania*" clade, respectively. In the 2408BUSCO dataset, we found 654 gene trees and 351 gene trees in "*Saccharomyces*" clade and "*Kazachstania*" clade, respectively.

In order to reduce bias of species tree and phylogenetic network inferences from small gene tree numbers, we only reported the phylogenetic networks for "*Saccharomyces*" (Figure 6A) and

"*Kazachstania*" clades (Figure 6B) for the 2408BUSCO dataset here. We found a signal of introgression in the 2408BUSCO dataset and some reticulate events occurred in the "non-robust internodes" of "*Saccharomyces"* genus and "*Kazachstania"* clades. In addition, introgression was detected for some lineages even when their internode supports were robust across different analyses.

Discussion

The era of big data and massive computing resources has allowed us to better understand species and population relationships (Allen et al., 2019; Bravo et al., 2019; Delsuc et al., 2005; Liu, Wu, & Yu, 2015). It is now possible to infer species trees and phylogenetic networks from hundreds of gene trees rather than concatenating a few loci to reconstruct a phylogeny. We developed a reliable and efficient tool called TREEasy to infer species trees and networks from molecular sequences directly. TREEasy is written in Python and can be run with multi-processors (https://github.com/MaoYafei/TREEasy).

First, the multiple threading module improved the running time in TREEasy. The running time improved 6.5 times with 12 processors compared to with 1 processor. One of possible reasons is that parallelization by the threading module in python is applied to sequence alignment and gene tree reconstruction in TREEasy. Meanwhile, with gene or taxon number increasing, both running time and maximum memory usage increased as expected. Interestingly, the increase of taxon number has a greater effect on running time compared to the increase of gene number while the increase of gene number has more effects on memory usage compared to the increase of taxon number. One possible reason is that searching tree space results in increases of running time, while heavy computation on gene tree reconstruction during parallelization leads to more memory requirements. Increasing taxon number quickly expands the tree search space of phylogenetic networks by order of magnitudes (Solís-Lemus et al., 2017; Wen et al., 2018).

Second, our pipeline provides an easy and robust way to infer species trees and phylogenetic networks. The empirical validation of five species of *Acropora* generated the same result as our previous study (Mao et al., 2018), and thus, this result suggests TREEasy is a reliable tool for species tree and network inferences. Next, in order to evaluate TREEasy with more taxa, we applied TREEasy to a newly-sequenced yeast genomic data (Shen et al., 2018). Our results both confirm previous results. First, the incongruence between the inferred species trees showed on two clades ("*Saccharomyces*" clade and "*Kazachstania*" clade) identical to the previous study. Moreover, the inferred yeast phylogenetic networks suggest introgression occurred in the two clades of yeasts and introgression is a possible reason to cause the incongruence between the inferred species trees. Interestingly, we also found some reticulate events occurred in the yeast lineages which had the same topology from different species tree inferences. These results show that TREEasy would be easily applied to a genomic study (as long as there is no missing data) on species relationships. Notably, we analyzed yeast evolution as a test case to show how novel patterns can be revealed from the workflow, however any conclusions about yeast evolution should be treated in a dedicated study by experts in those groups.

There are currently some limitations we need to mention, mostly due to constraints inherited from the underlying programs. First, the pipeline does not accept missing data due to limitations of the phylogenetic network inference programs (SNaQ and PhyloNet) (Solís-Lemus et al., 2017; Wen et al., 2018). In addition, the pipeline is better suited for analyses with a relatively small number of taxa, datasets including hundreds of taxa is not suitable to phylogenetic network inference. Moreover, to date, we did not implement gene tree concordance analysis in the tool, if users are interested in this we recommended users to use IQ-Tree (V2) to perform related analysis (http://www.iqtree.org/doc/Concordance-Factor).

Finally, we want to emphasize that while the intention of this pipeline is to make it easier to run workflows encompassing a variety of phylogenomic analyses, users still need to familiarize themselves with the constituent programs and their settings. In other words, the program should not be treated as a "black box". The main value of a pipeline program like TREEasy is to handle time-consuming and not scientifically-pertinent formatting and scripting tasks, but does not remove the need for expert knowledge of the included programs and their assumptions, settings, and limitations. To facilitate checking at different steps, TREEasy keeps all important intermediate results such as gene trees and log files, they can be inspected after for quality control. We would like to remind users that independent runs for the same dataset are necessary to reach conclusive results, as well as to cite all constituent programs if using the pipeline, not only TREEasy.

In all, this study presents a reliable and user-friendly tool to infer species trees and networks from molecular sequences directly, has the potential to be used widely in population genetic/genomic, phylogenomic and phylogeographic studies. We hope that this will lower barriers to analyses of evolutionary history and accelerate research in systematic biology.

ACKNOWLEDGEMENTS

This study was supported by OIST and was funded by JSPS grants (No.17J00557 to Y.M. and Kakenhi No.17K15180 to E.P.E.). The authors acknowledge support of the OIST Scientific Computing Support Section. We additionally thank Nitish Narula for helping to test the program.

AUTHOR CONTRIBUTIONS

Y.M. and E.P.E. conceived the study. Y.M. performed all analyses in this study. S.H. and J.S. built the GUI. Y.M. and E.P.E. wrote the initial manuscript and edited the final manuscript.

DATA ACCESSIBILITY

The following information was supplied regarding data availability:
GitHub: https://github.com/MaoYafei/TREEasy.

REFERENCES

Abascal, F., Zardoya, R., & Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res, 38*(Web Server issue), W7-13. doi:10.1093/nar/gkq291

Allen, J. M., Germain-Aubrey, C. C., Barve, N., Neubig, K. M., Majure, L. C., Laffan, S. W., . . . Soltis, P. S. (2019). Spatial Phylogenetics of Florida Vascular Plants: The Effects of Calibration and Uncertainty on Diversity Estimates. *iScience, 11*, 57-70. doi:10.1016/j.isci.2018.12.002

Barraclough, T. G., & Nee, S. (2001). Phylogenetics and speciation. *Trends Ecol Evol, 16*(7), 391-399. doi:10.1016/s0169-5347(01)02161-9

Bastide, P., Solis-Lemus, C., Kriebel, R., William Sparks, K., & Ane, C. (2018). Phylogenetic Comparative Methods on Phylogenetic Networks with Reticulations. *Syst Biol, 67*(5), 800-820. doi:10.1093/sysbio/syy033

Berner, D., & Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet, 31*(9), 491-499. doi:10.1016/j.tig.2015.07.002

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ, 4*, e1660. doi:10.7717/peerj.1660

Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., . . . Edwards, S. V. (2019). Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ, 7*, e6399. doi:10.7717/peerj.6399

Cerón-Romero, M. A., Maurer-Alcalá, X. X., Grattepanche, J.-D., Yan, Y., Fonseca, M. M., & Katz, L. A. (2019). PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. *Mol Biol Evol*.

de Oliveira Martins, L., & Posada, D. (2017). Species Tree Estimation from Genome-Wide Data with guenomu. *Methods Mol Biol, 1525*, 461-478. doi:10.1007/978-1-4939-6622-6_18

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet, 6*(5), 361-375. doi:10.1038/nrg1603

Edwards, S. V., Liu, L., & Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences, 104*(14), 5936-5941.

Fér, T., & Schmickl, R. E. (2018). HybPhyloMaker: target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics, 14*, 1176934317742613.

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 304*(1), 64-74.

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., . . . Sampaio, J. P. (2016). Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Current Biology, 26*(20), 2750-2761.

Huson, D. H., & Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol, 23*(2), 254-267.

Kozlov, A., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *bioRxiv*, 447110.

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol, 56*(1), 17-24.

Lambert, S. M., Reeder, T. W., & Wiens, J. J. (2015). When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. *Molecular Phylogenetics and Evolution, 82*, 146-155.

Leducq, J.-B., Nielly-Thibault, L., Charron, G., Eberlein, C., Verta, J.-P., Samani, P., . . . Landry, C. R. (2016). Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature Microbiology, 1*(1), 15003.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., & Gascuel, O. (2019). NGPhylogeny. fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res*.

Liu, L., Wu, S., & Yu, L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution, 53*(5), 380-390.

Liu, L., Xi, Z., Wu, S., Davis, C., & Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *arXiv preprint arXiv:1501.03578*.

Liu, L., & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Syst Biol, 60*(5), 661-667.

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *Bmc Evolutionary Biology, 10*(1), 302.

Ludlow, C. L., Cromie, G. A., Garmendia-Torres, C., Sirr, A., Hays, M., Field, C., . . . Dudley, A. M. (2016). Independent origins of yeast associated with coffee and cacao fermentation. *Current Biology, 26*(7), 965-971.

Mao, Y., Economo, E. P., & Satoh, N. (2018). The roles of introgression and climate change in the rise to dominance of Acropora Corals. *Current Biology, 28*(21), 3373-3382. e3375.

Marcet-Houben, M., & Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol, 13*(8), e1002220.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics, 30*(17), i541-i548.

Morel, B., Kozlov, A. M., & Stamatakis, A. (2018). ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics, 35*(10), 1771-1773.

Morrison, D. A. (2014). Is the tree of life the best metaphor, model, or heuristic for phylogenetics? *Syst Biol, 63*(4), 628-638.

Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics, 34*(14), 2490-2492.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol, 32*(1), 268-274.

Page, R. D. M., & Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution, 7*(2), 231-240.

Pei, J., & Wu, Y. (2017). STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. *Bioinformatics, 33*(12), 1789-1797.

Shaw, T. I., Ruan, Z., Glenn, T. C., & Liu, L. (2013). STRAW: species TRee analysis web server. *Nucleic Acids Res, 41*(W1), W238-W241.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., . . . Doering, D. T. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell, 175*(6), 1533-1545. e1520.

Shen, X.-X., Salichos, L., & Rokas, A. (2016). A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biology and Evolution, 8*(8), 2565-2580.

Solís-Lemus, C., Bastide, P., & Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol Biol Evol, 34*(12), 3292-3298.

Soltis, D., & Soltis, P. (2018). *The Great Tree of Life*: Academic Press.

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., & Ortí, G. (2015). Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS currents, 7*.

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Syst Biol, 67*(4), 735-740.

Wu, Y.-W. (2018). ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes. *BMC Genomics, 19*(1), 921.

Xu, S. (2000). Phylogenetic analysis under reticulate evolution. *Mol Biol Evol, 17*(6), 897-907.

Figure legends

Figure 1. Workflow in TREEasy. The orange oval represents inputs. Blue boxes represent subroutines and red boxes represent outputs.

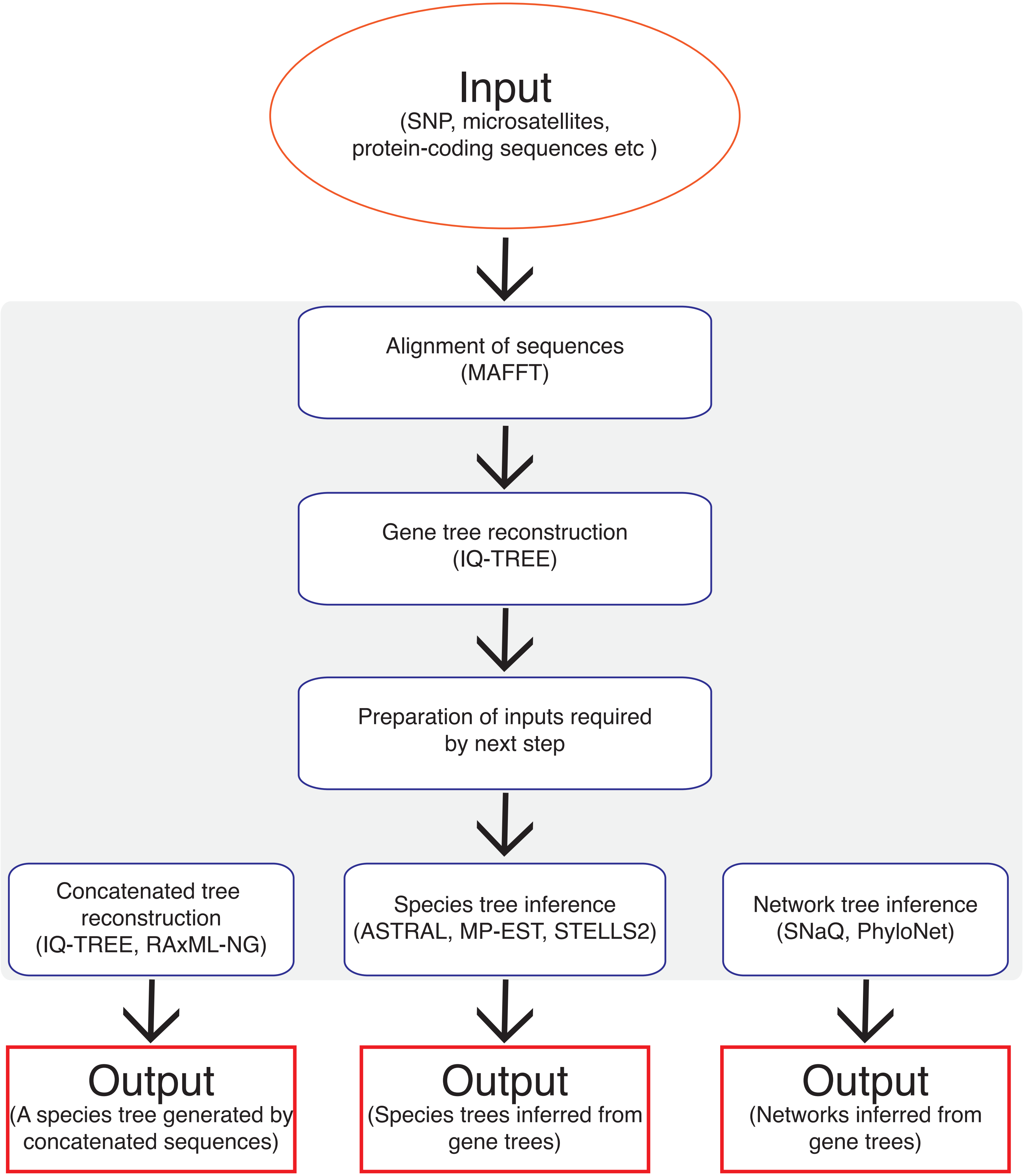
Figure 2. GUI windows of TREEasy. (A) The start window. (B) The window shows the simulated data example. (C) The window shows a successful run of the simulated data example.

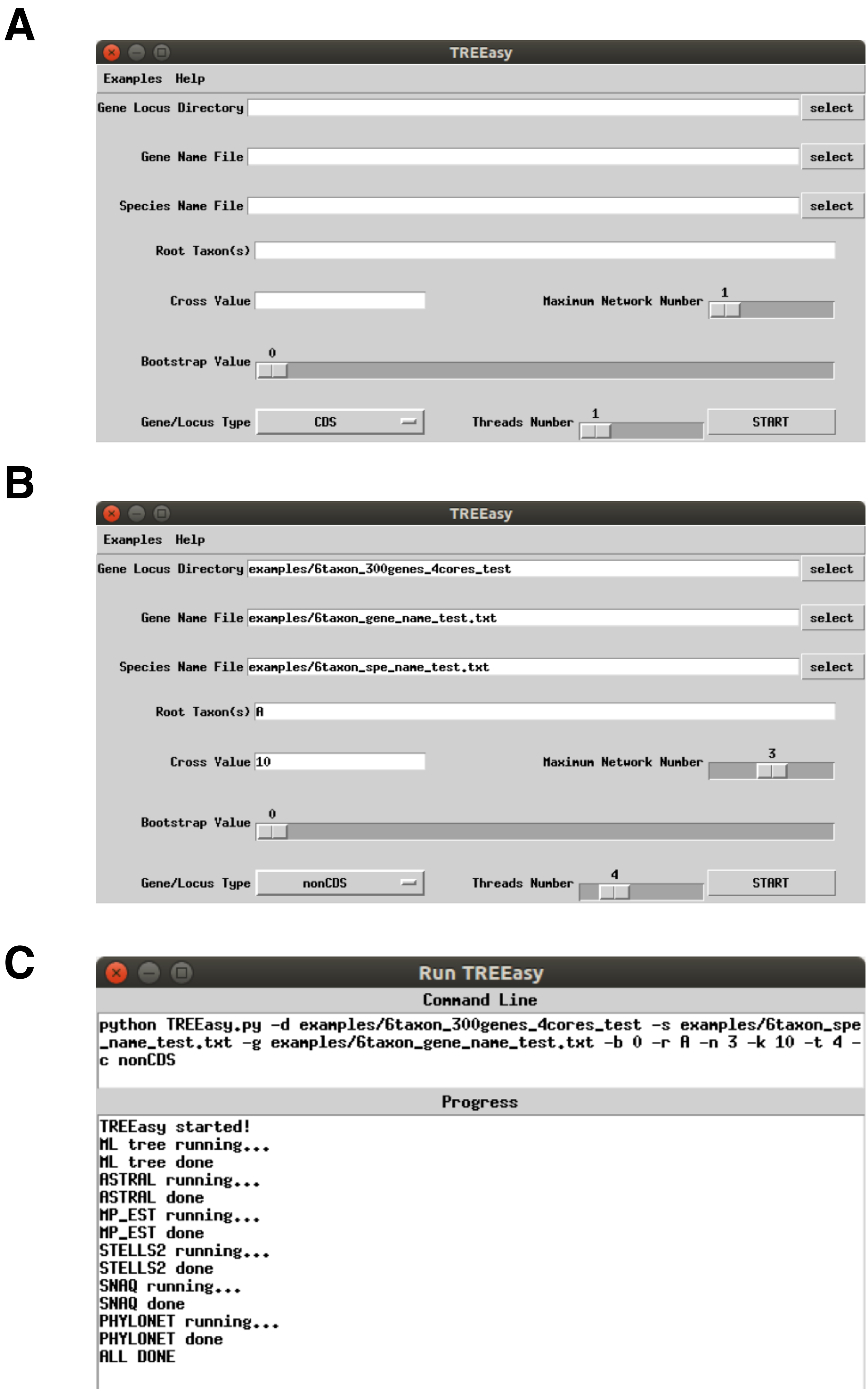
Figure 3. Estimation of TREEasy on simulated data. (A) Running time and maximum memory usage versus processor number (6 taxa and 300 genes). (B) Running time and maximum memory usage versus gene number (6 taxa and 4 processors). (C) Running time and maximum memory usage versus taxon number (300 genes and 12 processors).

Figure 4. Validation of TREEasy on empirical data. TREEasy running on 4, 945 single-copy orthologs of *Acropora* generated species trees by (A) concatenated method, (B) ASTRAL, (C) MP-EST, and (D) STELLS2; and phylogenetic networks by (E) SNaQ and (F) PhyloNet.

Figure 5. A case study of TREEasy on "WGD clade" of yeast genomic data. (A) The topology of "WGD clade" of yeast in the previous study. TREEasy running on two datasets generated species trees by (B, C) concatenated method, (D, E) ASTRAL, (F, G) MP-EST, and (H, I) STELLS2. The results (B, D, F, H) were generated by 1292BUSCO dataset and the results (C, E, G, I) were generated by 2408OG dataset. The blue shadows in (A) represents the "non-robust internodes" and the orange and yellow shadows represent two sub-clades: "*Saccharomyces*" clade and "*Kazachstania*" clade, for phylogenetic network analysis. The red branches represent the incongruences among phylogenetic trees.

Figure 6. Phylogenetic network inferences for two sub-clades of yeast genomic data. Phylogenetic networks inferred by SNaQ on "*Saccharomyces*" clade (A) and "*Kazachstania*" clade (B) from the 2408OG dataset. The orange shades represent reticulate events occurred in "non-robust internodes". The blue shades represent reticulate events occurred in lineages which had the same topology from different species tree inference methods.


Input
(SNP, microsatellites,
protein-coding sequences etc)
Alignment of sequences
(MAFFT)
Gene tree reconstruction
(IQ-TREE)
Preparation of inputs required
by next step
Concatenated tree
reconstruction
(IQ-TREE, RAxML-NG)
Species tree inference
(ASTRAL, MP-EST, STELLS2)
Network tree inference
(SNaQ, PhyloNet)
Output
(A species tree generated by
concatenated sequences)
Output
(Species trees inferred from
gene trees)
Output
(Networks inferred from
gene trees)



A

Gene Locus Directory
Gene Name File
Species Name File
Root Taxon(s)
Cross Value
Maximum Network Number 1
Bootstrap Value 0
Gene/Locus Type CDS
Threads Number 1

B

Gene Locus Directory examples/6taxon_300genes_4cores_test
Gene Name File examples/6taxon_gene_name_test.txt
Species Name File examples/6taxon_spe_name_test.txt
Root Taxon(s) A
Cross Value 10
Maximum Network Number 3
Bootstrap Value 0
Gene/Locus Type nonCDS
Threads Number 4

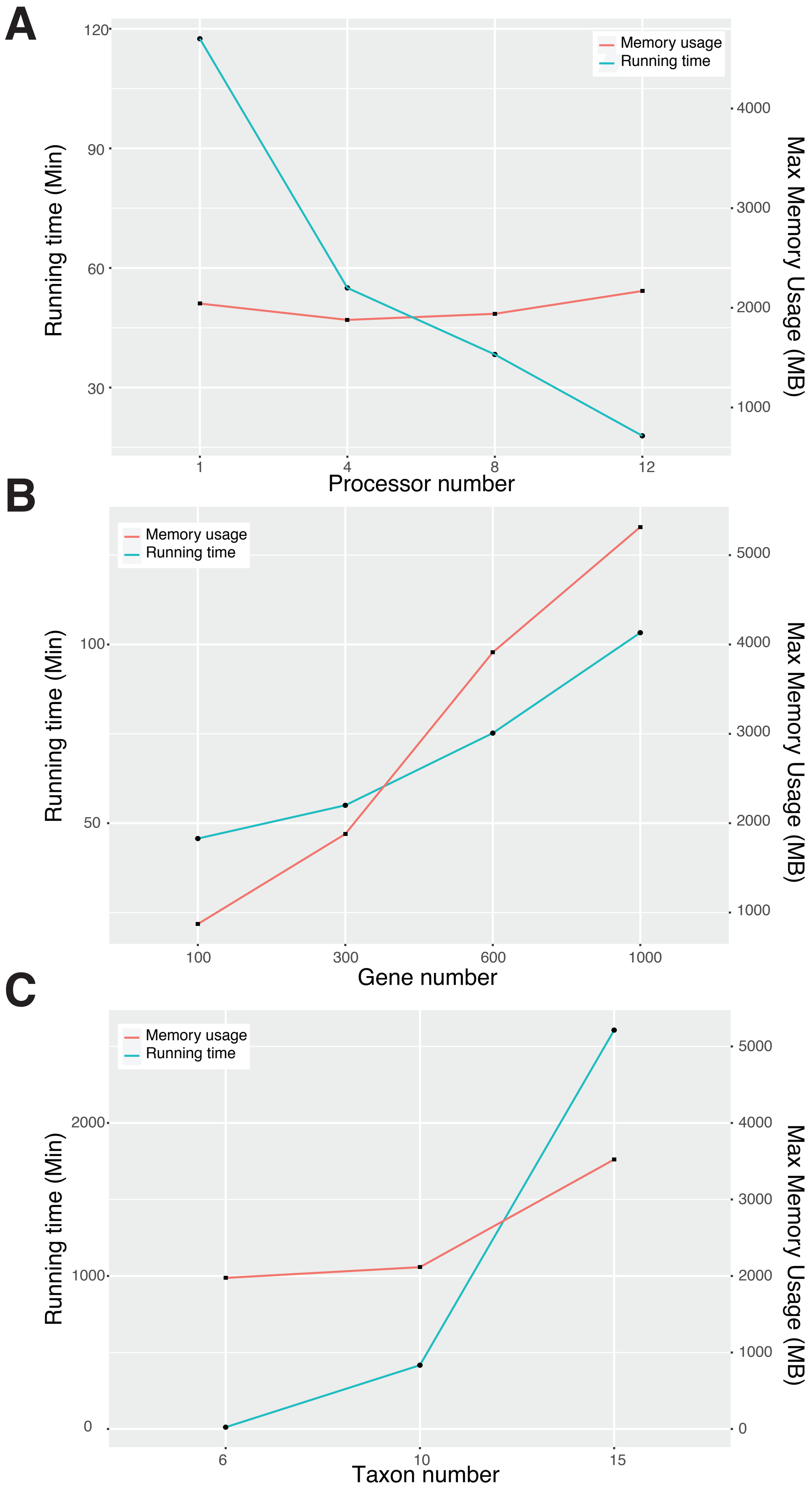
C

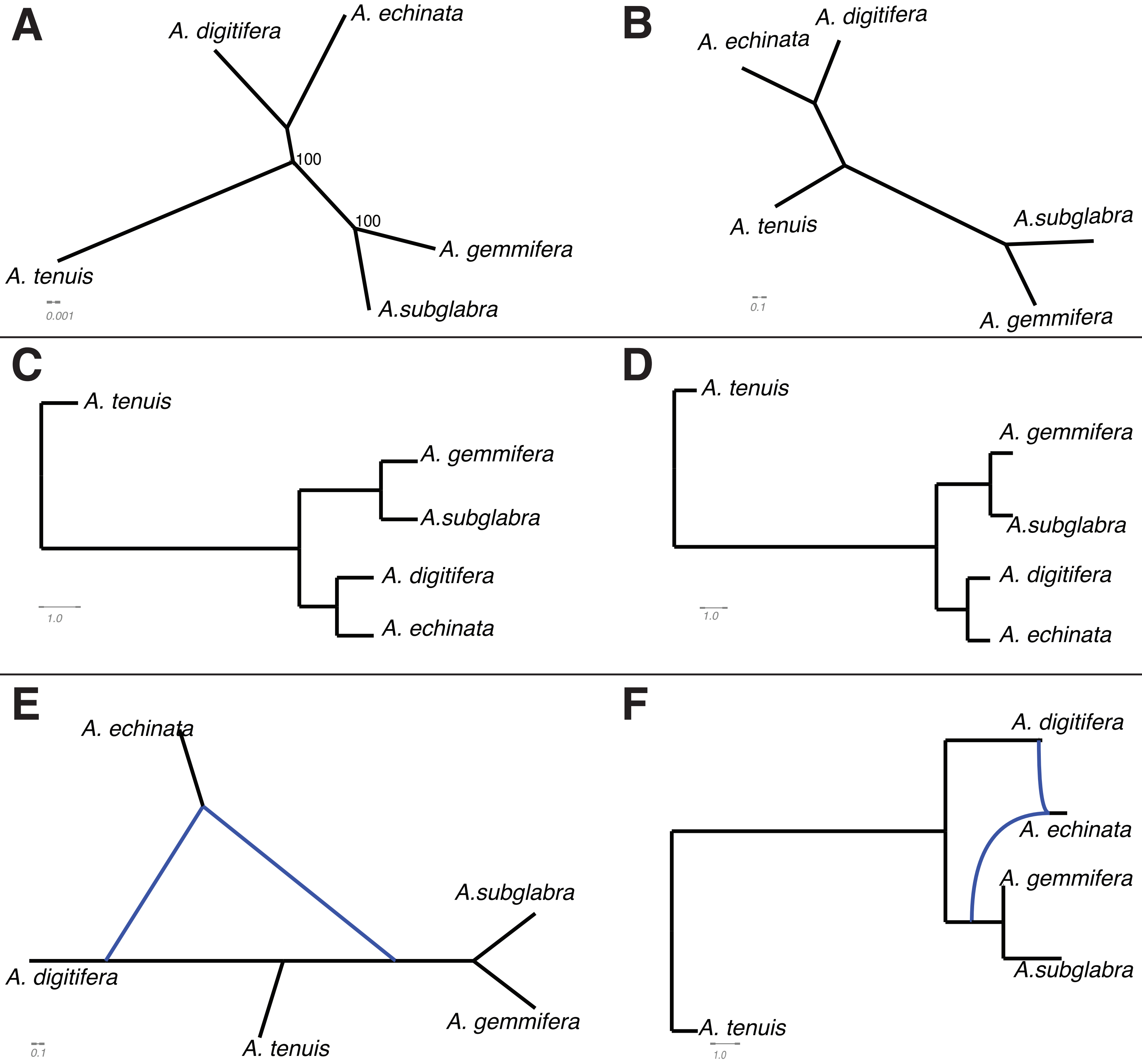
Command Line

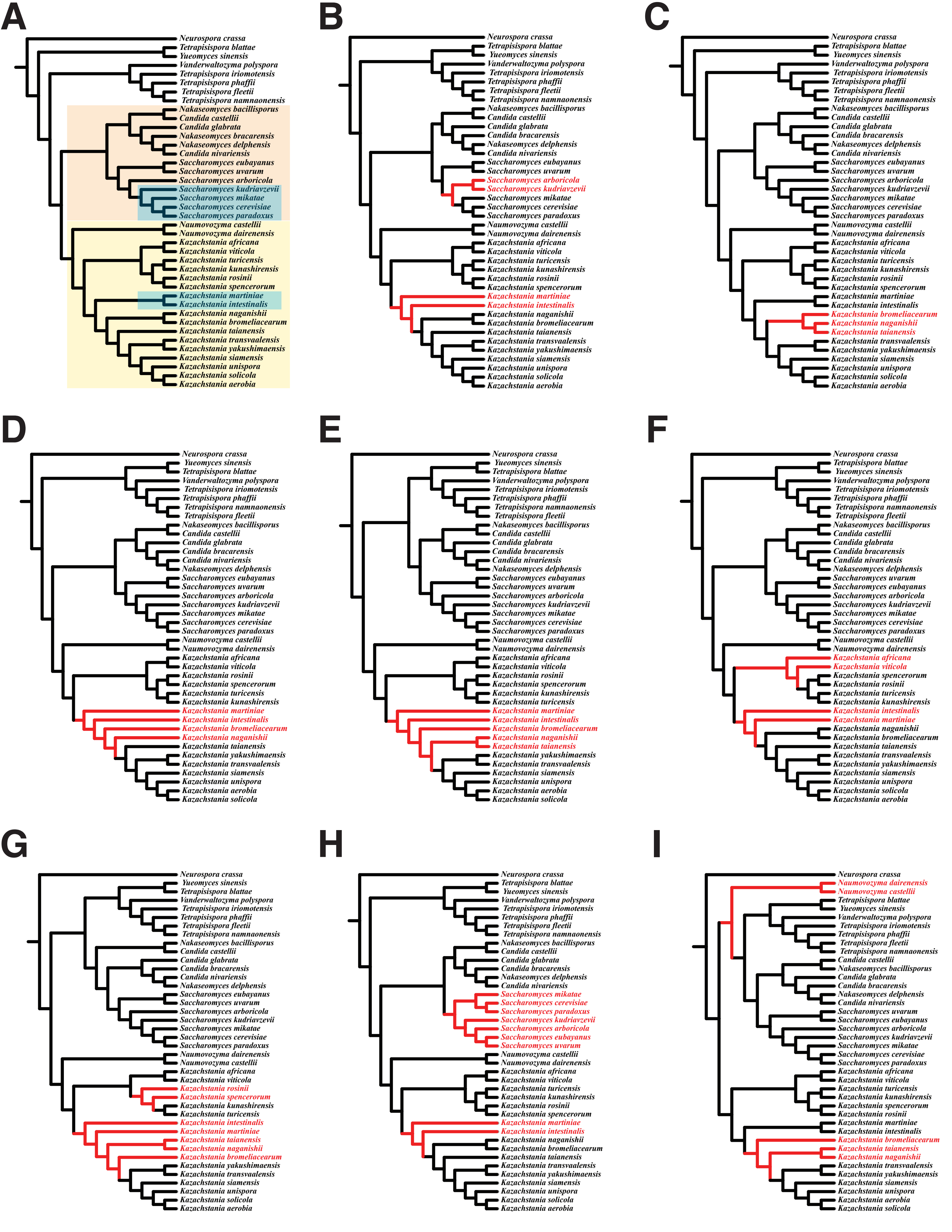
```
python TREEasy.py -d examples/6taxon_300genes_4cores_test -s examples/6taxon_spe
_name_test.txt -g examples/6taxon_gene_name_test.txt -b 0 -r A -n 3 -k 10 -t 4 -
c nonCDS
```

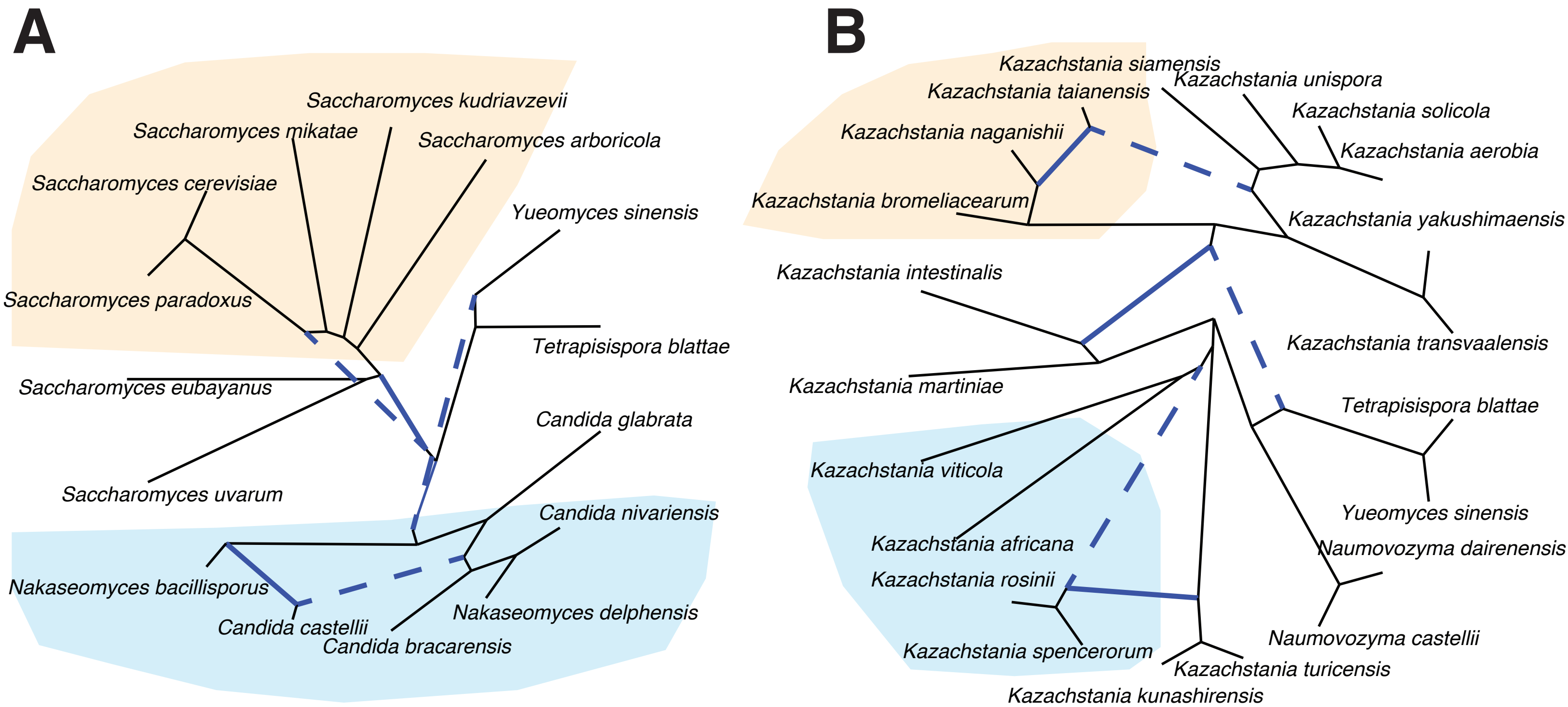
Progress

```
TREEasy started!
ML tree running...
ML tree done
ASTRAL running...
ASTRAL done
MP_EST running...
MP_EST done
STELLS2 running...
STELLS2 done
SNAQ running...
SNAQ done
PHYLONET running...
PHYLONET done
ALL DONE
```


A
120
90
60
30
Running time (Min)
Memory usage
Running time
4000
3000
2000
1000
Max Memory Usage (MB)
1
4
8
12
Processor number
B
100
50
Running time (Min)
Memory usage
Running time
5000
4000
3000
2000
1000
Max Memory Usage (MB)
100
300
600
1000
Gene number
C
2000
1000
0
Running time (Min)
Memory usage
Running time
5000
4000
3000
2000
1000
0
Max Memory Usage (MB)
6
10
15
Taxon number


A
A. echinata
A. digitifera
100
100
A. gemmifera
A. tenuis
A.subglabra
0.001
B
A. digitifera
A. echinata
A. tenuis
A.subglabra
A. gemmifera
0.1
C
A. tenuis
A. gemmifera
A.subglabra
A. digitifera
A. echinata
1.0
D
A. tenuis
A. gemmifera
A.subglabra
A. digitifera
A. echinata
1.0
E
A. echinata
A.subglabra
A. digitifera
A. gemmifera
A. tenuis
0.1
F
A. digitifera
A. echinata
A. gemmifera
A.subglabra
A. tenuis
1.0


A
Neurospora crassa
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Nakaseomyces bracarensis
Nakaseomyces delphensis
Candida nivariensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania naganishii
Kazachstania bromeliacearum
Kazachstania taianensis
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
B
Neurospora crassa
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Nakaseomyces delphensis
Candida nivariensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania naganishii
Kazachstania bromeliacearum
Kazachstania taianensis
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
C
Neurospora crassa
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Nakaseomyces delphensis
Candida nivariensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania bromeliacearum
Kazachstania naganishii
Kazachstania taianensis
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
D
Neurospora crassa
Yueomyces sinensis
Tetrapisispora blattae
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora namnaonensis
Tetrapisispora fleetii
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Candida nivariensis
Nakaseomyces delphensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania bromeliacearum
Kazachstania naganishii
Kazachstania taianensis
Kazachstania yakushimaensis
Kazachstania transvaalensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania aerobia
Kazachstania solicola
E
Neurospora crassa
Yueomyces sinensis
Tetrapisispora blattae
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora namnaonensis
Tetrapisispora fleetii
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Candida nivariensis
Nakaseomyces delphensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania kunashirensis
Kazachstania turicensis
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania bromeliacearum
Kazachstania naganishii
Kazachstania taianensis
Kazachstania yakushimaensis
Kazachstania transvaalensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania aerobia
Kazachstania solicola
F
Neurospora crassa
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora namnaonensis
Tetrapisispora fleetii
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Candida nivariensis
Nakaseomyces delphensis
Saccharomyces uvarum
Saccharomyces eubayanus
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania spencerorum
Kazachstania rosinii
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania intestinalis
Kazachstania martiniae
Kazachstania naganishii
Kazachstania bromeliacearum
Kazachstania taianensis
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
G
Neurospora crassa
Yueomyces sinensis
Tetrapisispora blattae
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Candida nivariensis
Nakaseomyces delphensis
Saccharomyces eubayanus
Saccharomyces uvarum
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Naumovozyma dairenensis
Naumovozyma castellii
Kazachstania africana
Kazachstania viticola
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania kunashirensis
Kazachstania turicensis
Kazachstania intestinalis
Kazachstania martiniae
Kazachstania taianensis
Kazachstania naganishii
Kazachstania bromeliacearum
Kazachstania yakushimaensis
Kazachstania transvaalensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
H
Neurospora crassa
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata
Candida bracarensis
Nakaseomyces delphensis
Candida nivariensis
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Saccharomyces kudriavzevii
Saccharomyces arboricola
Saccharomyces eubayanus
Saccharomyces uvarum
Naumovozyma castellii
Naumovozyma dairenensis
Kazachstania africana
Kazachstania viticola
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania rosinii
Kazachstania spencerorum
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania naganishii
Kazachstania bromeliacearum
Kazachstania taianensis
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania solicola
Kazachstania aerobia
I
Neurospora crassa
Naumovozyma dairenensis
Naumovozyma castellii
Tetrapisispora blattae
Yueomyces sinensis
Vanderwaltozyma polyspora
Tetrapisispora iriomotensis
Tetrapisispora phaffii
Tetrapisispora fleetii
Tetrapisispora namnaonensis
Candida castellii
Nakaseomyces bacillisporus
Candida glabrata
Candida bracarensis
Nakaseomyces delphensis
Candida nivariensis
Saccharomyces uvarum
Saccharomyces eubayanus
Saccharomyces arboricola
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Kazachstania africana
Kazachstania viticola
Kazachstania turicensis
Kazachstania kunashirensis
Kazachstania spencerorum
Kazachstania rosinii
Kazachstania martiniae
Kazachstania intestinalis
Kazachstania bromeliacearum
Kazachstania taianensis
Kazachstania naganishii
Kazachstania transvaalensis
Kazachstania yakushimaensis
Kazachstania siamensis
Kazachstania unispora
Kazachstania aerobia
Kazachstania solicola


A
Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces arboricola
Saccharomyces cerevisiae
Yueomyces sinensis
Saccharomyces paradoxus
Tetrapisispora blattae
Saccharomyces eubayanus
Candida glabrata
Saccharomyces uvarum
Candida nivariensis
Nakaseomyces bacillisporus
Nakaseomyces delphensis
Candida castellii
Candida bracarensis
B
Kazachstania siamensis
Kazachstania unispora
Kazachstania taianensis
Kazachstania solicola
Kazachstania naganishii
Kazachstania aerobia
Kazachstania bromeliacearum
Kazachstania yakushimaensis
Kazachstania intestinalis
Kazachstania transvaalensis
Kazachstania martiniae
Tetrapisispora blattae
Kazachstania viticola
Yueomyces sinensis
Kazachstania africana
Naumovozyma dairenensis
Kazachstania rosinii
Naumovozyma castellii
Kazachstania spencerorum
Kazachstania turicensis
Kazachstania kunashirensis