Biodiversity
Community
Integrated
Knowledge
**L**ibrary

BiCIKL

Curation of DNA Sequences

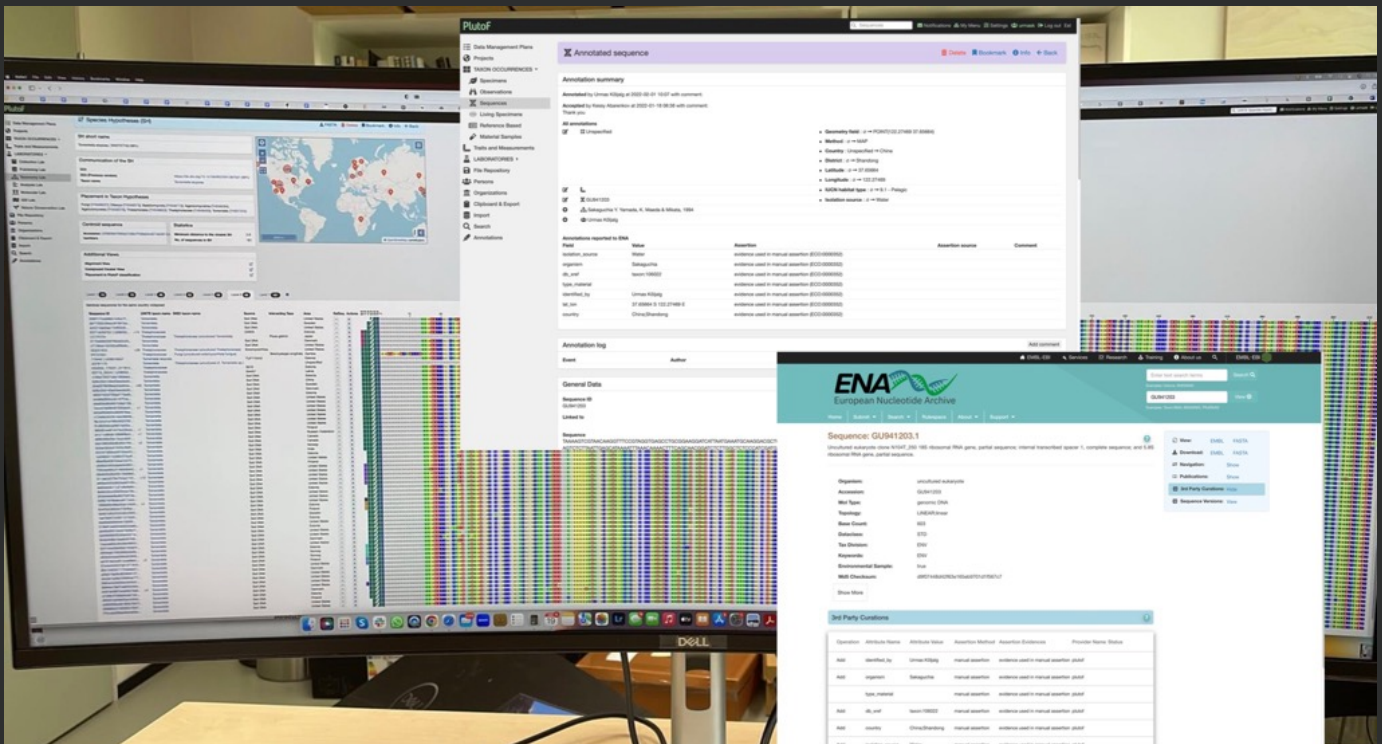# PlutoF

What is the Curation of DNA sequences service?

eDNA sequence curation service provided by the PlutoF platform offers the possibility to improve the quality of public DNA sequences and their source metadata (e.g., material source, geolocation, etc.). In collaboration with EMBL-EBI, errors and gaps in sequenced material source annotations on INSDC (International Nucleotide Sequence Database Collaboration) sequences in PlutoF are fed back to primary repositories through the operation of ELIXIR Contextual Data ClearingHouse. These corrections and updates are displayed on ENA (European Nucleotide Archive) sequence record pages alongside the original data.

Why do we need third-party curation of INSDC sequences?

Third-party curations are a valuable resource for improving the quality of public DNA sequences. For instance, sequences in INSDC often lack important features like taxon interactions, taxonomic identifications, information associated with material source, habitat, geolocation, literature, etc. A sequence may be tagged with the wrong country of origin, or the name of the host may be misspelled. This hampers the utilisation of the full power of such metadata in taxonomic, phylogenetic, and ecological studies. Therefore, initiatives to mine additional information from publications and link it to public DNA sequences has become common

practice. With the new third-party curation service developed as part of the BiCIKL project, it is now possible to publish these third-party curations to a common data warehouse in a standardised format, where these data are available for automated querying and downloading through RESTful web services.
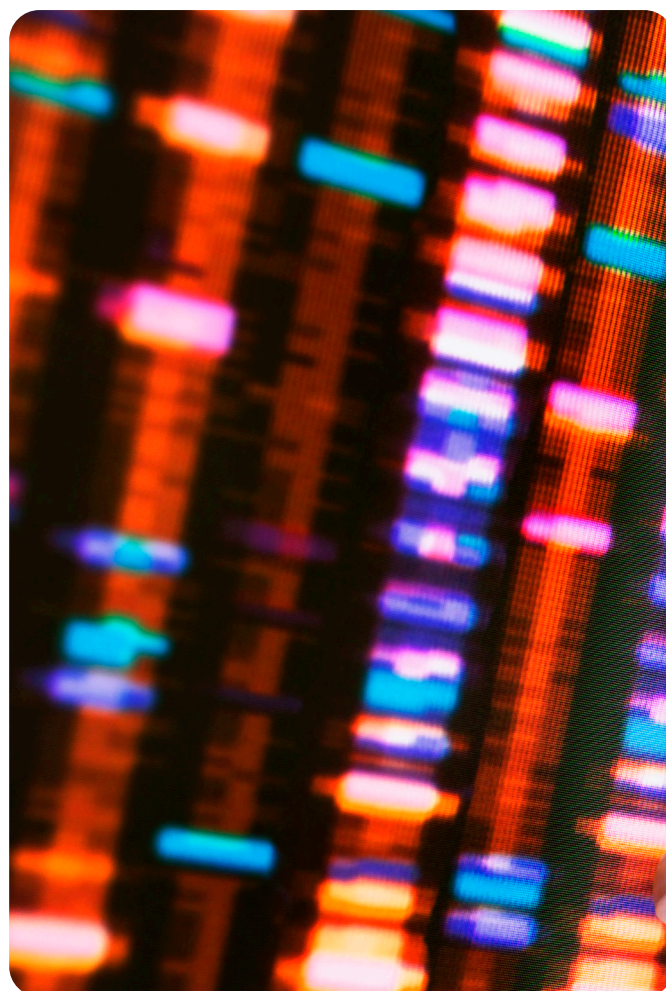
Monitor displaying the Species Hypotheses on the PlutoF platform. Annotating sequence window and ENA window displaying annotated sequence data sent from PlutoF platform.

Who is using it?

The DNA sequence curation service was primarily designed for individual researchers, research collaborations, and communities, as well as citizen scientists, but it is open for anyone wishing to improve the quality of public DNA sequences. One of the main users is the UNITE Community, which uses this service to curate INSDC rDNA ITS datasets. These datasets are utilised in calculating UNITE species hypotheses and are provided as reference datasets for the molecular identification of fungi and other eukaryotes.

Examples and current statistics

- **Example 3rd party curation:** https://www.ebi.ac.uk/ena/browser/view/KJ180283

- **Total number of 3rd party curations sent from PlutoF to ENA:** 4 624

- **The number of contributors:** 28

- **3rd party curations in PlutoF by annotated field:**

Annotated field	Count
Location	3 582
Taxon (name)	1 132
Measurements or Facts	433
Typification	55
Collector	16