

# iTree: a high-throughput phylogenomic pipeline

Ahmed Moustafa, Debashish Bhattacharya, and Andrew E. Allen

**Abstract**—Phylogenomics, conventionally defined as the intersection of phylogenetics and genomics, has become a key instrument in a wide spectrum of biological studies, including resolution of complex evolutionary relationships, assignment of taxonomic affiliation, prediction of protein molecular functions, and tracing horizontal gene transfer event. Here, we introduce an open-source phylogenomic pipeline, iTree, which automates the execution of phylogenetic analyses under multithreaded and grid-computing environments, providing a scalable high-throughput platform for performing genome-wide evolutionary analyses. Furthermore, we describe the results of two applications of using iTree: (1) taxonomic assignment of 16S ribosomal RNA sequences from human oral metagenomic samples and (2) detection of horizontal gene transfer in microbial genomes.

## I. INTRODUCTION

Phylogenomics, conventionally defined as the intersection of phylogenetics and genomics [1], has become a key instrument in a broad-spectrum of biological studies, including resolution of complex evolutionary relationships [2], assignment of taxonomic affiliation [3], prediction of protein molecular functions [4], and tracing horizontal gene transfer events [5]. In a typical phylogenetic analysis, the three main steps are (1) searching a reference dataset for homologous sequences for the sequence of interest (i.e., the query), (2) multiple sequence alignment of the query and its homologous sequences, and (3) tree reconstruction using a distance-, parsimony-, or maximum likelihood-based method [6].

A critical factor to infer a reliable phylogeny and meaningful evolutionary history is taxon sampling, i.e., the diversity and density of taxa included in the phylogenetic tree [7, 8]. Alignments with reduced taxon sampling, because of, for example, small reference dataset, too-

Manuscript was received on August 27, 2010. This work was supported partly by an Institutional National Research Service Award (T 32 GM98629) from NIH awarded to AM and two grants from NSF (EF 04-31117) and NIH (EF 04-31117) awarded to DB. The authors would like to acknowledge the valuable suggestions regarding the design of the pipeline made by Adrian Reyes-Prieto of the University of New Brunswick and the critical reading of the manuscript by Trina Norden-Krichmar of J. Craig Venter Institute. The authors would like to thank two anonymous reviewers for their constructive comments.

A. Moustafa was with J. Craig Venter Institute, San Diego, California, USA. He is now with the Department of Biology, American University in Cairo, New Cairo, Egypt (corresponding author; email: amoustafa@aucegypt.edu).

D. Bhattacharya is with the Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, New Jersey, USA (email: bhattacharya@aesop.rutgers.edu)

A. E. Allen is with the Department of Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, California, USA (email: aallen@jcv.i.org)

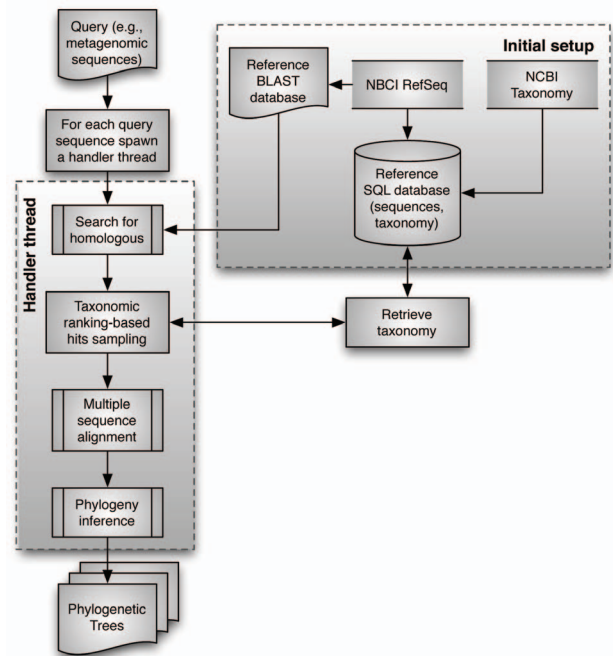


Fig. 1. Flowchart for the major steps of iTree. Key steps in the pipeline are indicated in the diagram into two key subgroups, “Initial setup” and “Handler thread”. “Initial setup” summarizes the steps required to prepare the SQL and BLAST databases that represent the backend of the pipeline. “Handler thread” is the central processing unit of the pipeline, where individual sequences are processed, starting with searching for homologous sequences in the database using BLAST and concluding with inferring the phylogeny of the query sequence.

stringent search criteria, or the nature of the homolog search strategy, often lead to inaccurate phylogenies and consequently misguided or incorrect conclusions. To circumvent the negative impact of taxon under-sampling, an intuitive approach would be to search for homologous sequences in a broad reference dataset and to employ permissive search criteria. However, as the number of sequences identified by the search step becomes large, the complexity of computing the alignments and inferring the phylogenetic trees becomes intractably intensive. To tackle this issue, we introduce a phylogenomic pipeline “iTree” that implements a simple method that increases taxon diversity at a non-dense sampling. We collect only the top hits that belong to different taxonomic ranks, resulting in a comprehensive coverage of the taxa that to be included in subsequent analyses, while maintaining a reasonable alignment size. Thus, the generated alignment can be used to infer an accurate phylogeny under practical computing power and time requirements. We employed iTree in several studies and in some cases we identified novel and significant

findings (e.g., [5] and [9]) that were not apparent using other approaches and pipelines. Furthermore, iTree also addresses several troublesome issues with the other phylogenomic pipelines such as difficulty of installation and setting up, scalability in both dimensions, number of queries and size of the reference dataset, and hard-coded features such as alignments and tree constructions algorithms.

iTree runs as a multithreaded application to exploit the multi-core architecture in modern processors, allowing simultaneous processing for multiple query sequences in a single-computing environment. In addition, a single-threaded version of iTree is available for grid-computing environments.

In this report, we discuss the implementation details of the pipeline and provide two case studies using iTree. The first is identification of taxonomic affiliation of metagenomic data from the human oral microbiome. The second is prediction of endosymbiotically-transferred genes in the nuclear genome of marine picoeukaryotes.

## II. IMPLEMENTATION

There are two components in the implementation of iTree: a relational database (RDB) and Perl programs. The iTree RDB holds a replicate of the BLAST database, which is searched for putative homologs of the query sequences during the first step in the pipeline (Fig. 1). Unlike pipelines that load the entire BLAST database sequences in the random-access memory (RAM) (e.g., PhyloGenie [10]), replicating the BLAST sequences into a RDB allows iTree to be independent of the size of the BLAST database by eliminating the requirement for allocating extensive RAM. The RDB also retains the taxonomic classification of the reference sequences, based on the National Center for Biotechnology Information (NCBI) taxonomy [11]. The taxonomy component in iTree partially inherits the taxonomy entities of the BioSQL schema [12], which represents the hierarchical structure of the taxonomic relationships in a nested sets model [13]. This approach maintains the taxonomic hierarchies and provides an efficient mechanism for traversing the taxonomic ranks. The main Perl program runs in single- and multi-threaded fashions. In the multithreaded version, a processing thread is spawned for each entry in the query set. The processing thread performs the steps of the phylogenetic reconstruction (1) searching for homologs using BLAST, (2) aligning a query and its homologs, and (3) inferring a phylogenetic tree.

In our implementation, after identifying the hits based on the BLAST results and prior to the multiple sequence alignment, we determine the full taxonomic ranking (i.e., kingdom, phylum, class, order, family, genus, and species) of each hit. According to a configurable set of parameters, we then retrieve only the top subset of the hits that belong to each of the taxonomic levels. Thus, we collect homologous sequences from most of the taxonomic levels represented in the BLAST hits that account for the taxonomic diversity in

the original full set without having to include all BLAST results. This strategy maximizes taxon-coverage in respect of the diversity, while avoiding computationally intractable taxon sampling.

To assess the overall performance of the pipeline running as a multithreaded process, we analyzed the 182 proteins encoded by the smallest known bacterial genome, the 160 Kb genome of the  $\gamma$ -proteobacterium insect endosymbiont *Carsonella ruddii* [14]. Here, the performance of a single thread was compared to the performance of ten additional runs in which the number of threads was increased from 10 to 100 by increments of 10 threads. Benchmarking was conducted on a dedicated two-2.4 GHz quad-core (8-core) Intel Xeon “Westmere” Mac Pro (Fig. 2). As a single-thread, iTree finished processing the 182 proteins in >5 hours. After increasing the number of threads to 10, iTree finished in less than an hour with a dramatic improvement of more than 80% over the single-thread run. Furthermore, by increasing the number of threads to 20, there was an additional 10% processing time improvement compared to the 10-thread run. However, further increases in the concurrent threads from 30 to 100 lowered overall performance almost linearly from about 8% to 100%, respectively, compared to the 20-thread run. This is likely explained by the increasing overhead cost of managing the shared resources on the computer and competition among the concurrent threads over these resources.

To prepare iTree for the public release, we populated a MySQL database (i.e., the iTree RDB) with NCBI RefSeq Release 41 [15] and associated taxonomic information. We generated a FASTA-formatted database of the RefSeq sequences after processing the sequences and renaming them according to a pattern that specifies the GenBank accession numbers and taxonomic classifications (i.e., supergroup,

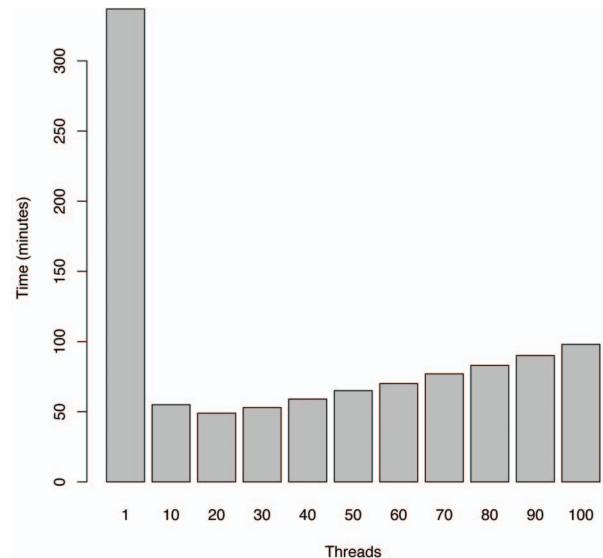


Fig. 2. Phylogenomic processing time of the genome of *Carsonella ruddii*. The processing time of the 182 proteins encoded by the smallest bacterial genome, *Carsonella ruddii*, on the y-axis as a function of the number of concurrent handler threads on the x-axis.

genus, and species). iTree is supported on Linux, UNIX, and Mac OS. It was tested successfully on standalone workstations, running Mac OS X and Ubuntu and on a high-performance computing cluster, running Sun Grid Engine, in multi- and single-threaded modes, respectively. The MySQL and FASTA databases and the Perl programs are published as open-source at <http://itree.sourceforge.net> under the terms of the GNU General Public License (GPL).

### III. APPLICATIONS

#### A. Assignment of taxonomic affiliation of human oral metagenomic-derived 16S ribosomal RNA sequences

The rapid advancements in sequencing technologies and the accumulation of massive amounts of nucleic acid sequences that provide “deep” coverage of samples under investigation have led the emergence of new venues of research and applications. Of particular importance is the application of next-generation sequencing methods to study environmental samples in medical [16], ecological [17], and energy [18] studies. Upon sequencing environmental DNA, the first issue to be addressed is the “identities” of the inhabitants of the environment. The simplest and most straightforward method is to search a database of genome sequences with “known” identities (i.e., references) for homologs to the environmental “unknown” sequences using BLAST [19] and to transfer the identities of the best hits to the queries. However, it is known that the BLAST “best-hit” is not necessarily the nearest phylogenetic neighbor [20] and; therefore, this approach could lead to misidentification of the query sequences. This is a major issue, particularly in medical studies where it is critical to accurately identify the taxonomic composition of the environmental samples in order to classify and associate taxonomic composition to the conditions of health and disease.

To illustrate the application of iTree in metagenomic studies and to compare it to the “best-hit” approach, we used the available 16S ribosomal RNA (rRNA) sequences from the Human Oral Microbiome Database (HOMD) [21] that have already been assigned taxonomic identities (we identified 1,500 “fully” characterized sequences out of a total 1,647 in the reference HOMD) as a positive control to compare the accuracy of our pipeline against the “best-hit” approach solely using BLAST. The target database in both cases (i.e., best-hit and iTree), was a subset of the SILVA rRNA database [22] after removing all sequences with any form of uncertainty at any taxonomic rank; i.e., sequences that were labeled as “environmental”, “uncultured”, or “unidentified” were excluded from the original SILVA database to prepare a taxonomically well-annotated target database.

By using the 16S rRNA human oral sequences as queries and searching SILVA with “blastn”, the taxonomy of the best hit for each query was transferred directly to the queries. Next, using iTree, the 16S rRNA human oral sequences were processed against the same database to generate a phylogenetic tree for each sequence. Thereafter,

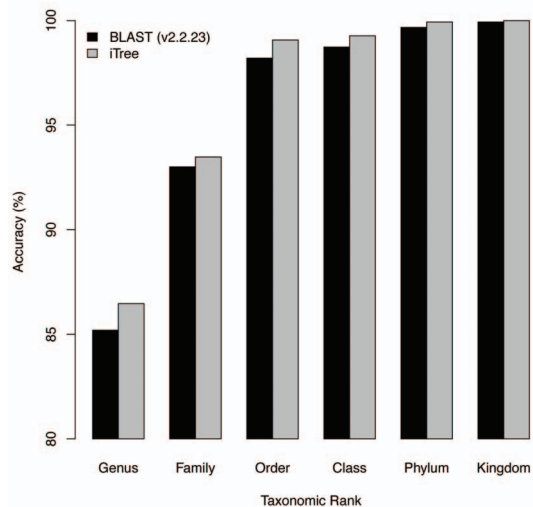


Fig. 3. Accuracy of taxonomy assignment. The accuracy (y-axis) in determining the taxonomic origin of metagenomic-derived rRNA data using iTree and BLAST “best-hit” approach at the different taxonomic rankings (x-axis).

RAxML [23] maximum likelihood (ML) inferred trees were searched for the nearest phylogenetic neighbors from the query sequences using PhyloSort [24]. Thus, the taxonomies of the nearest neighbors were used to predict the taxonomy of the queries. By comparing the best-hit and iTree assignments to the control HOMD taxonomic information at the ranks from genus to kingdom (Fig. 3), it is apparent that iTree provides more accurate and sensitive assignments for the 16S rRNA sequences throughout all of the taxonomic ranks with the largest difference at the genus level (about 1.5% better accuracy). Although the largest advantage of iTree over BLAST was at the genus level, given the large data volumes generated by next-generation sequencing technologies iTree is potentially to correctly call the taxonomy of about thousands of reads that are likely to be miscalled by the BLAST best-hit approach.

#### B. Identification of endosymbiotically-transferred genes in the nuclear genome of a picoeukaryote

Horizontal gene transfer (HGT) has been widely recognized as a major factor in shaping genomes and the tree of life [25], [26]. Therefore, it has become essential to estimate the extent of horizontally transferred genes, especially in the genomes of microbial organisms, and their functional contributions to the biology of the host cells. Phylogenetic-based approaches have been typically used to search for potential HGT scenarios in all domains of life [5], [27], [28]. Briefly, a phylogenetic “gene”-based tree is reconstructed then its topology is compared to the canonical “species” tree and cases of incongruence are interpreted as evidence of HGT [25].

Endosymbiotic gene transfer (EGT) is a specific case of HGT, which results from a long-term association between a host cell and an endosymbiont [29]. During such a host-endosymbiont interaction, genetic materials can transfer from the endosymbiont to the nucleus of the host. During

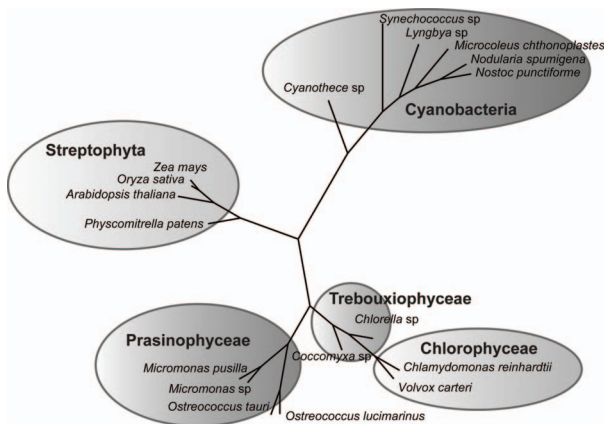


Fig. 4. Maximum likelihood phylogenetic tree of RuBisCO activase. An unrooted radial phylogenetic tree of RuBisCO activase as a sample of the trees generated by iTree while searching for cases of endosymbiotic gene transfer from the cyanobacterial endosymbiont to the nuclear genome of the ancestor of the photosynthetic eukaryotes during the process of “primary endosymbiosis”. The tree shows that RuBisCO activase is strictly encoded by the genomes of photosynthetic organisms where it was originated in cyanobacteria (photosynthetic bacteria) then was transferred to plants and green algae (photosynthetic eukaryotes; Prasinophyceae, Trebouxiophyceae, Chlorophyceae, and Streptophyta) via endosymbiotic gene transfer.

this process, deleterious (with respect to host fitness) transfers are lost, whereas advantageous transfers are fixed and become integrated into host biology. During the early evolution of the eukaryotic cell, two milestone endosymbiosis events took place that left profound signatures on almost all forms of life on our planet. The first was the endosymbiosis of an  $\alpha$ -proteobacterium-like cell >2 billion years ago (Ga) by an Archaea-like host. Over time, the  $\alpha$ -proteobacterial symbiont evolved into the energy producing organelle, the “mitochondrion”, in contemporary eukaryotic cells [30]. The second was the engulfment of a cyanobacterium (i.e., a photosynthetic bacterium) by a heterotrophic eukaryotic host about 1.5 Ga. This event introduced and established photosynthesis in the eukaryotic domain through the evolution of the cyanobacterium symbiont into the photosynthetic organelle, the “plastid” [31], [32].

In the case of the plastid endosymbiosis, we showed in a previous study that about 4% of the nuclear genome of the freshwater unicellular green alga *Chlamydomonas* was contributed by genes that were transferred *via* EGT from the cyanobacterium endosymbiont to the nuclear genome of the eukaryotic host [24]. Here we compare our taxonomically-“filtered” top-hits method, implemented in iTree, against the typical “straight” top-hits approach in identifying homologs for inferring phylogenies. This approach is used to search for cyanobacterial genes in the marine pico-eukaryotic unicellular green alga *Micromonas* that contains a highly reduced nuclear genome [33]. Using iTree, we analyzed the complete genome of *Micromonas* against a comprehensive genome database that was assembled from RefSeq, the Joint Genome Institute microbial complete genomes, and partial

genomes from expressed sequence tag (EST) libraries (for evolutionarily, important organisms that currently do not have publicly available complete genome data; e.g., red algae and dinoflagellates).

First, we enabled the “filtered” top-hits feature and searched the generated PhyML [34] phylogenetic trees for clades that support the monophyly of *Micromonas* (optionally, along with other photosynthetic eukaryotes) and cyanobacteria. At a bootstrap support cut-off of  $\geq 75\%$ , we identified 418 nuclear-encoded proteins of cyanobacterial origin, representing 4.1% of the genome of *Micromonas*. This result agrees with our previous estimate using a green alga with a typical genome size [24]. However, given the reduced nature of the *Micromonas* genome [33], our data suggest an interesting aspect of EGT that, unlike some of the “native” genes that were lost during the process of genome reduction, endosymbiotically transferred genes, appear to have survived the process of gene loss, perhaps because of the key functions they provide to the host.

Second, we disabled the “filtered” feature and allowed the pipeline to collect simply the top “100” hits for the queries and use them in the subsequent steps. We reprocessed the *Micromonas* genome as described above. We found that 40 genes of the 418 cyanobacterial genes (i.e., 9.6%) could not be recovered under the “straight” (100) top-hits scheme. By manually inspecting the trees of the missing genes, we found that the main reasons why these genes did not satisfy the monophyly and confidence criteria were either that they were highly conserved among eukaryotes, especially among the photosynthetic lineages or that there were multiple copies of these genes in the organisms. Both cases caused the alignments of these missing genes to be filled with eukaryotic sequences and subsequently the phylogenetic trees did not include any bacterial or, specifically, cyanobacterial sequences. Among the missing genes are two genes that were shown to be *bona fide* cyanobacterial genes in the nuclear genomes of photosynthetic eukaryotes. The first is RuBisCO activase (RCA), a Calvin-Cycle enzyme localized in the plastid thylakoid membrane with ATP binding activity (Fig. 4), which was reported to be of cyanobacterial origin [35], agreeing with our results with the “filtered” feature enabled. The second is a carbon fixation enzyme, nuclear-encoded plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH), which was also previously shown to be of cyanobacterial origin and transferred to the nuclear genome of eukaryotes through EGT [36]. Thus, our “filtered” top-hits approach has a true-positive rate higher than the typical “straight” top-hits approach, in particular in automated analyses executed against a comprehensive genome dataset.

#### IV. CONCLUSIONS

We designed and developed iTree, an open-source, easy-to-use, high-throughput, phylogenomic pipeline. We introduced our “filtered” top-hits approach, which achieves a balance between taxon diversity and taxon density, in

particular when conducting analyses against extensive genome datasets. We also showed that our pipeline runs as a multithreaded-application in a standalone multi-core environment and a single-threaded application in a grid-computing environment. Finally, we provided practical examples using iTree in active areas of investigation in evolutionary and genome biology, metagenomics and horizontal gene transfer.

## REFERENCES

- [1] J. A. Eisen, "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis," *Genome Res*, vol. 8, no. 3, pp. 163-7, Mar, 1998.
- [2] C. W. Dunn, A. Hejnal, D. Q. Matus *et al.*, "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, pp. 745-9, Apr 10, 2008.
- [3] D. Wu, P. Hugenholtz, K. Mavromatis *et al.*, "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea," *Nature*, vol. 462, no. 7276, pp. 1056-60, Dec 24, 2009.
- [4] K. Sjolander, "Phylogenomic inference of protein molecular function: advances and challenges," *Bioinformatics*, vol. 20, no. 2, pp. 170-9, Jan 22, 2004.
- [5] A. Moustafa, B. Beszteri, U. G. Maier *et al.*, "Genomic footprints of a cryptic plastid endosymbiosis in diatoms," *Science*, vol. 324, no. 5935, pp. 1724-6, Jun 26, 2009.
- [6] C. J. Harrison, and J. A. Langdale, "A step by step guide to phylogeny reconstruction," *Plant J*, vol. 45, no. 4, pp. 561-72, Feb, 2006.
- [7] T. A. Heath, S. M. Hedtke, and D. M. Hillis, "Taxon sampling and the accuracy of phylogenetic analyses," *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 239-257, May, 2008.
- [8] D. J. Zwickl, and D. M. Hillis, "Increased taxon sampling greatly reduces phylogenetic error," *Syst Biol*, vol. 51, no. 4, pp. 588-98, Aug, 2002.
- [9] A. Reyes-Prieto, H. S. Yoon, A. Moustafa *et al.*, "Differential Gene Retention in Plastids of Common Recent Origin," *Molecular Biology and Evolution*, vol. 27, no. 7, pp. 1530-1537, Jul, 2010.
- [10] T. Frickey, and A. N. Lupas, "PhyloGenie: automated phylome generation and analysis," *Nucleic Acids Res*, vol. 32, no. 17, pp. 5231-8, 2004.
- [11] E. W. Sayers, T. Barrett, D. A. Benson *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D5-15, Jan, 2009.
- [12] J. E. Stajich, and H. Lapp, "Open source tools and toolkits for bioinformatics: significance, and where are we?," *Brief Bioinform*, vol. 7, no. 3, pp. 287-96, Sep, 2006.
- [13] J. Celko, *Joe Celko's Trees and hierarchies in SQL for smarties*, Amsterdam ; Boston: Morgan Kaufmann, 2004.
- [14] A. Nakabachi, A. Yamashita, H. Toh *et al.*, "The 160-kilobase genome of the bacterial endosymbiont Carsonella," *Science*, vol. 314, no. 5797, pp. 267, Oct 13, 2006.
- [15] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D61-5, Jan, 2007.
- [16] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald *et al.*, "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, pp. 1027-31, Dec 21, 2006.
- [17] D. B. Rusch, A. L. Halpern, G. Sutton *et al.*, "The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific," *PLoS Biol*, vol. 5, no. 3, pp. e77, Mar, 2007.
- [18] F. Warnecke, P. Luginbuhl, N. Ivanova *et al.*, "Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite," *Nature*, vol. 450, no. 7169, pp. 560-5, Nov 22, 2007.
- [19] S. F. Altschul, T. L. Madden, A. A. Schaffer *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389-402, Sep 1, 1997.
- [20] L. B. Koski, and G. B. Golding, "The closest BLAST hit is often not the nearest neighbor," *J Mol Evol*, vol. 52, no. 6, pp. 540-2, Jun, 2001.
- [21] T. Chen, W. H. Yu, J. Izard *et al.*, "The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information," *Database (Oxford)*, vol. 2010, pp. baq013.
- [22] E. Pruesse, C. Quast, K. Knittel *et al.*, "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Res*, vol. 35, no. 21, pp. 7188-96, 2007.
- [23] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688-90, Nov 1, 2006.
- [24] A. Moustafa, and D. Bhattacharya, "PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*," *BMC Evol Biol*, vol. 8, pp. 6, 2008.
- [25] J. P. Gogarten, and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nat Rev Microbiol*, vol. 3, no. 9, pp. 679-87, Sep, 2005.
- [26] R. Jain, M. C. Rivera, and J. A. Lake, "Horizontal gene transfer among genomes: the complexity hypothesis," *Proc Natl Acad Sci U S A*, vol. 96, no. 7, pp. 3801-6, Mar 30, 1999.
- [27] F. Burki, K. Shalchian-Tabrizi, M. Minge *et al.*, "Phylogenomics reshuffles the eukaryotic supergroups," *PLoS One*, vol. 2, no. 8, pp. e790, 2007.
- [28] V. Daubin, M. Gouy, and G. Perriere, "A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history," *Genome Res*, vol. 12, no. 7, pp. 1080-90, Jul, 2002.
- [29] J. N. Timmis, M. A. Ayliffe, C. Y. Huang *et al.*, "Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes," *Nat Rev Genet*, vol. 5, no. 2, pp. 123-35, Feb, 2004.
- [30] M. W. Gray, G. Burger, and B. F. Lang, "Mitochondrial evolution," *Science*, vol. 283, no. 5407, pp. 1476-81, Mar 5, 1999.
- [31] D. Bhattacharya, and L. Medlin, "Algal phylogeny and the origin of land plants," *Plant Physiology*, vol. 116, no. 1, pp. 9-15, Jan, 1998.
- [32] S. D. Dyall, M. T. Brown, and P. J. Johnson, "Ancient invasions: From endosymbionts to organelles," *Science*, vol. 304, no. 5668, pp. 253-257, Apr 9, 2004.
- [33] A. Z. Worden, J. H. Lee, T. Mock *et al.*, "Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*," *Science*, vol. 324, no. 5924, pp. 268-72, Apr 10, 2009.
- [34] S. Guindon, F. Delsuc, J. F. Dufayard *et al.*, "Estimating maximum likelihood phylogenies with PhyML," *Methods Mol Biol*, vol. 537, pp. 113-37, 2009.
- [35] A. Reyes-Prieto, and D. Bhattacharya, "Phylogeny of Calvin cycle enzymes supports Plantae monophyly," *Mol Phylogenet Evol*, vol. 45, no. 1, pp. 384-91, Oct, 2007.
- [36] W. Martin, H. Brinkmann, C. Savonna *et al.*, "Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes," *Proc Natl Acad Sci U S A*, vol. 90, no. 18, pp. 8692-6, Sep 15, 1993.