NEXT GENERATION PHENOMICS FOR THE DIATOM TREE OF LIFE
-USING THALASSIOSIARLES AS AN EXAMPLE

Mengjie Yu¹, Edward C Theriot¹, Andrew J Alverson², Matthew Julius³, J. Gordon Burleigh⁴, Kenzey Alphonse⁵, Holly M Bik⁶, Carrine Blank⁷, Andrea L Cirranello⁸, Hong Cui⁹, Marymegan Daly¹⁰, Thomas G Dietterich¹¹, Gail Gasparich¹², Jed Irvine¹¹, Seth Kaufman¹⁶, Edith Law¹³, Jing Liu⁴, Lisa Moore¹⁴, Maureen A O'Leary¹⁵, Maria Passarotti¹⁶, Sonali Ranade⁹, Nancy B Simmons⁸, Dennis W. Stevenson¹⁷, Robert W Thacker¹⁸, Sinisa Todorovic¹¹, Paul M. Velazco¹⁹, Ramona L Walls²⁰, Joanna M Wolfe¹⁹

¹ Section of Integrative Biology, University of Texas at Austin, ² Department of Biological Sciences, University of Arkansas, ³ Department of Biological Sciences, St. Cloud State University, ⁴ Department of Biology, University of Florida, ⁵ KenX Technology, Medford, New York, ⁶ UC Davis Genome Center, ⁷ University of Montana, ⁸ Department of Mammalogy, American Museum of Natural History, ⁹ School of Information Resources & Library Science, University of Arizona, ¹⁰ Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, ¹¹ School of Electrical Engineering and Computer Science, Oregon State University, ¹² Department of Biological Sciences, Towson University, ¹³ Center for Research on Computation and Society, School of Engineering and Applied Sciences, Harvard University, ¹⁴ University of Southern Maine, ¹⁵ Department of Anatomical Sciences, Stony Brook University, ¹⁶ Whirl-i-Gig, Greenport, New York, ¹⁷ The New York Botanical Garden, ¹⁸ University of Alabama Birmingham, ¹⁹ American Museum of Natural History, ²⁰ University of Arizona

The present revolution in the field known as genomics has been made possible in large part by advances in computer sciences in the area called bioinformatics. This revolution has been applied to diatom systematics with great success. Millions to billions of data points about multiple diatom genomes can be generated, sorted, synthesized and interpreted in the time it once took to prepare and study one diatom sample for LM, SEM and TEM. Between this, the relative shortage of number of diatom specialists and the tremendous diversity of diatoms, our ability to generate these data are far outstripping our ability to understand and interpret them in any but the most coarse-grained way. Not only are new morphological data difficult to generate, it is at least equally difficult to gather and synthesize existing data. This is because of non-standard use of terms, non-standard presentation of images, and the qualitative nature of morphology. These problems are more or less the same across the entire Tree of Life. The Theriot Lab, with collaborators Matt Julius and Andy Alverson, were invited to participate in a project lead by Maureen O’Leary and funded the by NSF Assembling and Visualizing the Tree of Life program. A team of informaticians, artificial intelligence specialists, computer visualization experts and other computer scientists have been assembled to coordinate with working taxonomists to develop working and prototypical models for integrated computer visualization, natural language processing, and crowd-sourcing strategies to automate and accelerate the gathering of morphological data (and other phenotypic/epigenetic data such as physiological data). We will present an overview of the project and explain our efforts to build a model standardized reference system for diatoms using our combined expertise in the Thalassiosirales as a starting point.