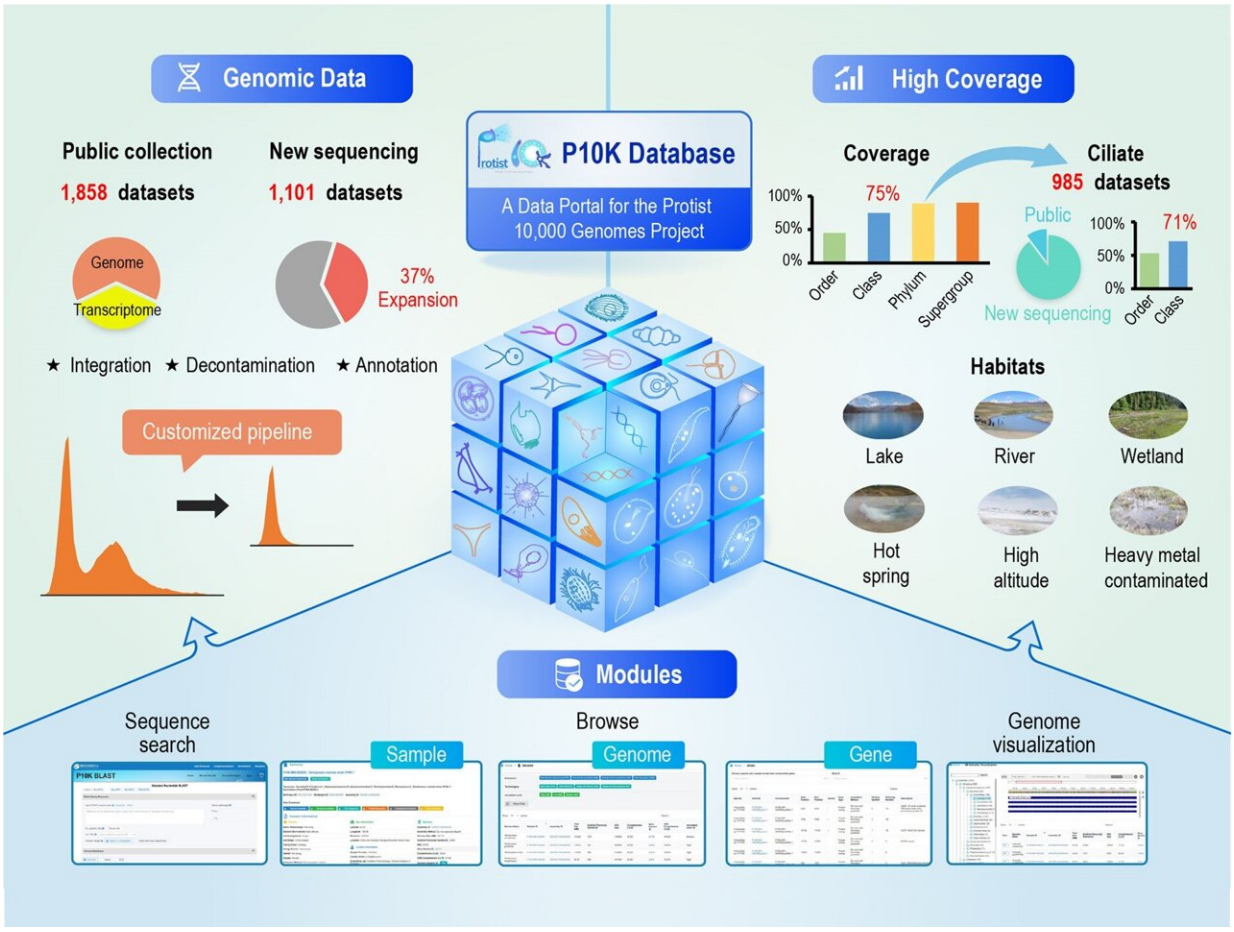
Researchers release initial dataset for protist genomes project

January 3 2024, by Liu Jia



Graphical abstract. Credit: *Nucleic Acids Research* (2023). DOI: 10.1093/nar/gkad992

Protists, single-celled eukaryotic organisms encompassing unicellular algae and protozoans, inhabit aquatic environments. Functioning as primary producers and oxygen generators, they play crucial roles in the carbon cycle and serve as vital sources of human nutrition, bioenergy, and food for aquatic animals. However, they can also pose challenges, causing harmful algal blooms and red tides, acting as both pathogens and beneficial partners in symbiotic relationships.

The NCBI taxonomy system has documented over 60,000 identified protist species. In December 2019, a group of scientists led by the Institute of Hydrobiology (IHB) of the Chinese Academy of Sciences (CAS) launched the Protist 10,000 Genomes Project (P10K). The primary aim of this project is to create a comprehensive genetic resource database for protists.

Recently, Prof. Miao Wei's team at the IHB and Prof. Zhang Zhang's team from the Beijing Institute of Genomics of CAS (China National Center for Bioinformation) released the initial dataset from the P10K project which is now available, and the related paper was published in *Nucleic Acids Research*.

The inaugural data released from the P10K comprises a comprehensive set of 2,959 protist datasets, featuring 1,601 genomes and 1,358 transcriptomes. Among these datasets, 1,858 were integrated from public databases. The P10K team undertook new sequencing for 1,101 datasets with a primary focus on ciliates. The newly sequenced data contributed to a substantial 37% expansion in the overall size of the protist dataset.

To overcome the analytical challenges posed by large-scale single-cell omics data, the P10K team developed a standardized analysis pipeline tailored for single-cell sequencing data of protists. This pipeline encompasses the assembly, decontamination, species identification, gene annotation, and evaluation processes.

Quality assessments revealed that genomes annotated through this pipeline exhibit a similar proportion of medium and high-quality data compared to those available in public databases.

The researchers believe the P10K database will promote research on eukaryotic origins, diversity, and microbial interactions, and the applications of protist genetic resources in ecological conservation, pollutant degradation, nutrition, health, and disease prevention. In addition, the database will support the identification of planktonic organisms based on environmental DNA (eDNA), facilitating aquatic ecological health assessments.

More information: Xinxin Gao et al, The P10K database: a data portal for the protist 10 000 genomes project, *Nucleic Acids Research* (2023). DOI: 10.1093/nar/gkad992

Provided by Chinese Academy of Sciences