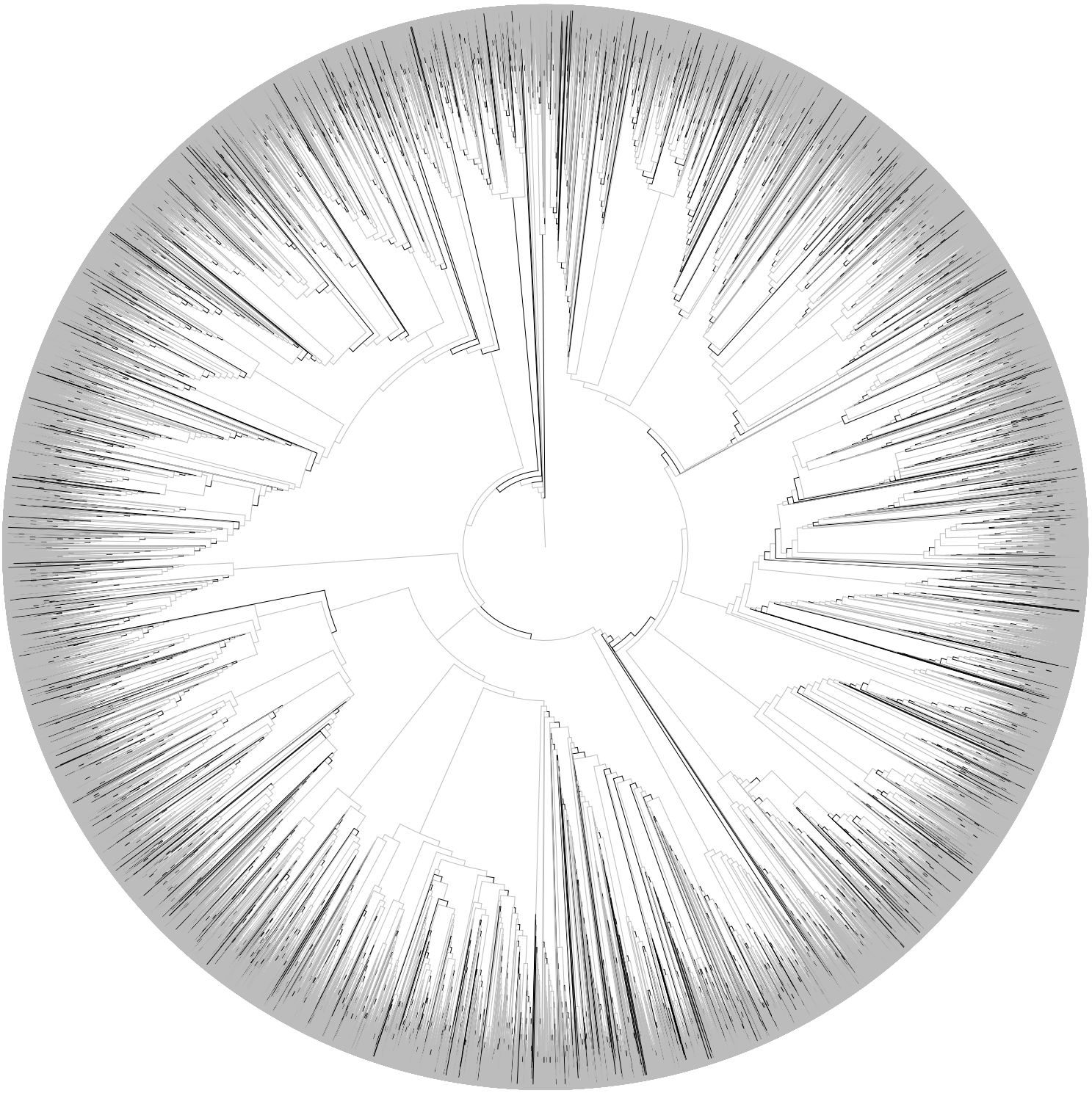
Smith, Stephen A., Jeremy M. Beaulieu, Alexandros Stamatakis, and Michael J. Donoghue. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* 98(3): 404-414.

Appendix S1. The consensus tree of the maximum likelihood phylogenies for 55,473 species of seed plants and the location of significant shifts in diversification rate across the tree. The data matrix was constructed using the mega-phylogeny method and includes DNA sequences for six gene regions: *atpB, matK, trnK, trnL, rbcL,* and nuclear ribosomal ITS. Black branches denote inferred shifts in diversification rate based on Δi statistic.



Smith, Stephen A., Jeremy M. Beaulieu, Alexandros Stamatakis, and Michael J. Donoghue. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* 98(3): 404-414.

Appendix S2. *Phylogenetic construction.*

We constructed the phylogenetic dataset for 55,473 seed plants using PHLAWD and the methodology described in Smith et al., (2009). The dataset included flowering plants and Acrogymnospermae (for rooting) using available data from GenBank for the gene regions atpB, matK, trnK, trnL, rbcL, and ITS. These gene regions were concatenated creating a dataset with 9853 sites.

Phylogenetic inference has been facing two major challenges over the last couple of years (i) the immense pace of accumulation of molecular data and (ii) the multi-core revolution, that is, the introduction of parallel computing at the laptop and desktop level for building ever more powerful computers. In general, work on phylogeny reconstruction methods over the past years can be regarded as a continuous struggle to maintain program scalability with input dataset growth, both with respect to increasing numbers of taxa and genes. In fact, the memory consumption of phylogenomic analyses, especially for memory-intensive likelihood-based (Maximum Likelihood and Bayesian) programs, is becoming an increasing concern (RAxML memory requirements of up to 190GB have been reported; Ziheng Yang, pers. comm., May 2010), in particular with the announcement of large-scale genome sequencing projects such as the 10K Vertebrate project (http://www.genome10k.org/). With the increasing number of cores available on current multi-core and supercomputing systems, the fine-grained parallelization of the likelihood function is becoming more important for accelerating programs, handling their memory requirements, and exploiting modern computer and accelerator architectures such as Graphics Processing Units (GPUs) (Ott et al., 2008; Stamatakis and Ott, 2008; Pratas et al., 2009; Suchard and Rambaut, 2009). With respect to heuristic tree search approaches, there has been a convergence of the strategies deployed (Guindon et al., 2010) in programs such as GARLI

(Zwickl, 2006), PHYML (Guindon et al., 2010), or RaxML (Stamatakis, 2006b) (and future versions of MrBayes (Ronquist and Huelsenbeck, 2003); John Huelsenbeck, pers. comm., May 2010) that are all based on some flavor of so-called lazy SPR moves (Stamatakis et al., 2005). Algorithmic progress has also been achieved with fast approaches for computing support values such as the RAxML rapid bootstrap algorithm (Stamatakis et al., 2008) or the aLRT (Anisimova and Gascul, 2006) test (implemented in PHYML 3.0 [Guindon et al., 2010], FastTree 2.0 [Price et al., 2010], and RAxML v727). Finally, the community is also facing two technical challenges (i) the increasing code complexity of phylogeny reconstruction programs that offer a plethora of substitution models and can analyze binary, DNA, morphological, secondary structure, protein, and Codon data. This increase in code complexity poses several difficult software engineering challenges. (ii) The absence of checkpoint mechanisms in most programs, that is, the ability to stop and restart long-running analyses may hinder such analyses in the future. The personal view of the authors is that the most urgent issue that needs to be addressed is memory consumption though (see Stamatakis and Alachiotis, 2010 for a novel solution to reduce memory consumption for "gappy" phylogenomic datasets).

Phylogenetic trees were inferred using the Pthreads-based and SSE3-vectorized RAxML (Stamatakis, 2006b) version 7.2.6. The post-analysis steps (consensus tree building, evaluating the final trees under the GTR+GAMMA model etc.) were carried out with RAxML v7.2.7. We used the standard RAxML search algorithm with the asymptotic stopping rule and the low memory consumption flag (-F and -D options) to infer 223 ML trees on the original alignment under the GTR+CAT approximation of rate heterogeneity (Stamatakis, 2006a) and a partitioned model (we estimated the GTR and alpha parameters separately for each gene) with a joint branch length estimate. The usage of the GAMMA model of rate heterogeneity was not possible on all multi-core systems we used for the analysis because of memory limitations (a run under GTR+CAT required approximately 30GB of main

memory, a run under GAMMA requires approximately four times more memory). Branch lengths and likelihood scores under GTR+GAMMA for all 223 ML trees were computed using the -f n option.

We also inferred 244 bootstrap trees using the RAxML rapid bootstrap algorithm (Stamatakis et al., 2008). We then plotted BS support values onto the best-scoring ML tree and also computed strict, majority-rule, and extended majority rule consensus trees for the bootstrap replicates and the ML trees on the original alignment. We also applied the bootstopping (bootstrap convergence) tests (Pattengale et al., 2010) a posteriori to the bootstrap trees. The test indicated that an insufficient number of BS replicates has been computed to guarantee stable support values. Finally we computed pair-wise Robinson Foulds (RF) distances between all ML trees (average relative RF: 21.79%) and all bootstrap trees (average relative RF: 53.32\%).

For our analyses we used two 32-core Sun x4600 multi-core systems with 64GB of main memory and several AMD Barcelona 16-core systems with 64GB and 128GB of main memory at the Laboratory for Computational Biology and Bioinformatics at the Swiss Federal Institute of Technology and the Texas Advanced Computing Center (TACC). We executed one multi-threaded RAxML run per node at a time (note that, a single ML search required approximately 6 days on 16 cores). In order to make use of the resources at TACC which is a typical supercomputer system with job run-time restrictions, we deployed appropriate scripts and the newly integrated restart capability of RAxML that makes use of the dmtcp library (http://dmtcp.sourceforge.net/). To compute the likelihood scores and estimate the branch lengths of the final ML trees under GTR+GAMMA we completely re-designed the numerical scaling procedure in RAxML that is used to prevent numerical underflow in the likelihood function. For trees of this size it turned out that, the previous scaling procedure implemented in RAxML was not sufficient for obtaining numerically stable results. Since the adapted scaling procedure is more

compute-intensive (more arithmetic operations are required for scaling) the RAxML version that contains this scaling procedure for large-scale datasets can be compiled using a separate Makefile and has been made available in RAxML v7.2.7. Finally, the consensus tree building step was also conducted with RAxML v7.2.7 that uses a faster tree parsing mechanism as well as some algorithmic optimizations for building extended majority rule consensus trees (Aberer et al., 2010).

Supplementary References

Aberer, A. J., N. D. Pattengale, and A. Stamatakis. 2010. Parallel computation of phylogenetic consensus trees. Procedia Computer Science 1:1059-1067.

Anisimova, M. and O. Gascuel. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Systematic Biology 55: 539-552.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59: 307.

Ott, M., J. Zola, S. Aluru, A.D. Johnson, D. Janies and A. Stamatakis. 2008. Large-scale Phylogenetic Analysis on Current HPC Architectures. Scientific Programming 16: 255-270.

Pattengale, N. D., M. Alipour, O. R. P. Bininda-Emonds, B. M. E. Moret, and A. Stamatakis. 2008. How many bootstrap replicates are necessary? Journal of Computational Biology 17: 337-354.

Pratas, F., P. Trancoso, A. Stamatakis, and L. Sousa. 2009. Fine-grain Parallelism for the Phylogenetic Likelihood Functions on Multi-cores, Cell/BE, and GPUs. Proceedings of ICPP 2009.

Price, M. N., P. S. Dehal and A. P. Arkin. 2010. FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments. PloS One.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Stamatakis, A. 2006a. Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. Proceedings of IPDPS 2006. HICOMB Workshop, Proceedings on CD, Rhodos, Greece.

Stamatakis, A. 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stamatakis, A. and N. Alachiotis. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. Bioinformatics 26: 1132.

Stamatakis, A. and M. Ott. 2008. Exploiting Fine-Grained Parallelism in the Phylogenetic Likelihood Function with MPI, Pthreads, and OpenMP: A Performance Study. PRIB eds. Madhu Chetty and Alioune Ngom and Shandar Ahmad. Pgs. 424-435. Lecture Notes in Computer Science.

Stamatakis, A., T. Ludwig and H. Meier. 2005. RAxML-III: A Fast Program for Maximum Likelihood-based Inference of Large Phylogenetic Trees. Bioinformatics 21:456-463.

Stamatakis, A., P. Hoover and J. Rougemont. 2008. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Systematic Biology 57:758-771.

Suchard, M. A. and A. Rambaut. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25: 1370.

Zwickl, D. 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. Ph. D. Thesis University of Texas at Austin.