Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

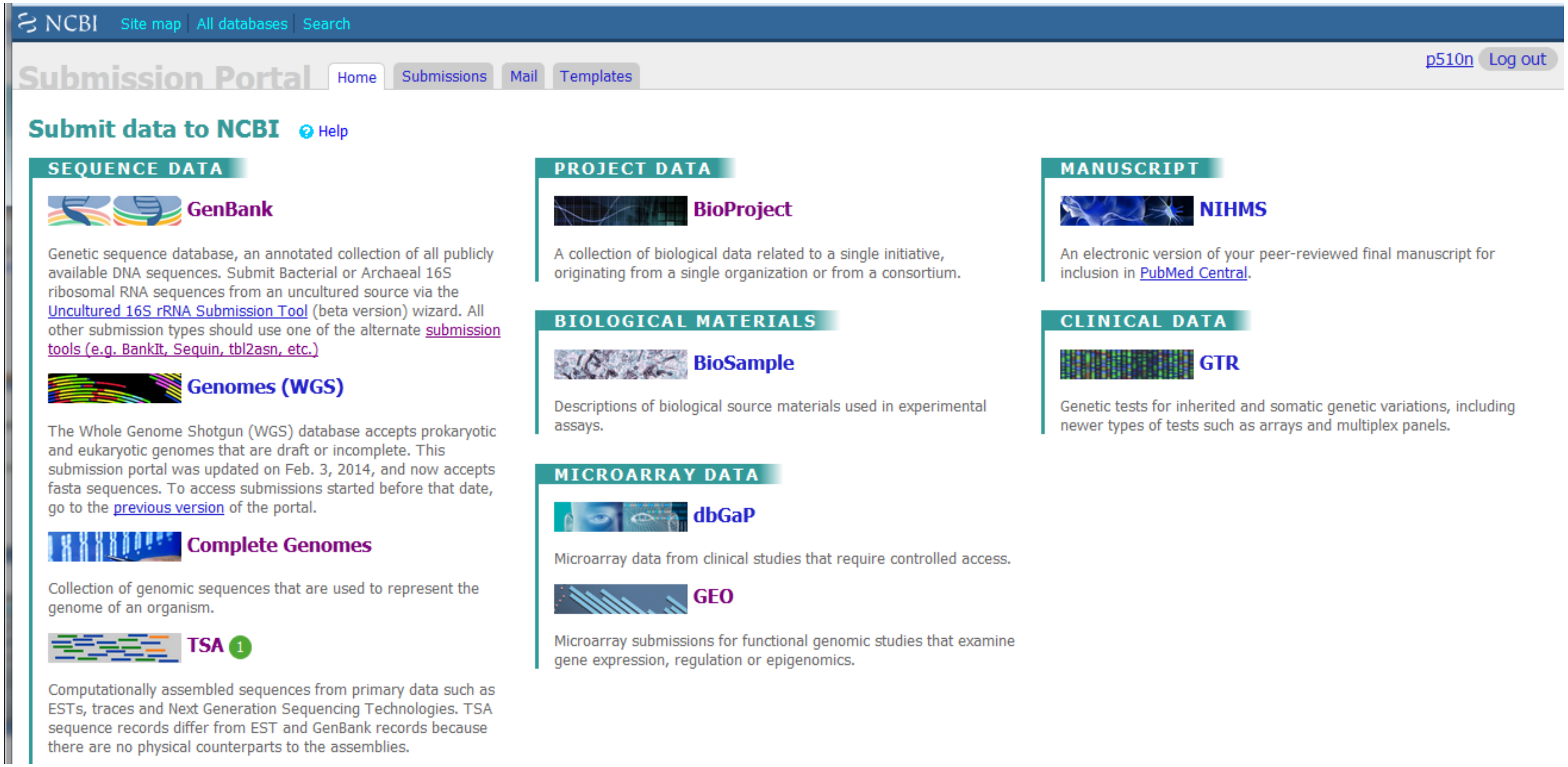
Isak Sylvin

Data submission

Repositories, tools and types

Where do I start?

https://submit.ncbi.nlm.nih.gov/



Submit data to NCBI

Help

SEQUENCE DATA

GenBank

Genetic sequence database, an annotated collection of all publicly available DNA sequences. Submit Bacterial or Archaeal 16S ribosomal RNA sequences from an uncultured source via the Uncultured 16S rRNA Submission Tool (beta version) wizard. All other submission types should use one of the alternate submission tools (e.g. BankIt, Sequin, tbl2asn, etc.)

Genomes (WGS)

The Whole Genome Shotgun (WGS) database accepts prokaryotic and eukaryotic genomes that are draft or incomplete. This submission portal was updated on Feb. 3, 2014, and now accepts fasta sequences. To access submissions started before that date, go to the previous version of the portal.

Complete Genomes

Collection of genomic sequences that are used to represent the genome of an organism.

TSA 1

Computationally assembled sequences from primary data such as ESTs, traces and Next Generation Sequencing Technologies. TSA sequence records differ from EST and GenBank records because there are no physical counterparts to the assemblies.

PROJECT DATA

BioProject

A collection of biological data related to a single initiative, originating from a single organization or from a consortium.

BIOLOGICAL MATERIALS

BioSample

Descriptions of biological source materials used in experimental assays.

MICROARRAY DATA

dbGaP

Microarray data from clinical studies that require controlled access.

GEO

Microarray submissions for functional genomic studies that examine gene expression, regulation or epigenomics.

MANUSCRIPT

NIHMS

An electronic version of your peer-reviewed final manuscript for inclusion in PubMed Central.

CLINICAL DATA

GTR

Genetic tests for inherited and somatic genetic variations, including newer types of tests such as arrays and multiplex panels.

Main data categories

- WGS and Genomes
- ”Annotated sequences”
    - DDBJ/EMBL/Genbank
        - Continuous
        - Nucleotide data
        - >200 nucleotides
        - Physical counterpart
- Microarray
    - GEO
- SRA
- dbSNP & dbVar
    - <50 and >50 BP seq respectively
- Data for specific purposes (Clinical etc)

Main metadata categories

- BioSample
  - Biological source materials
  - Enviroment

- Bioproject
  - Details about the lab samples
  - Includes BioSample

Genbank submission tools

Tool	Function
Bankit	Web-based, easy, weak
Sequin	Stand-alone, graphical, easy, works offline
Tbl2asn	Stand-alone, command line, strong, offline; **Good for big data**
Barcode	Like Bankit but for Barcode of Life projects based on the COI gene.

About project size

- There are always batch alternatives
- Only use graphical tools on small projects
- Or when unsure

About the hands-on

You will test uploading to:

- Genbank, BankIt
- Genbank, Sequin
- BioSample
- BioProject
- GEO
- SRA

About the hands-on

1. We will not submit anything
2. Play around with as many fields as possible
3. Data can be uploaded to other sections
4. If you are unsure about what a field does; **ASK!**