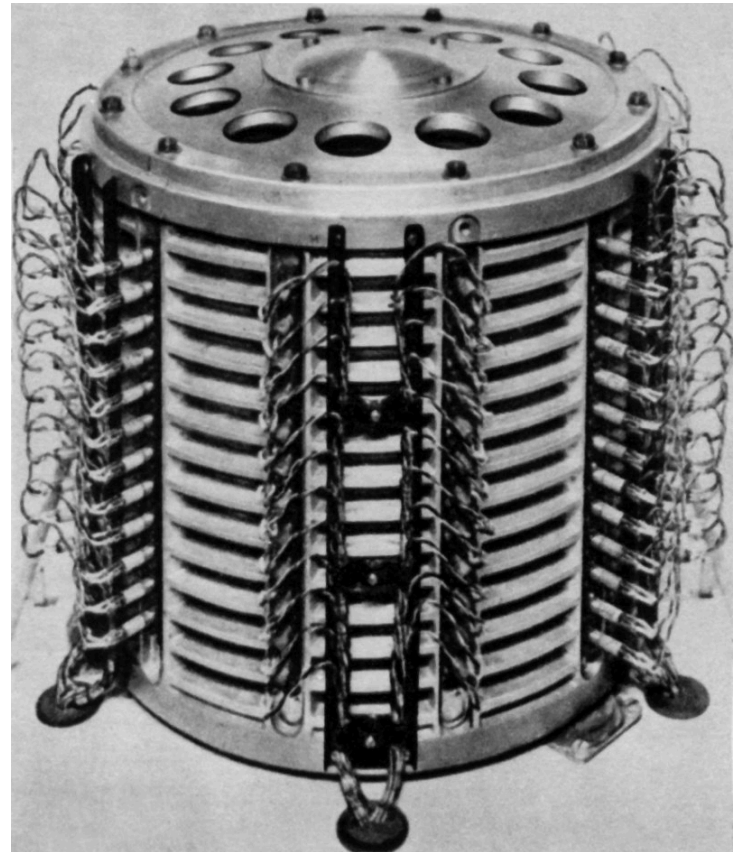


Online Resources

Why do we need databases?

- To archive and preserve information
- To put all the things in one place (facilitate discovery)
- To enforce and maintain format standards
- To allow reuse of data (its expensive, use it more than once!)
- To prevent fraud in research
- To have reproducibility of research
- To store metadata



Metadata

- Metadata is data about data
- Where did the data come from?
 - Organism or substrate, experimental conditions, location, time and date, tissue
- How was the data collected?
 - Field methods, lab methods, instruments, calibration
- How has the data been processed?
 - Normalization, removal of “bad” data, any processing at all
- User rights and management for the data
 - Is this open for additional publication or is it embargoed?
 - Does it carry a license?



A Few Types of Databases

- I. International, Primary Repositories
- II. Protein DB resources
- III. Community DBs

These are quite possibly totally unrelated to your data of interest. Publications and internet searches will help you identify the right database for you.

Sometimes there just isn't a home. Non-human metabolomics data?

International Nucleotide Sequence Database (INSD)

- Consists of the following 3 dbs:
 - DDBJ (DNA Data Bank of Japan)
 - EMBL (European Molecular Biology Laboratory)
 - NCBI (National Center for Biotechnology Information)
- repositories for nucleotide sequence data from all organisms
- all three databases accept nucleotide sequence submissions, and then exchange new and updated data on a daily basis
- Primary database = house original sequence data



NCBI

- Discover
- Download
- Submit

The screenshot shows the NCBI homepage layout. At the top is the NCBI logo and a search bar. Below the logo is a vertical list of resource categories. The main content area features a 'Welcome to NCBI' message and a grid of six action buttons: Submit, Download, Learn, Develop, Analyze, and Research. Each button includes an icon, a title, and a brief description of the service.

NCBI Home
National Center for Biotechnology Information

All Databases

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

<http://www.ncbi.nlm.nih.gov/>

Discover “Entrez”

<http://www.ncbi.nlm.nih.gov/gquery/>

Search NCBI databases

[Help](#)

Results found in 22 databases for "Fraxinus"

Literature

Books	7	books and reports
MeSH	2	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	654	scientific & medical abstracts/citations
PubMed Central	1,060	full-text journal articles

Health

ClinVar	0	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	6	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

Genes

EST	12,100	expressed sequence tag sequences
Gene	5	collected information about gene loci
GEO DataSets	4	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	240	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

Proteins

Conserved Domains	0	conserved protein domains
Protein	809	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	0	experimentally-determined biomolecular structures

Chemicals

Download

<http://www.ncbi.nlm.nih.gov/home/download.shtml>

Download

The majority of NCBI data are available for downloading, either directly from the NCBI FTP site or by using software tools to download custom datasets.



ADDITIONAL LINKS

[How to download custom data sets](#)

[Large Data Download Best Practices](#)

[SRA Download Reference](#)

FTP

Download data from the NCBI FTP site



Aspera

High-speed downloads provided by Aspera software



Download Tools

Tools and APIs for downloading customized datasets





- Private software owned by IBM
- Free for clients
- Can be hundreds of times faster than http and ftp
- For NCBI, you need to download and install a web browser plug in, Aspera Connect




NCBI

Fast Aspera Download [How to setup Aspera.](#)

Please ensure you are running a current version of AsperaConnect. It is available at [Aspera Connect](#) under the "RESOURCES" tab.

Set your bandwidth rate and continue increasing it until the data transfer rate plateaus. Many sites can transfer data at 200-500Mbps. and nearly all sites can transfer at faster than 10Mbps.

Please refer to [Aspera Transfer Guide](#) and [Aspera's documentation](#) for more information.

Name	Total size	Content	Last update
 SRR292241	956.32 Mb	1 file	2015-06-28 01:33
└─ SRR292241.sra	956.32 Mb		2015-06-28 01:33

Sequence Read Archive

- GenBank was the original name for the database to store all sequence reads
- GenBank now encompasses the Sequence Read Archive (SRA), which accepts next generation sequence data
 - Raw sequencing data
 - Alignment information
- <http://www.ncbi.nlm.nih.gov/sra>

Emphasis on metadata

- This is a new paradigm from the old Trace Archive
 - **Study** – A study is a set of experiments and has an overall goal.
 - **Experiment** – An experiment is a consistent set of laboratory operations on input material with an expected result.
 - **Sample** – An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
 - **Run** – Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.

Hierarchical Design

Study

Experiment

Experiment

Sample

Sample

Sample

Sample

Sample

Run

Run

Run

Run

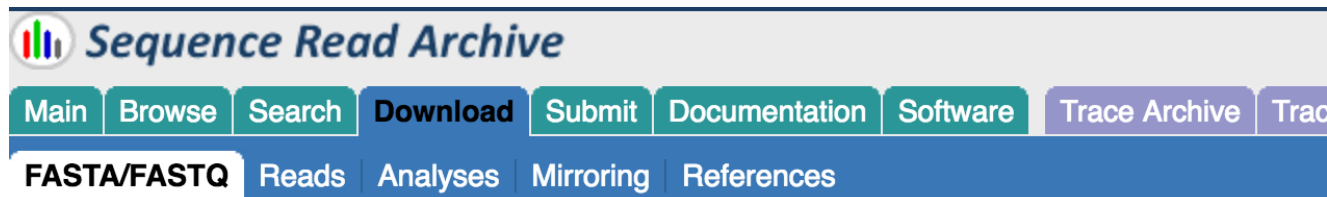
Run

Run

Sequence Read Archive Format

- Their own format : SRA format
- (this is why a lot of our lessons use FASTQ files sourced from EMBL)
- there is a web tool for downloading fastq files if you have a list of accessions and want to do this over the web:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_seq_name



Downloading SRA data in either fasta or fastq format

Experiment(s):

[? What can be entered in this field?](#)

SRA Toolkit

- CLI tool for downloading and converting to/from SRA format
- Two most important commands:
 - **fastq-dump**: Convert SRA data into fastq format
 - **prefetch**: Allows command-line downloading of SRA, dbGaP, and ADSP data

Submitting to the SRA

- Journals will require that you submit all NGS data to SRA or ENA or somewhere legit (and most other 'omic data forms somewhere!)
- Collect all data while the experiment is being done
- Know what data you need – they have spreadsheets!
- Start the submission process early (especially if you have a lot of data)

Hierarchical Design

Study

Experiment

Experiment

Sample

Sample

Sample

Sample

Sample

Run

Run

Run

Run

Run

Run

BioProjects

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

- Biosamples
- Raw reads
- Genome assembly
- Transcriptome assembly
- Genome annotation
- Markers

- You can create a BioProject page at the very beginning of the project (without data)
- Provide ongoing updates

Castanea mollissima strain:Vanuxem Targeted Locus (Loci)

The integrated genetic and physical map for Chinese chestnut was utilized to identify bacterial artificial clones (BACs) located in the three previously identified QTL regions conferring blight resistance. [More...](#)

See [Genome](#)
Information for
Castanea mollissima

Related Resources:

- [Link to assembly results](#)

NAVIGATE ACROSS

2 additional projects
are related by
organism.

Project Data Type: Targeted Locus (Loci)

Attributes: Scope: Monoisolate; Material: Genome; Capture: Targeted Locus Loci; Method type: Sequencing

Relevance: Environmental

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	8
OTHER DATASETS	
BioSample	1

▼ SRA Data Details

Parameter	Value
Data volume, Gbases	7
Data volume, Mbytes	7147

Lineage: *Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Castanea; Castanea mollissima* [Taxonomy ID: 60419]

Submission:

Registration date: 8-Nov-2014

[University of Tennessee Institute of Agriculture](#)

NCBI Data Submission: Earlier is Better!

For submitting any processed data such as a genome sequence: NCBI does a lot of contamination screening, so plan ahead!

1. submit and make sure you pass their QA screens
2. Receive an accession number but keep the data private
3. Do downstream analysis
4. When you are ready to publish, make data public

If you don't do this, then they may ask you to completely change your data, which will (possibly) invalidate your downstream analysis.



Lets go look at NCBI and download some data!

When The Sequencing Data Comes In
#WhatShouldWeCallGradSchool



<http://whatshouldwecallgradschool.tumblr.com/post/127656695585/when-the-sequencing-data-comes-in>