

Leveraging skewed transcript abundance by next-generation sequencing to increase the genomic depth of the tree of life

Chris Todd Hittinger^{1,2}, Mark Johnston^{1,2}, John T. Tossberg³, Antonis Rokas³

¹Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

²Center for Genome Sciences, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, MO 63108, USA

³Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

Assembling the tree of life is a major goal of biology, but the difficulty and expense of obtaining sufficient orthologous DNA hinders progress toward fully resolved phylogenies. Next-generation DNA sequencing technologies could accelerate progress, but genome sequencing remains impractical for most of the 1.8 million described species. Eukaryotic transcriptomes are smaller, easier to assemble from short reads, and biased toward highly-expressed housekeeping genes that tend to be conserved, meaning they could provide a rich set of phylogenetic characters. Here we sampled the non-normalized transcriptomes of 10 mosquito species by assembling 36 base-pair sequence reads into a massive phylogenomic data matrix containing nearly one million orthologous nucleotides from over 2,500 genes. Analysis using several different procedures yielded well-supported phylogenetic inferences that were robust to rare orthology assignment errors (assessed by comparison with two annotated mosquito genome sequences to be between 1% and 18% depending on the stringency of assumptions). Due to the high effective coverage of a subset of genes that were highly expressed in all species, surprisingly few sequence reads were required to assemble large data matrices and make strong phylogenetic inferences. This approach is more data-rich, efficient, and economical than traditional methods, and provides a scalable phylogenomic strategy to infer the branches and twigs of the tree of life.

OPEN SOURCE NOTE: This abstract is about the utility, efficacy, and cost of a specific approach for the generation of phylogenomic data matrices and does not describe a software package. Therefore, the Open Source License requirements are not applicable. We used several publicly available third-party software packages and are happy to provide full methods and/or a list of the software used upon request. All short read sequence data have been deposited with the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) and are publicly accessible as SRP001532 of SRA010237.