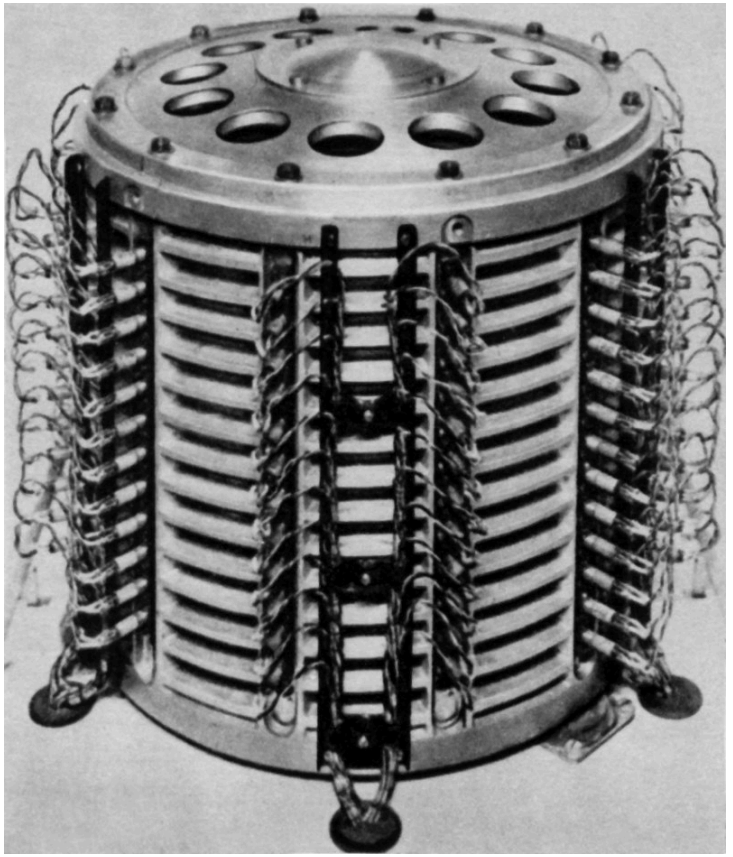
Online Resources for Nucleotide Data

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?

http://twitpic.com/6ry6ar

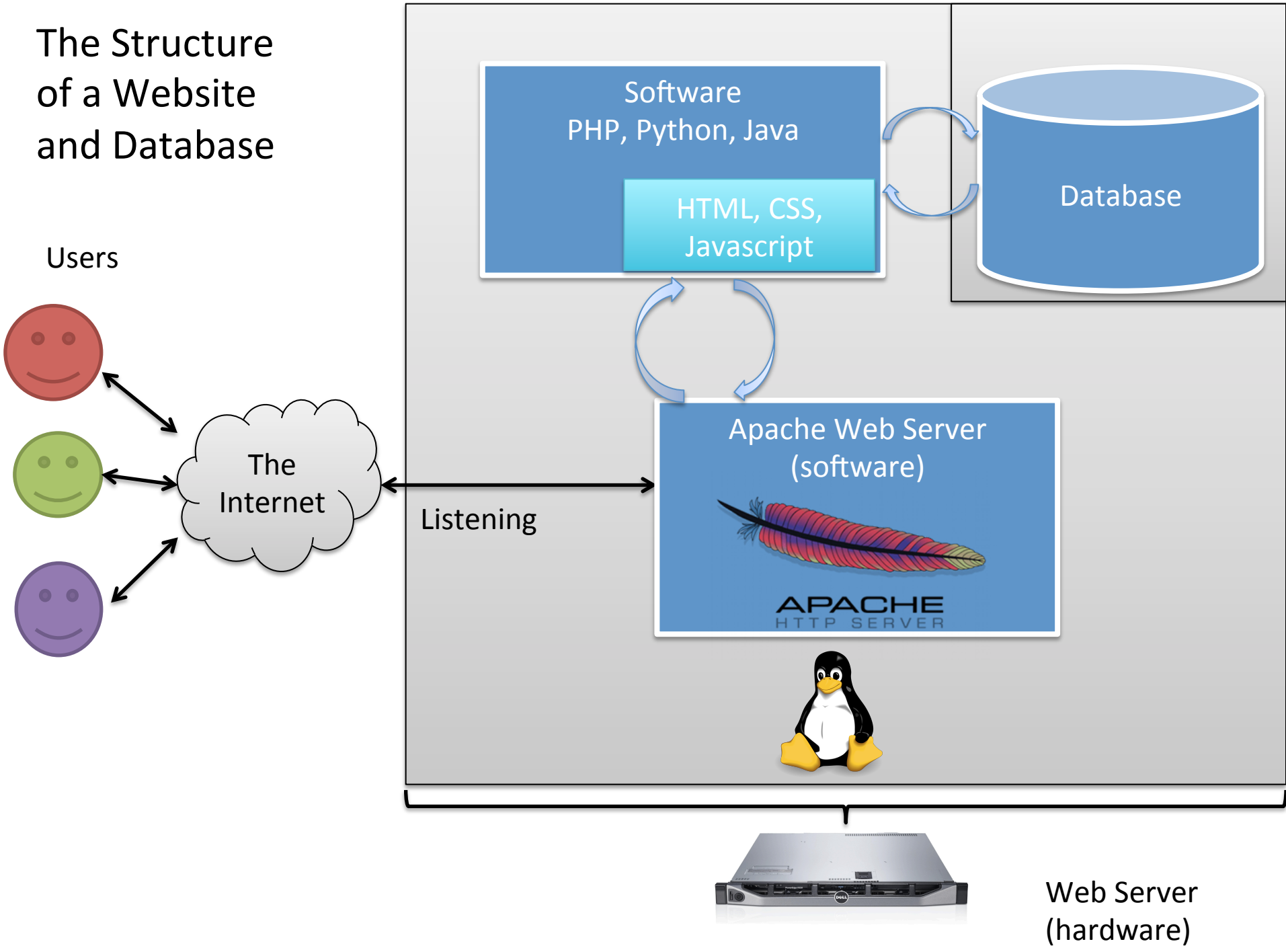
A Few Types of Databases

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.

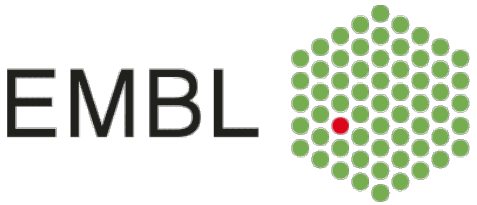
Sometimes there just isn't a home. Non-human metabolomics data?

The Structure of a Website and Database



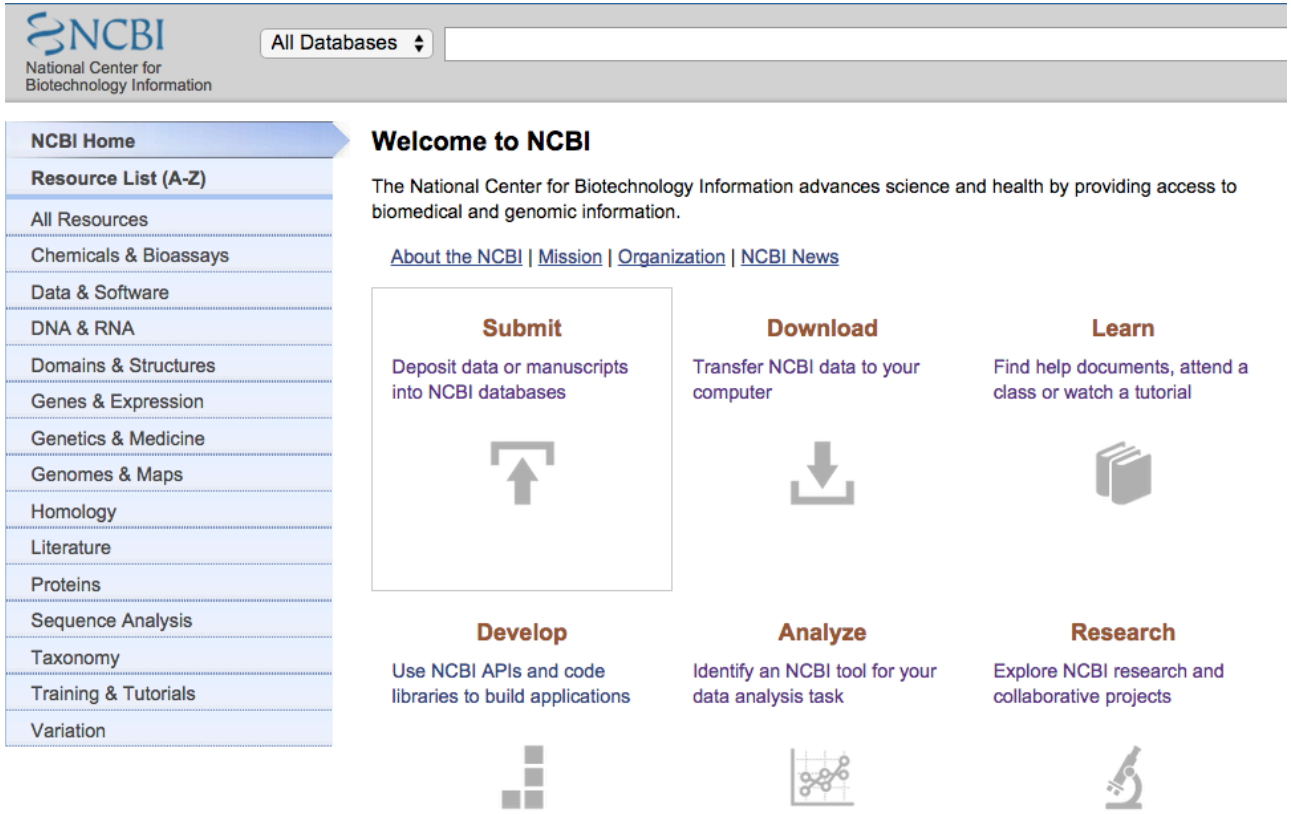
International Nucleotide Sequence Database (INSD)

- 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
- Analyze



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

Discover
“Entrez”
http://www.ncbi.nlm.nih.gov/gquery/

Search NCBI databases Help

Fraxinus Search

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.

Collapse tree

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

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

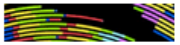
Submit data to NCBI

Help

SEQUENCE DATA

GenBank

Genetic sequence database, an annotated collection of all publicly available DNA sequences. Submit Bacterial or Archaeal 16S ribosomal RNA sequences via the 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

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.

SRA

The Sequence Read Archive (SRA) stores raw and aligned sequence data from NextGen sequencing platforms.

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

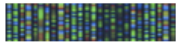
Microarray submissions for functional genomic studies that examine gene expression, regulation or epigenomics.

MANUSCRIPT

NIHMS

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

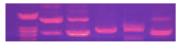
CLINICAL DATA



GTR

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

VARIATION



Variation

Variation resources at NCBI: dbSNP, dbVar and ClinVar. dbSNP represents short variation in any organism including single nucleotide variants, insertions, deletions, and microsatellites. dbVar represents genomic structural variations from studies submitted on any organism or phenotype. ClinVar aggregates information about human sequence variation and its relationship to human health.

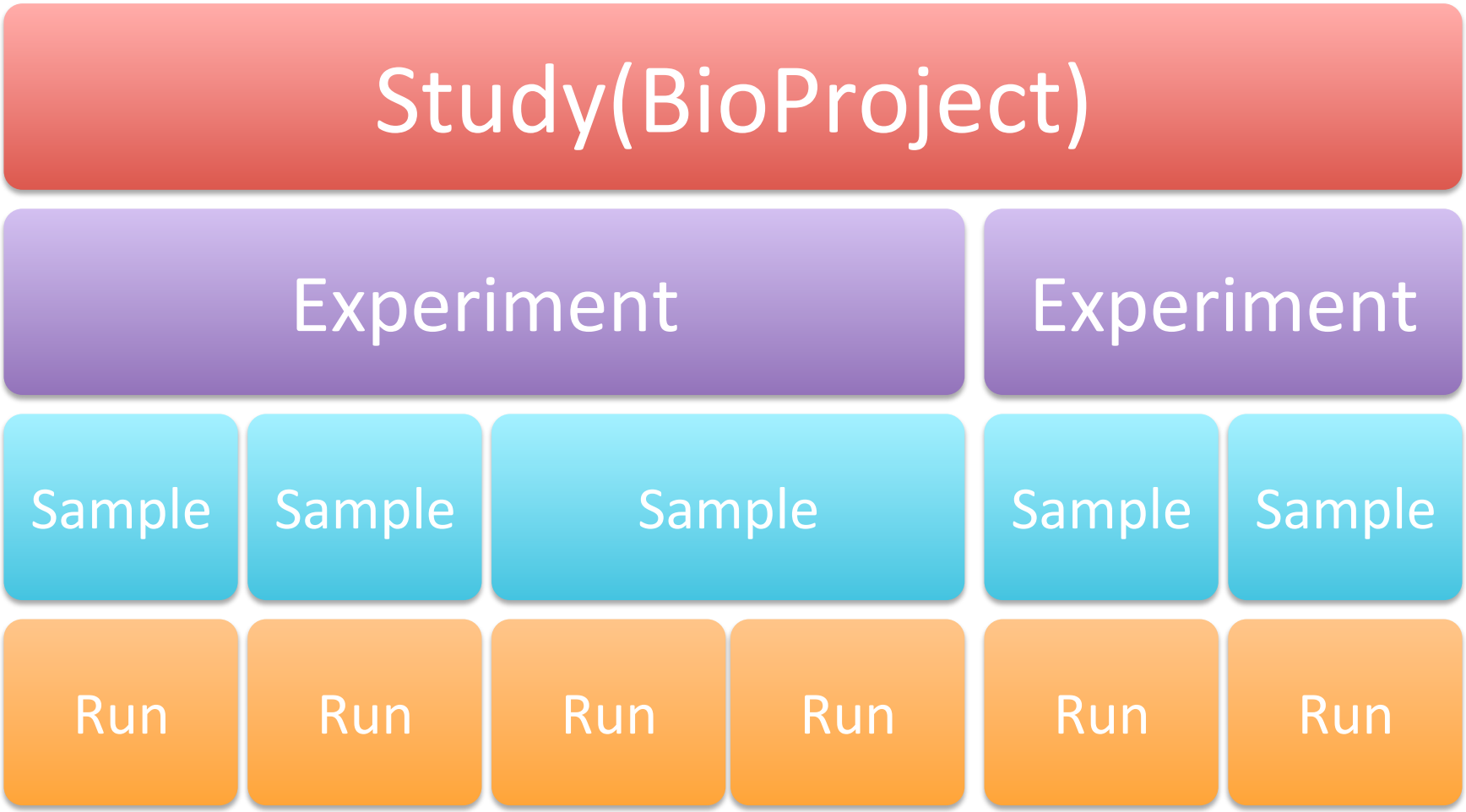
Sequence Read Archive

- 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 (BioProject)** – 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



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.

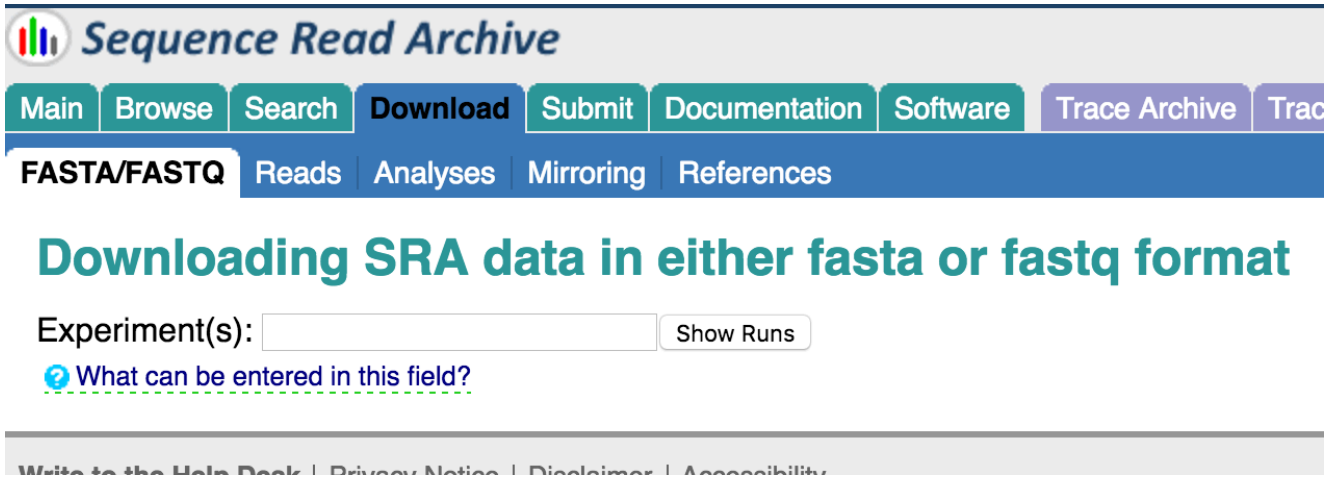
- Biosamples
- Raw reads
- Genome assembly
- Transcriptome assembly
- Genome annotation
- Markers

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

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



SRA Toolkit

- CLI tool for downloading and converting to/ from SRA format
- Most important command:

  - **fastq-dump**: Download SRA data from the internet and convert into fastq format

Submitting to the SRA

- Journals will require that you submit all NGS data to SRA (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)

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.

Submitting to the SRA

Before you begin

Gather information

Why did you perform your analysis?

- Project title and abstract
- Aims and objectives
- Organism(s) sequenced
- Optional: Funding sources, publications, etc.

What did you sequence?

- Descriptive sample information
- Tabular format is ideal
- Examples: Organism(s), age(s), gender(s), location data, cell line(s), etc.

How did you sequence your samples?

- Sequencing methods
- Kits used
- Instrument model(s)

What is your data file format?

- Files in acceptable format(s): BAM, FASTQ, etc.
- MD5 checksum for each file
- Minimum of 1 unique dataset per sample

Register metadata

BioProject → **SRA Study**

- A description of the research effort
- "Why" you sequenced your samples

BioSample → **SRA Sample**

- A description of biologically or physically unique specimens
- "What" you sequenced

Provide technical details

SRA Experiment

- A description of a sample-specific sequencing library
- "How" you performed the sequencing
- Multiple Experiments can "point" to a single Sample, but not vice-versa

Transmit data files

SRA Run

- All files linked to a Run are "merged" into a single dataset
- Files are converted to SRA format
- Files submitted by FTP or Aspera once steps 1 and 2 are complete

Castanea mollissima strain:Vanuxem (Chinese chestnut)

Accession: PRJNA266703 ID: 266703

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

European Nucleotide Archive

- www.ebi.ac.uk/ena
- Mirrors all data in NCBI
- Can submit/download from here
- They use fastq format – no conversion needed!
- Can be slower for large datasets b/c they are far away

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