BEST PRACTICES ON SUBMITTING DATA TO THE PUBLIC SEQUENCE DATABASES (ID 391)

Presenter
Conrad L. Schoch

BEST PRACTICES ON SUBMITTING DATA TO THE PUBLIC SEQUENCE DATABASES

Conrad L. Schoch

NCBI Taxonomy

National Library of Medicine
National Center for Biotechnology Information

International Nucleotide Sequence Database Collaboration

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI.

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

NCBI

ENA European Nucleotide Archive

DDBJ DNA Data Bank of Japan

Newly introduced missing value reporting at INSDC

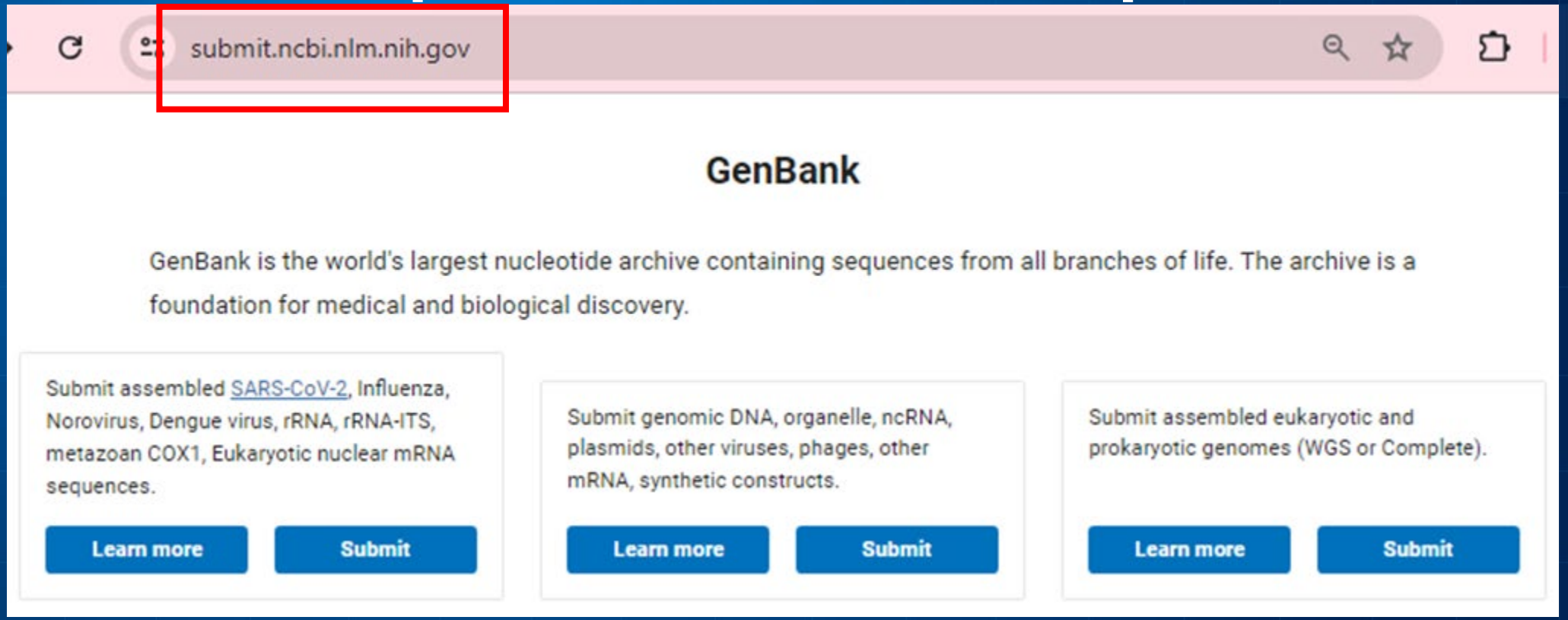
Location of collection: the locality of isolation of the sequenced sample should be indicated to country level at least and should be provided in terms of political names for nations, oceans or seas using values from the controlled vocabulary at http://www.insdc.org/documents/country-qualifier-vocabulary

Date/time of collection: the date and time at which the specimen was collected should be provided, at least to the nearest year.

Newly introduced missing value reporting at INSDC

INSDC term (top level)	INSDC term (lower level)	Definition	INSDC term (reporting level)	Definition
not applicable		Information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately	control sample	Information is not applicable as the sample represents a negative control sample collected in a lab.
			sample group	Information is not applicable as the sample represents a group of samples that do not have a single origin. E.g. for co-assembly or transcriptome assembly.

Before publication: choose portal



Specific workflows for the sequence types listed above. rRNA, rRNA-ITS, metazoan COX1, and specific viruses have automated annotation!

Submit sequences not listed to the left or right to **BankIt**. *Note: BankIt will be going away within a year or two as we consolidate submission tools.*

Submission Portal Genomes – submit Prokaryotic and Eukaryotic Genomes

Runs of unassembled reads, (e.g. Illumina), can be submitted to the Sequence Read Archive (SRA).

Submission Questions? Write to: gb-admin@ncbi.nlm.nih.gov

Before publication

- Annotate protein markers, else they will flagged UNVERIFIED and exclude from blast
- Trim sequences appropriately – ensure vectors and low quality stretches removed
- Add valid voucher information
- Documentation is available and always ask questions if more info is needed.

Before publication: type material

From: Renner et al. (10.1093/sysbio/syad068)

After publication

- Ensure taxonomic names are updated (send pdf if possible)
- Update publication on nucleotide records (send pdf if possible)
- Ensure type material and all taxonomic data elements are correct in NCBI TaxBrowser