Genomes on a Tree (GoaT): A centralized resource for eukaryotic genome sequencing initiatives

Cibele Gomes Sotero-Caio‡, Richard Challis‡, Sujai Kumar‡, Mark Blaxter‡

‡ Wellcome Sanger Institute, Hinxton, United Kingdom

Corresponding author: Cibele Gomes Sotero-Caio (cs43@sanger.ac.uk)

Abstract

Genomic data are transforming our understanding of biodiversity and, under the umbrella of the Earth BioGenome Project (EBP - https://www.earthbiogenome.org), many initiatives seek to generate large numbers of reference genome sequences. The distributed nature of this work makes coordination essential to ensure optimal synergy between projects and to prevent duplication of effort. While public sequence databases hold data describing completed projects, there is currently no global source of information about projects in progress or planned. In addition, the scoping and delivery of sequencing projects benefits from prior estimates of genome size and karyotype, but existing data are scattered in the literature. To address these issues, the Tree of Life programme (https://www.sanger.ac.uk/programme/tree-of-life/) has developed Genomes on a Tree (GoaT), an ElasticSearch-powered, taxon-centred database that collates observed and estimated genome-relevant metadata—including genome sizes and karyotypes—for eukaryotic species. Missing values for individual species are estimated from phylogenetic comparison. GoaT also holds declarations of actual and planned activity, from priority lists and in-progress status, to submissions to the International Nucleotide Sequence Database Collaboration (INSDC https://www.insdc.org/), across genome sequencing consortia. GoaT can be queried through a mature API (application programming interface), and we have developed a web front-end that includes data summary visualisations (see https://goat.genomehubs.org/). We are currently transitioning this service into the Tree of Life production pipeline. GoaT currently reports priority lists from the Darwin Tree of Life project (focussed on the biodiversity of Britain and Ireland). We are actively soliciting additional data concerning progress and intent from other projects so that GoaT displays a real-time summary of the state of play in reference genome sequencing, and thus facilitates collaboration and cooperation among projects. We are developing standard formats and procedures so that any project can make explicit its intent and progress. Cross referencing to other data systems such as the INSDC sequence databases, the BOLD DNA barcodes resource and Global Biodiversity Information Facility- and Open Tree of Life-related taxonomic and distribution databases

©

will further enhance the system's utility. We also seek to incorporate additional kinds of metadata, such as sex chromosome systems, to augment the utility of GoaT in supporting the global genome sequencing effort.

Keywords

database, Earth BioGenome Project, genome size, karyotype, metadata, species list, sequencing status

Presenting author

Cibele Gomes Sotero-Caio

Presented at

TDWG 2021

Conflicts of interest