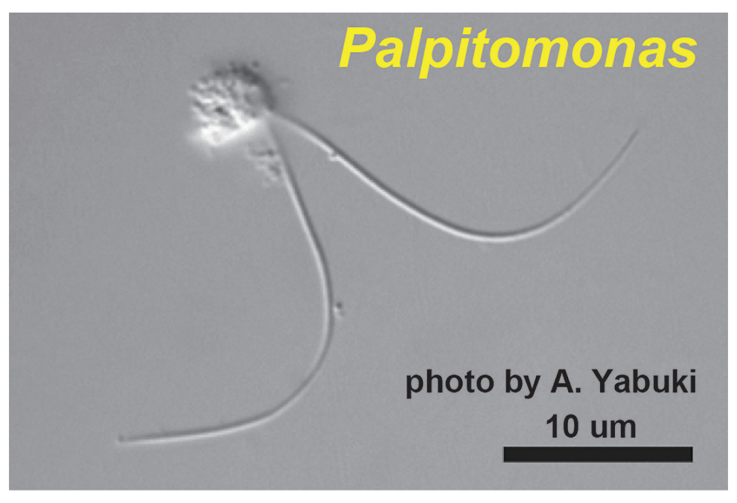
HPC for Phylogenetic Tree Inference

Phylogenetic Inference on the Universal Tree of Life

The central focus of our research is to gain insights into the origin and early evolution of eukaryotes. This requires i) to reconstruct the Universal Tree of Life including diverse organisms on Earth based on large-scale sequence datasets, and ii) to apply parameter-rich models to obtain appropriate evaluation of complex evolution in a disparate range of species.

Phylogeneitc position of a newly identified protist, *Palpitomonas bilix*, inferred from a large-scale multigene dataset

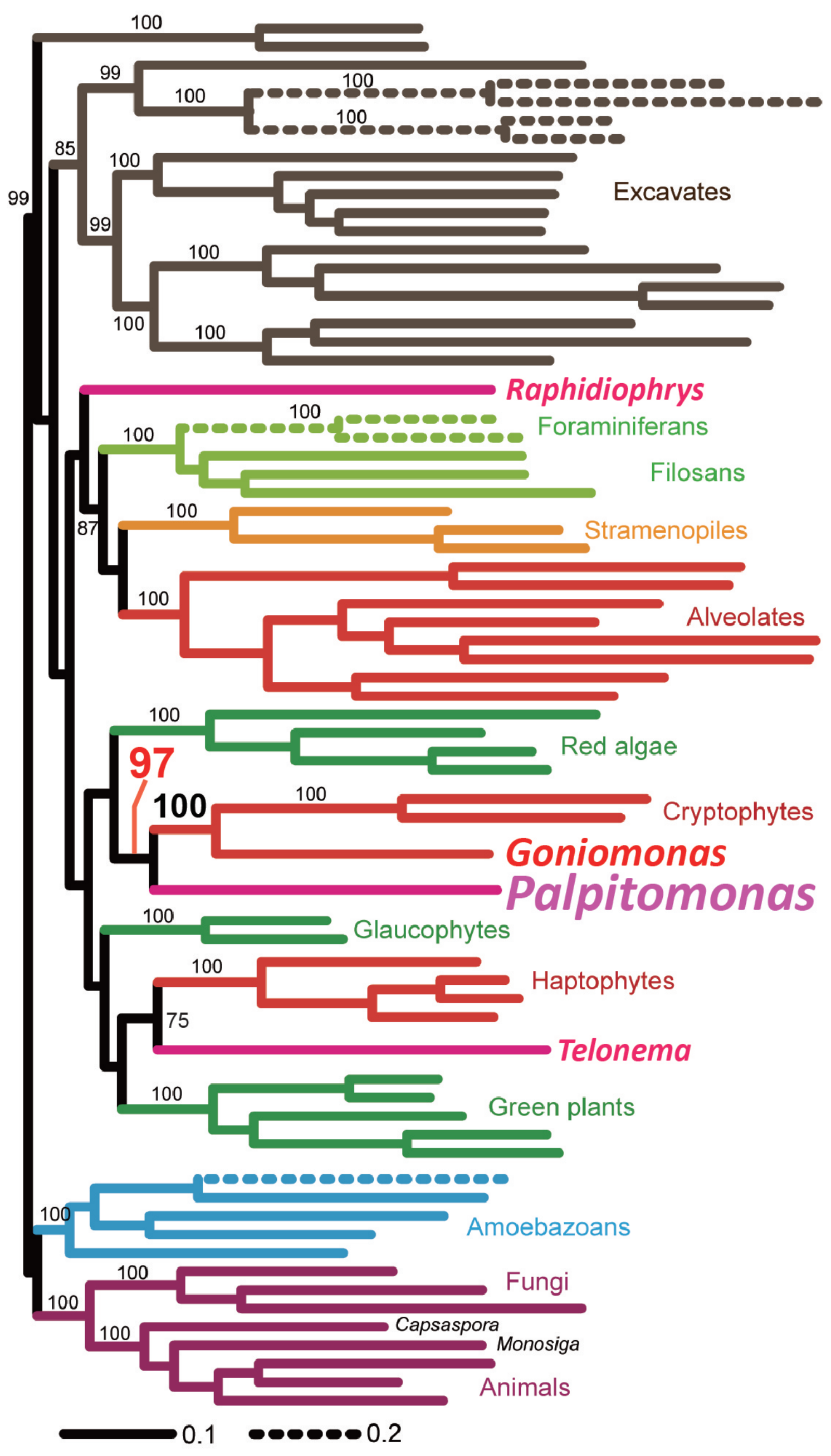


Introduction: ***Palpitomonas bilix*** is a recently identified single-cell protist and considered to provide key information regarding the early evolution of major groups of photosynthetic eukaryotes. However, the phylogenetic position of ***Palpitomonas*** has yet to be resolved with confidence.

Results: We conducted massive expressed sequence tag analyses on ***Palpitomonas bilix*** using 454 pyrosequencing technology, and prepared a 159-gene dataset including 69 taxa and 42,744 amino acid positions. In the 159-gene analyses, ***Palpitomonas*** was robustly placed at the ancestral position of the clade of cryptophytes and ***Gonionmonas***. Interestingly, our result showed that two photosynthetic groups, haptophytes and cryptophytes were NOT monophyletic, which was contradictory to the hypothesis proposed prior to our analyses.

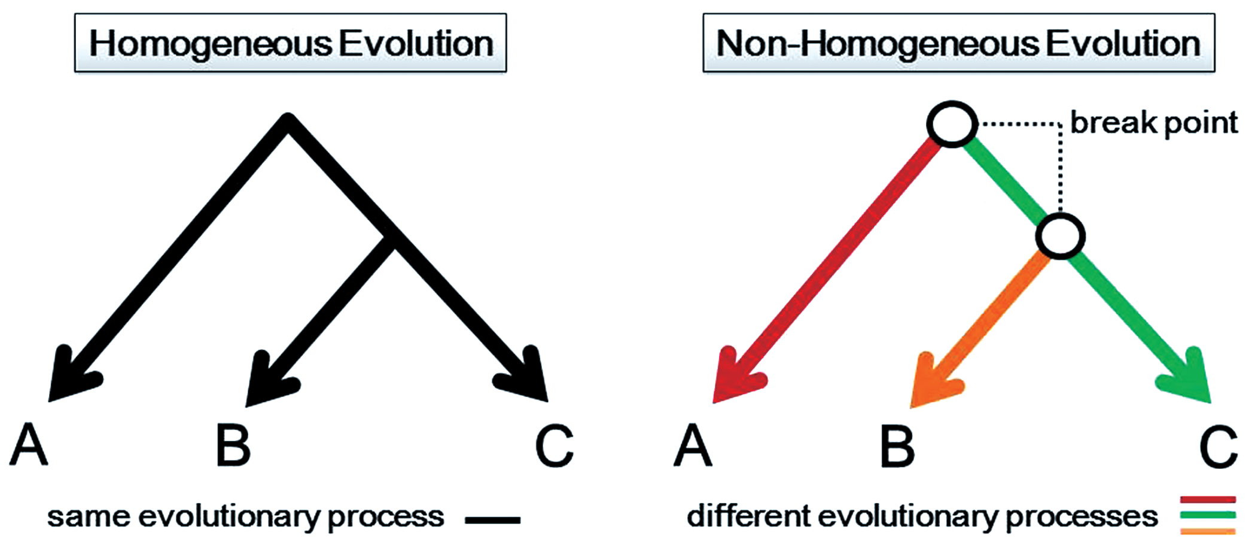
Discussion: Our 159-gene analyses successfully provided the first clues for the putative phylogenetic position of ***Palpitomonas***. In addition, our phylogenetic analysis disfavored the hypothesis assuming the intimate relationship between cryptophytes and haptophytes.

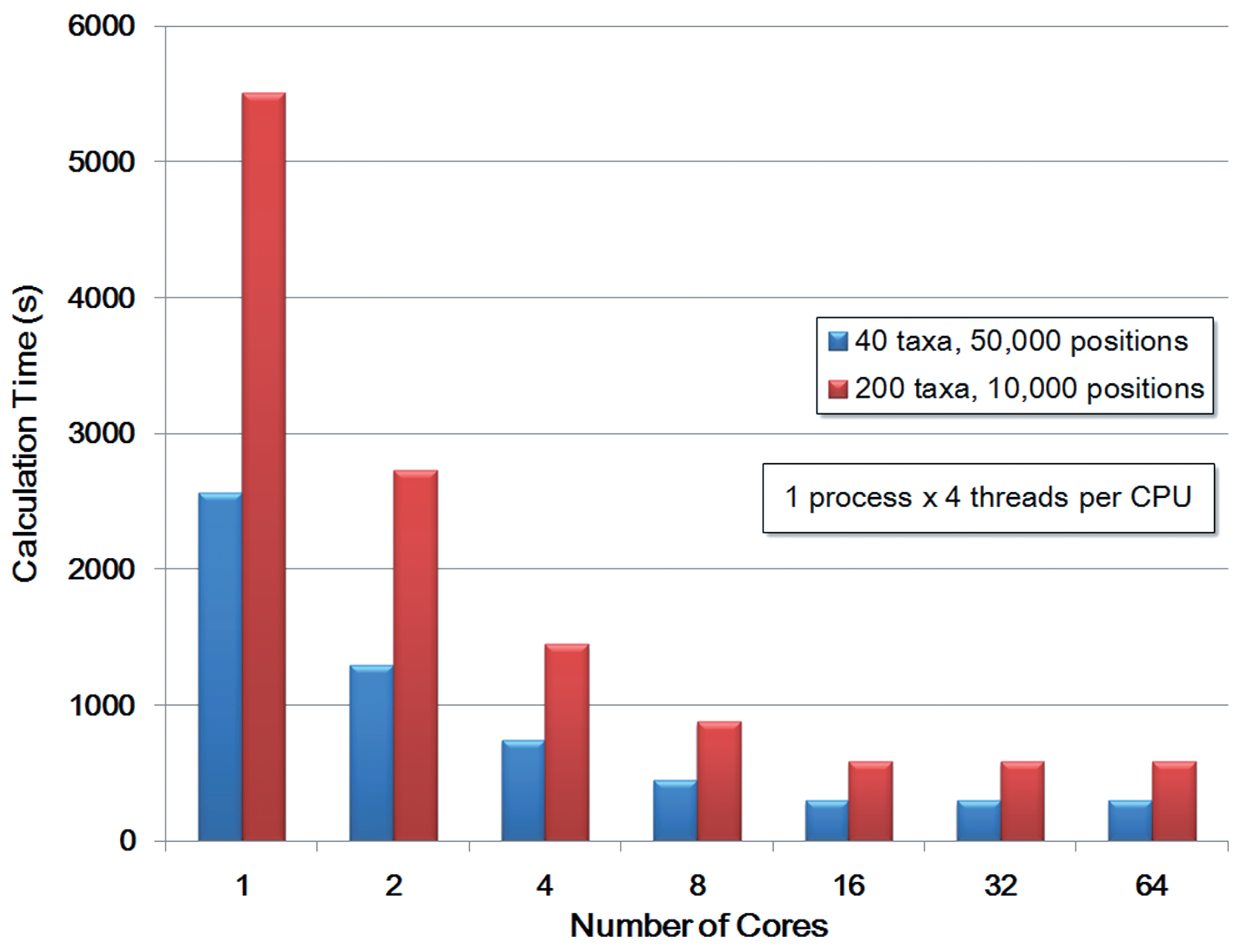
Methods: The 159-gene dataset was analyzed by RAxML with LG+Γ+F model. Maximum-likelihood bootstrap values were presented on nodes. Branch lengths indicate the average number of substitutions per position.



Reference: Yabuki et al. "***Palpitomonas bilix*** gen. et sp. nov.: A novel deep-branching Heterotroph possibly related to Archaeplastida or Hacrobia." Protist. 2010. 161:523-538.

Parallelization of the phylogenetic inference with the non-homogeneous substitution models


Homogeneous Evolution
A
B
C
same evolutionary process
Non-Homogeneous Evolution
break point
A
B
C
different evolutionary processes



Introduction: The nucleotide and amino acid sequences in distantly related species certainly evolve under different evolutionary processes (non-homogeneous evolution). Thus, to infer phylogenetic trees from real-world sequence data, non-homogeneous (NH) models, in which allocate different model parameters on each node of the tree to evaluate, are more realistic than homogeneous models which enforce a single set of model parameters to all branches in the tree. However, the analyses with NH models can be computationally intense, as enormous amount of model parameters need to be optimized.

Results: We applied two approaches for parallel computing, OpenMP and MPI, for estimating the parameters in the maximum-likelihood (ML) method with a NH model. This HYBRID parallel programming enabled to finish the likelihood calculation of single phylogenetic tree in 8 ~ 9 times faster than the control. However, the analyses using more than 16 cores showed no contribution for the likelihood calculation.

Discussion: In this study, we only parallelized the parameter estimation of the NH model on a single tree. However, ML phylogenetic inferences in general require the likelihood calculations of a numerous number of tree topologies during heuristic tree search. Therefore, we need to achieve the higher level of parallel computing on the likelihood calculation of multiple trees than that presented here.

Methods: We simulated two nucleotide sequence data, i) 40 taxa and 50,000 positions, and ii) 200 taxa and 10,000 positions, based on different model trees. Each sequence dataset was subjected to the likelihood calculation of the corresponding model tree with the NH model. Data analyses on multiple processors were run on T2K Tsukuba (~ 4 nodes/64 cores; each node is composed of 4 quad-core CPUs).

Reference: Galtier and Gouy. "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis." Molecular Biology and Evolution. 1998. 15:871-9.