Abstract: The Microbial Genomes Atlas Science Gateway -- MiGA @ XSEDE: A Searchable Database of Prokaryotic Genomes for Taxonomic Identification and Diversity Cataloguing

The diversity of prokaryotic microbes on the planet is very large, estimated at over one billion species of bacteria, and most of it remains undiscovered. As genome sequencing can help characterizing this diversity and has recently become routine, most microbial scientists have been overwhelmed by the amount of genomic data that were made recently available. Tools that can help direct researchers to the most "interesting" genomes among thousands of candidates will be important, including for identification (diagnostics) of microbial disease agents in addition to diversity discovery. However, current tools to analyze metagenomic data are clearly lagging behind the development of sequencing technologies (and data), and are typically limited to genome assembly and gene annotation. In response, we recently introduced the Microbial Genomes Atlas (MiGA) (Rodriguez-R et al., 2018), a genomic data processing and management system that uses whole-genome comparisons for the identification of relatives and taxonomic classification, and provides several tools for genome quality evaluation and genome clustering for novel microorganisms. Together with the MiGA infrastructure, we also released the MiGA Online webserver, an online system that allows users evaluating, comparing, and classifying their own genome sequences against different reference databases including the collection of all complete prokaryotic genomes in NCBI (NCBI_Prok, ~14,000 genomes), all reference genomes in RefSeq (RefSeq; ~2,000 genomes), and two large collections of metagenome-assembled genomes (MAGs and Parks8, with ~3,000 and ~8,000 genomes, respectively), among several others. MiGA is currently being used by hundreds of users, and has already processed about 9,000 query genomes, which is remarkable for a new resource (less than 1 year since the publication of the MiGA manuscript) and a testament that MiGA fulfils a critical need of contemporary research and education. Indeed, MiGA has already been used for the proposal of novel taxa, the classification and evaluation of microbial genomes, or to discuss data-driven microbial taxonomy. Notably, MiGA is currently unique among similar efforts by others in that it allows external users to query their own sequences against MiGAâ€™s internal databases and not only provides taxonomic classification but also assessment of genome quality, completeness, and gene content variation.

MiGA includes a series of heuristics to allow the rapid identification of closest relatives using whole-genome comparisons. However, indexing and processing query datasets remains computationally challenging, given the size and growth rate of the databases. We currently invest around 240 thousand CPU hours each month on this task, the equivalent of over 300 dedicated CPUs at 100% capacity. Our current expansion plans include a new database including all named species with available genomes (complete or draft) as well as high-resolution databases for species with multiple available genomes. Our projections indicate that about 1.4 million SSUs are required to index these databases and support online querying. In this startup allocation project, we request access to the computational resources of Comet in XSEDE to index and offer online querying of these databases though a MiGA Science Gateway. In addition to the benefits for the research community derived from the availability of these new databases, the availability of the MiGA infrastructure and theses databases in XSEDE will allow high-throughput access to the MiGA resources by any researcher in the XSEDE system for both research and education purposes. We have successfully delivered bioinformatics workshops using the MiGA Command Line Interface in the past focused on processing genomic and metagenomic data, at a small scale (processing ~10 genomes), using commercial cloud computing (Amazon Web Services). With MiGA available in XSEDE, these educational materials would also be available to any person with access to the XSEDE system, and at a large scale