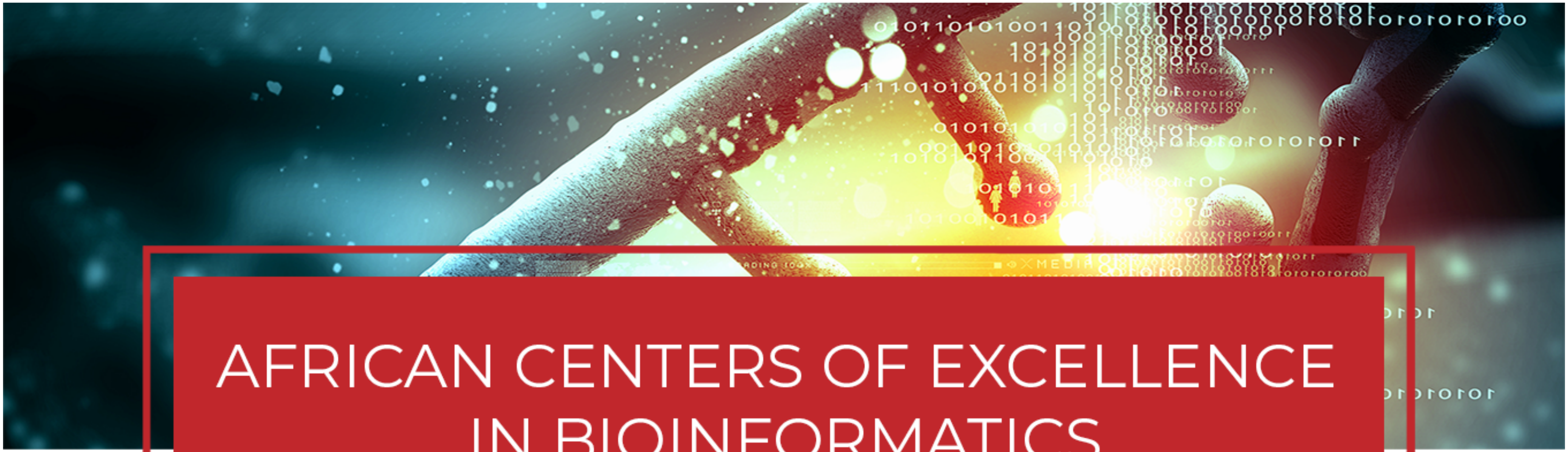


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

SUPPLEMENTAL TRAINING IN BIOINFORMATICS
Primary Sequence Databases
(Practical learning - March 2021)

Today’s Instructor

Mariam Quinones, Ph.D.

Computational Biologist

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor’s email:
    - mariam.quinones@nih.gov

Objectives

- Serve as supplemental training in bioinformatics to students enrolled in course MSB 7104: Online Bioinformatics and Sequence Database (Kampala, Uganda).

- Provide additional background content to help the student get familiar with databases of wide interest

- Describe and use functionality for search and download data of various file types for analysis

Agenda

Part I: Lecture and practical

1. Review of general concepts on sequence databases
2. Get familiar with methods for searching and download of data from Primary databases
3. Participate in practical exercises for working with primary databases (NCBI, EMBL)

Part II: Mostly practical

1. Explore specialized databases for microbial and non-mammalian organisms
2. Use database tools for comparative genomics

Sequence Databases

Primary Sequence / Genome Databases

(Databases that receive, archive and share nucleotide and protein sequence databases derived from experiments. Some also provide tools for mining and analysis)

1. NCBI (GenBank, SRA) , EMBL-EBI, DDBJ (nucleotide sequence) - all three are part of INSDC
2. ArrayExpress and GEO (functional genomics data)
3. Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)

Secondary databases

(Databases that use data from Primary Databases for analysis, annotation, curation, visualization and more)

1. InterPro (protein families, motifs and domains)
2. UniProt Knowledgebase (sequence and functional information on proteins)
3. Ensembl (variation, function, regulation and more layered onto whole genome sequences)

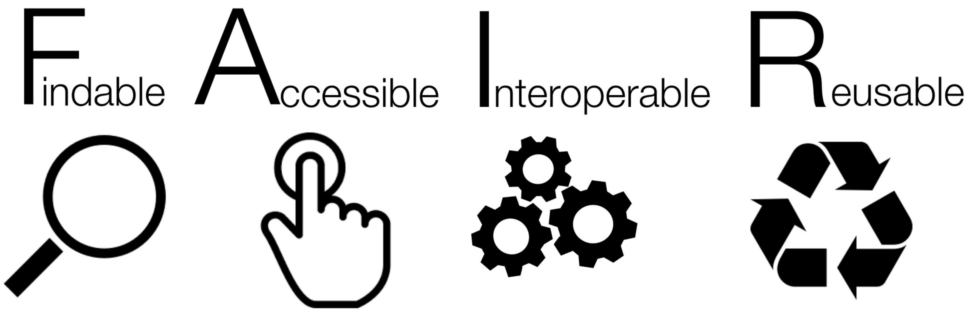
Hybrid databases and families of databases

These are databases that receive experimental data but also perform curation or analysis (e.g. UniProt)

It’s important to share data for reuse

FAIR Principles

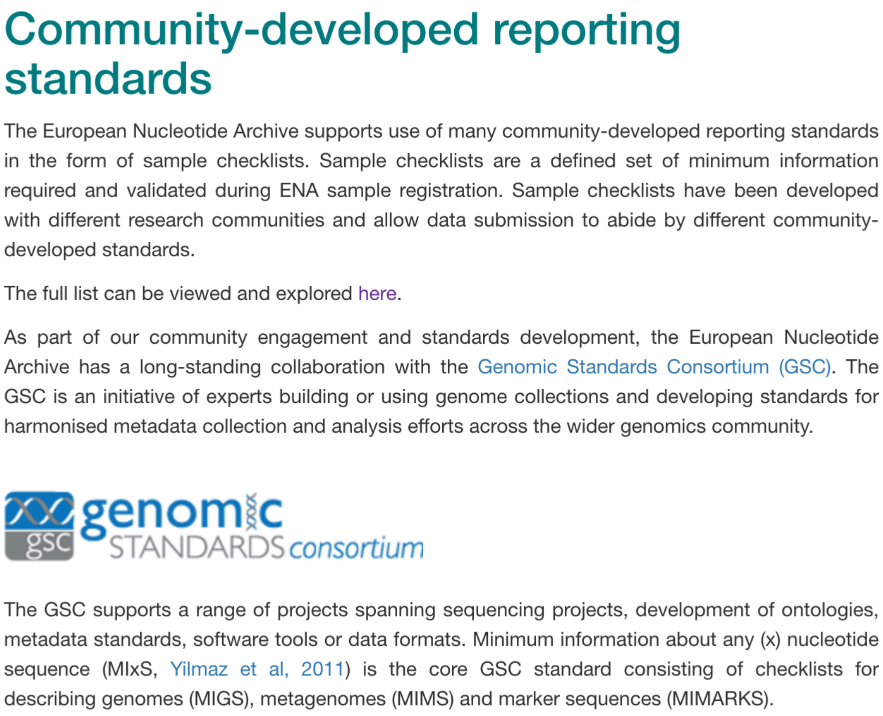
Learn more in:
https://www.go-fair.org/fair-principles/
https://fairsharing.org/



Collaborations such as the INSDC (International Nucleotide Sequence Database Collaborations) and the UniProt Consortium are examples of large collaborative projects making sequence data FAIR. Other projects such as the Human Microbiome Project the Human Genome Project benefit from the collaborations established but also make data FAIR.

https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg#/media/File:FAIR_data_principles.jpg

In order to make data FAIR, Standards are necessary!



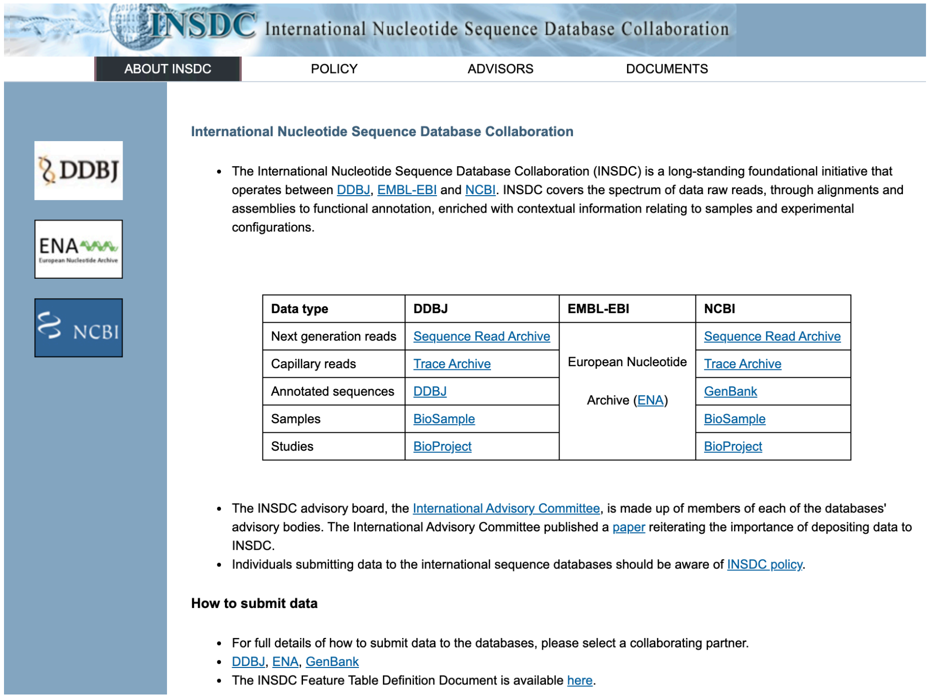
Community-developed reporting standards

The European Nucleotide Archive supports use of many community-developed reporting standards in the form of sample checklists. Sample checklists are a defined set of minimum information required and validated during ENA sample registration. Sample checklists have been developed with different research communities and allow data submission to abide by different community-developed standards.

The full list can be viewed and explored here.

As part of our community engagement and standards development, the European Nucleotide Archive has a long-standing collaboration with the Genomic Standards Consortium (GSC). The GSC is an initiative of experts building or using genome collections and developing standards for harmonised metadata collection and analysis efforts across the wider genomics community.

The GSC supports a range of projects spanning sequencing projects, development of ontologies, metadata standards, software tools or data formats. Minimum information about any (x) nucleotide sequence (MIxS, Yilmaz et al, 2011) is the core GSC standard consisting of checklists for describing genomes (MIGS), metagenomes (MIMS) and marker sequences (MIMARKS).



INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
DDBJ
ENA
NCBI

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.

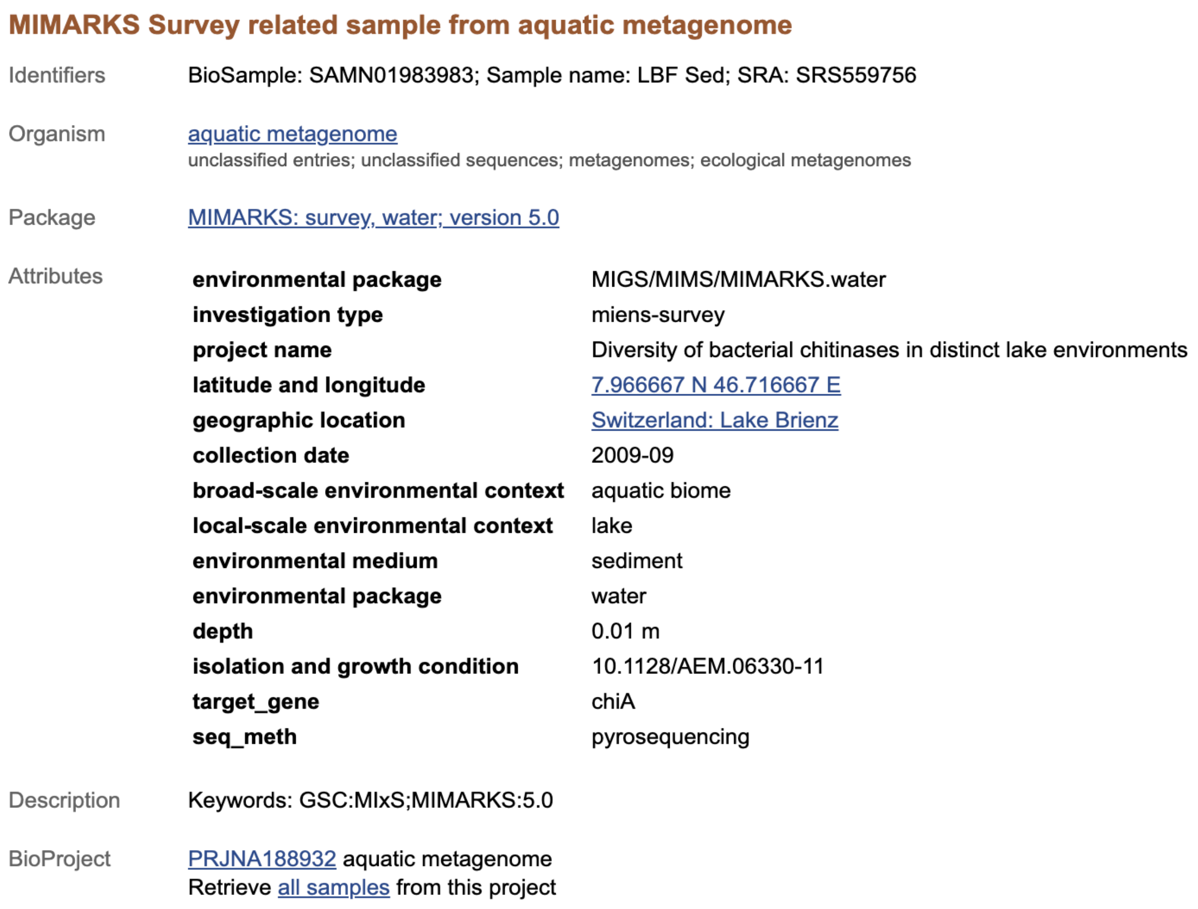
Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- The INSDC advisory board, the International Advisory Committee, is made up of members of each of the databases' advisory bodies. The International Advisory Committee published a paper reiterating the importance of depositing data to INSDC.
- Individuals submitting data to the international sequence databases should be aware of INSDC policy.

How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- DDBJ, ENA, GenBank
- The INSDC Feature Table Definition Document is available here.

BioSample database uses the GSC standards



MIMARKS Survey related sample from aquatic metagenome

Identifiers BioSample: SAMN01983983; Sample name: LBF Sed; SRA: SRS559756

Organism aquatic metagenome
unclassified entries; unclassified sequences; metagenomes; ecological metagenomes

Package MIMARKS: survey, water; version 5.0

Attributes

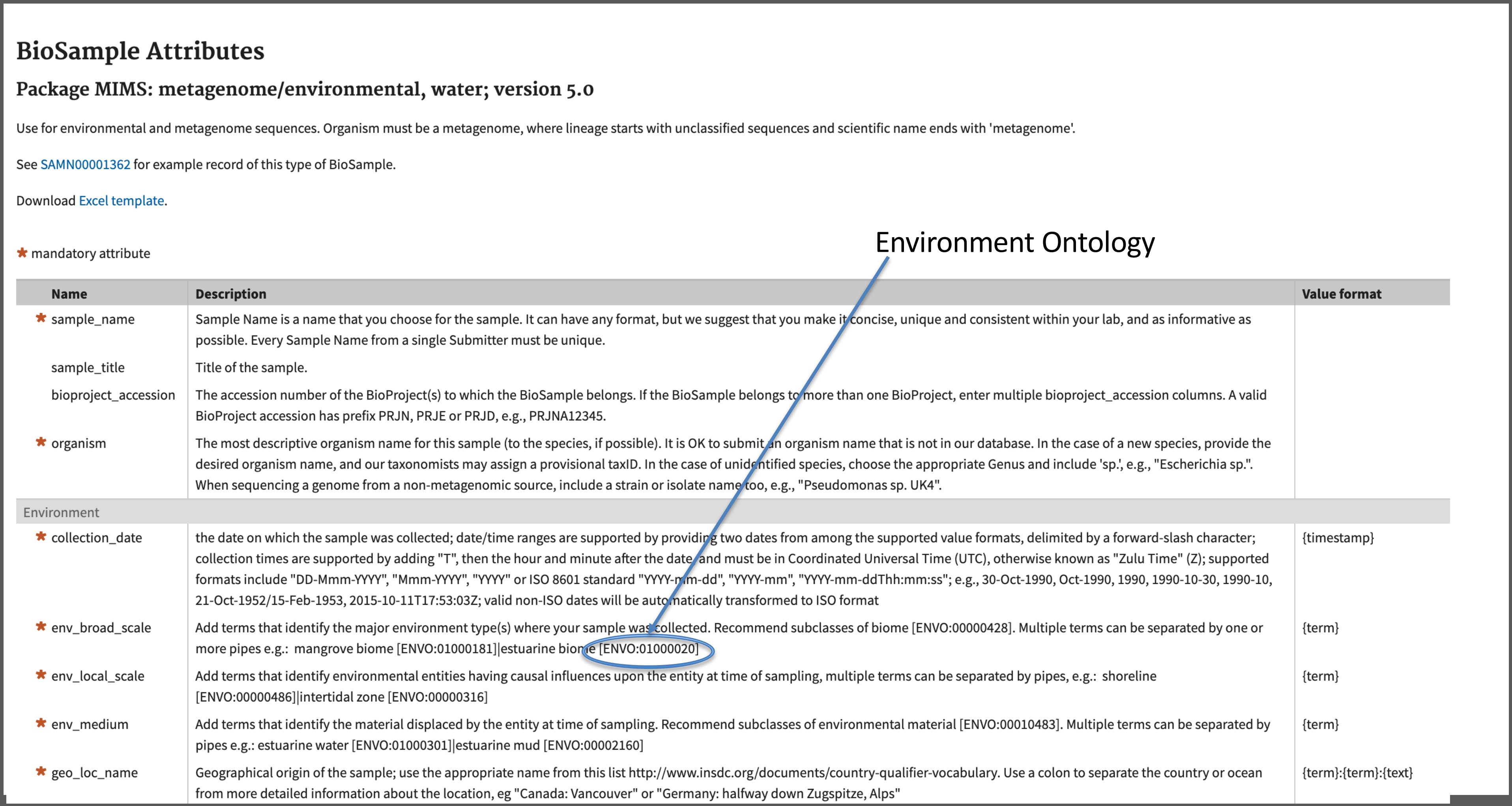
environmental package	MIGS/MIMS/MIMARKS.water
investigation type	miens-survey
project name	Diversity of bacterial chitinases in distinct lake environments
latitude and longitude	7.966667 N 46.716667 E
geographic location	Switzerland: Lake Brienz
collection date	2009-09
broad-scale environmental context	aquatic biome
local-scale environmental context	lake
environmental medium	sediment
environmental package	water
depth	0.01 m
isolation and growth condition	10.1128/AEM.06330-11
target_gene	chiA
seq_meth	pyrosequencing

Description Keywords: GSC:MIxS;MIMARKS:5.0

BioProject PRJNA188932 aquatic metagenome
Retrieve all samples from this project

https://www.ncbi.nlm.nih.gov/biosample/1983983

Description of attributes



BioSample Attributes

Package MIMS: metagenome/environmental, water; version 5.0

Use for environmental and metagenome sequences. Organism must be a metagenome, where lineage starts with unclassified sequences and scientific name ends with 'metagenome'.

See SAMN00001362 for example record of this type of BioSample.

Download Excel template.

* mandatory attribute

Name	Description	Value format
* sample_name	Sample Name is a name that you choose for the sample. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible. Every Sample Name from a single Submitter must be unique.	
sample_title	Title of the sample.	
bioproject_accession	The accession number of the BioProject(s) to which the BioSample belongs. If the BioSample belongs to more than one BioProject, enter multiple bioproject_accession columns. A valid BioProject accession has prefix PRJN, PRJE or PRJD, e.g., PRJNA12345.	
* organism	The most descriptive organism name for this sample (to the species, if possible). It is OK to submit an organism name that is not in our database. In the case of a new species, provide the desired organism name, and our taxonomists may assign a provisional taxID. In the case of unidentified species, choose the appropriate Genus and include 'sp.', e.g., "Escherichia sp.". When sequencing a genome from a non-metagenomic source, include a strain or isolate name too, e.g., "Pseudomonas sp. UK4".	
Environment		
* collection_date	the date on which the sample was collected; date/time ranges are supported by providing two dates from among the supported value formats, delimited by a forward-slash character; collection times are supported by adding "T", then the hour and minute after the date, and must be in Coordinated Universal Time (UTC), otherwise known as "Zulu Time" (Z); supported formats include "DD-Mmm-YYYY", "Mmm-YYYY", "YYYY" or ISO 8601 standard "YYYY-mm-dd", "YYYY-mm", "YYYY-mm-ddThh:mm:ss"; e.g., 30-Oct-1990, Oct-1990, 1990, 1990-10-30, 1990-10, 21-Oct-1952/15-Feb-1953, 2015-10-11T17:53:03Z; valid non-ISO dates will be automatically transformed to ISO format	{timestamp}
* env_broad_scale	Add terms that identify the major environment type(s) where your sample was collected. Recommend subclasses of biome [ENVO:00000428]. Multiple terms can be separated by one or more pipes e.g.: mangrove biome [ENVO:01000181]|estuarine biome [ENVO:01000020]	{term}
* env_local_scale	Add terms that identify environmental entities having causal influences upon the entity at time of sampling, multiple terms can be separated by pipes, e.g.: shoreline [ENVO:00000486]|intertidal zone [ENVO:00000316]	{term}
* env_medium	Add terms that identify the material displaced by the entity at time of sampling. Recommend subclasses of environmental material [ENVO:00010483]. Multiple terms can be separated by pipes e.g.: estuarine water [ENVO:01000301]|estuarine mud [ENVO:00002160]	{term}
* geo_loc_name	Geographical origin of the sample; use the appropriate name from this list http://www.insdc.org/documents/country-qualifier-vocabulary. Use a colon to separate the country or ocean from more detailed information about the location, eg "Canada: Vancouver" or "Germany: halfway down Zugspitze, Alps"	{term}:{term}:{text}

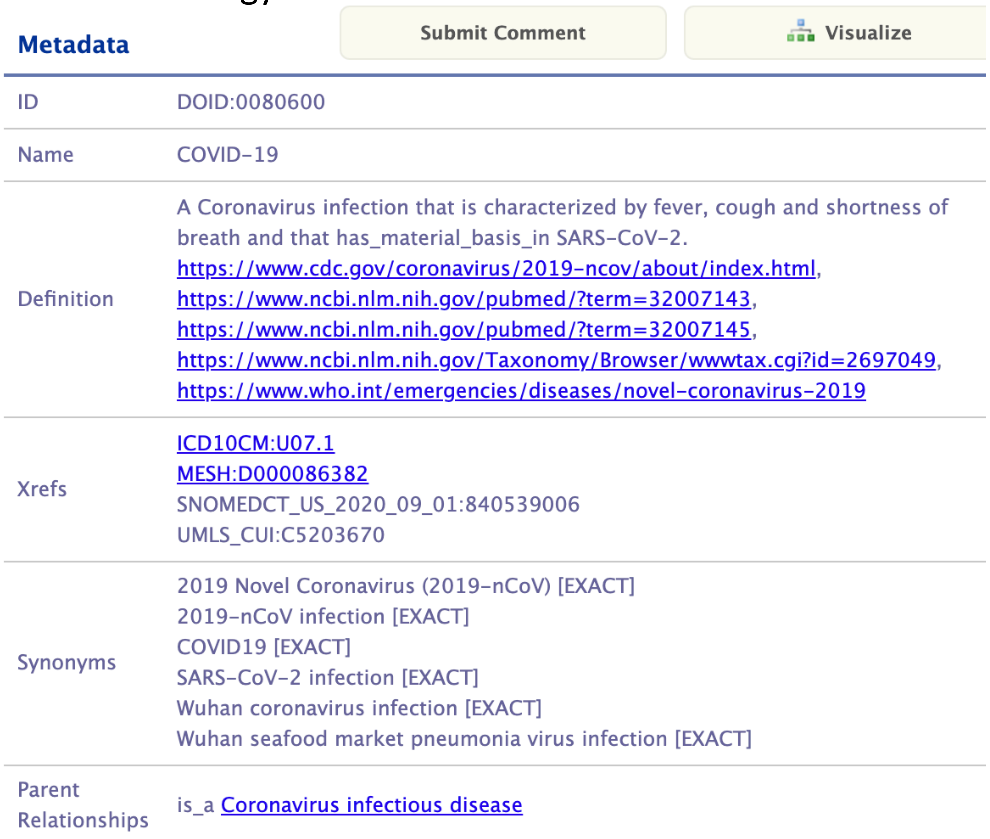
Environment Ontology

Standards leverage many domain specific ontologies

1. Experimental Factor Ontology (EFO)
2. Biomedical Investigation Ontology (OBI)
3. Information Artifact Ontology (IAO)
4. Environment Ontology
5. NCBI Taxonomy
6. Chemical Entities of Biological Interest (ChEBI)
7. Disease Ontology
8. …. Many others

Do we need controlled vocabulary?
Play Game Here

Disease Ontology



Metadata	
ID	DOID:0080600
Name	COVID-19
Definition	A Coronavirus infection that is characterized by fever, cough and shortness of breath and that has_material_basis_in SARS-CoV-2. https://www.cdc.gov/coronavirus/2019-ncov/about/index.html, https://www.ncbi.nlm.nih.gov/pubmed/?term=32007143, https://www.ncbi.nlm.nih.gov/pubmed/?term=32007145, https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2697049, https://www.who.int/emergencies/diseases/novel-coronavirus-2019
Xrefs	ICD10CM:U07.1 MESH:D000086382 SNOMEDCT_US_2020_09_01:840539006 UMLS_CUI:C5203670
Synonyms	2019 Novel Coronavirus (2019-nCoV) [EXACT] 2019-nCoV infection [EXACT] COVID19 [EXACT] SARS-CoV-2 infection [EXACT] Wuhan coronavirus infection [EXACT] Wuhan seafood market pneumonia virus infection [EXACT]
Parent Relationships	is_a Coronavirus infectious disease

How to archive sequence data?

Databases have tools for submission of the various types of sequences.

For example:

1- NCBI Submission Portal https://submit.ncbi.nlm.nih.gov/

a) Full Genomes → submit to GenBank using BankIt

b) Raw sequence reads (e.g. fastq) can be submitted via the SRA submission web interface or with tools such as METAGENOTE which use an API to facilitate annotation, validation and generation of SRA records

2- ENA Webin tool facilitates submission to the EMBL-EBI

These submission tools will require or encourage users to provide metadata following standards and ontologies.

JGI
IMG/M
INTEGRATED MICROBIAL GENOMES & I
PATRIC 3.6.9
ViPR
Virus Pathogen Resource
JGI
MycoCosm
THE FUNGAL GENOMICS RESOURCE
WormBase
FungiDB
Release 50 beta
17 Dec 2020
Fungal & Oomycete Informatics Resources
VEuPathDB
Release 50 beta
17 Dec 2020
Eukaryotic Pathogen, Vector & Host Informatics Resources

Examples of Specialized Databases

Microbial (Bacteria and Viruses)

- Integrated Microbial Genomes and Microbiomes (IMG/M)
- PATRIC BACTERIAL BIOINFORMATICS RESOURCE CENTER (PATRIC)
- Virus Pathogen Resource (ViPR)
- Los Alamos HIV Databases

Fungal, Oomycete and Worms

- MycoCosm (e.g. Schleroderma, Aspergillus)
- FungiDB (e.g. Candida, Aspergillus, Cryptococcus)
- Saccharomyces Genome Database (yeast)
- WormBase (e.g. *C. elegans, B. malayi*)

Eukaryotic Pathogens

- VEuPathDB (e.g. Plasmodium, Giardia, Toxoplasma, Acanthamoeba)

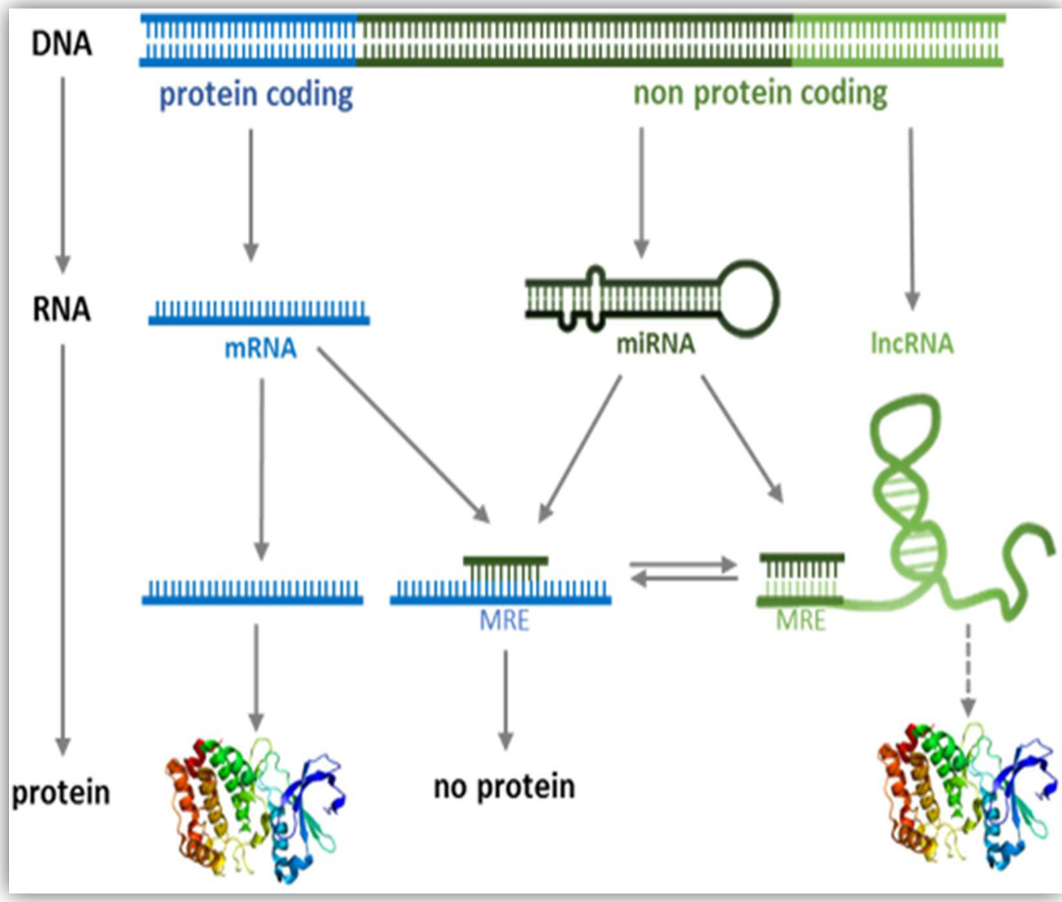
Mouse, Human

- Mouse Genome Informatics (MGD)
- Genome Reference Consortium

What types of sequences are typically stored?

- Raw Sequences (reads from whole genomes, targeted regions, cDNA)
- Assembled sequences (e.g. contigs, genomes, transcripts)
- miRNAs and lncRNA
- Motifs (e.g. TF binding sites, promoters)
- Protein sequences
- Variants (SNVs, INDELS)
- expressed sequence tag (ESTs)

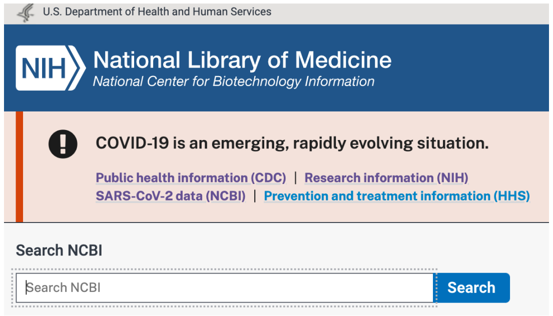
Central dogma



https://www.frontiersin.org/files/Articles/446580/fgene-10-00281-HTML/image_m/fgene-10-00281-g001.jpg

How to initiate a search?

Methods 1: Keyword search (e.g. Gene name/symbol)

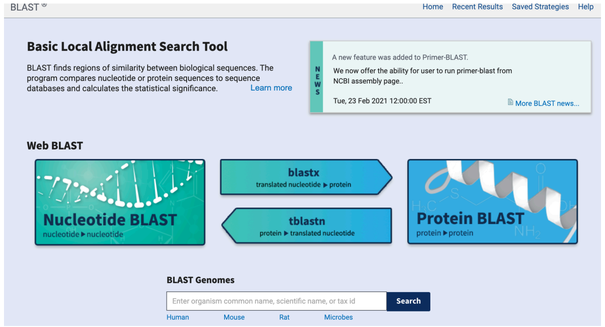

U.S. Department of Health and Human Services
NIH
National Library of Medicine
National Center for Biotechnology Information

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

Search NCBI

Method 2: Sequence search (through BLAST)



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new feature was added to Primer-BLAST.

We now offer the ability for user to run primer-blast from NCBI assembly page..

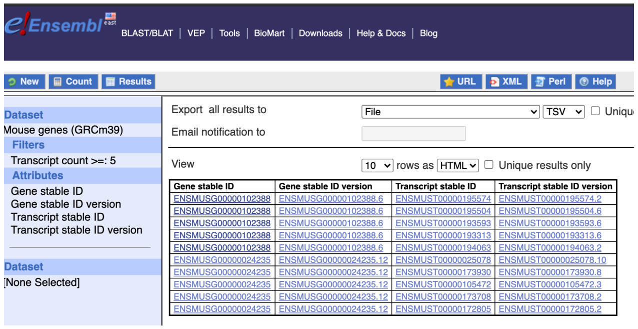
Tue, 23 Feb 2021 12:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

Method 3: Metadata filters (e.g. EMBL Biomart)


e!Ensembl east
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results
URL
XML
Perl
Help

Dataset

Mouse genes (GRCm39)

Filters

Transcript count >=: 5

Attributes

Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version

Dataset

[None Selected]

Export all results to File TSV □ Uniqu

Email notification to

View 10 rows as HTML □ Unique results only

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version
ENSMUSG00000102388	ENSMUSG00000102388.6	ENSMUST00000195574	ENSMUST00000195574.2
ENSMUSG00000102388	ENSMUSG00000102388.6	ENSMUST00000195504	ENSMUST00000195504.6
ENSMUSG00000102388	ENSMUSG00000102388.6	ENSMUST00000193593	ENSMUST00000193593.6
ENSMUSG00000102388	ENSMUSG00000102388.6	ENSMUST00000193313	ENSMUST00000193313.6
ENSMUSG00000102388	ENSMUSG00000102388.6	ENSMUST00000194063	ENSMUST00000194063.2
ENSMUSG00000024235	ENSMUSG00000024235.12	ENSMUST00000025078	ENSMUST00000025078.10
ENSMUSG00000024235	ENSMUSG00000024235.12	ENSMUST00000173930	ENSMUST00000173930.8
ENSMUSG00000024235	ENSMUSG00000024235.12	ENSMUST00000105472	ENSMUST00000105472.3
ENSMUSG00000024235	ENSMUSG00000024235.12	ENSMUST00000173708	ENSMUST00000173708.2
ENSMUSG00000024235	ENSMUSG00000024235.12	ENSMUST00000172805	ENSMUST00000172805.2

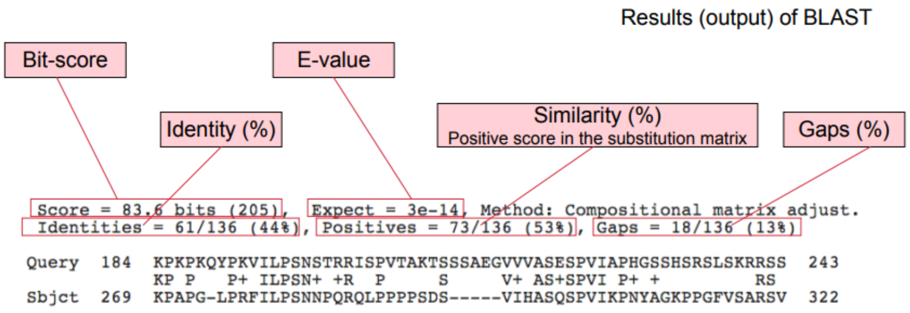
A preview on BLAST

Reference: Chapter 9BLAST QuickStart
https://www.ncbi.nlm.nih.gov/books/NBK1734/

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between protein or nucleotide sequences. The program compares nucleotide or protein sequences to sequence in a database and calculates the statistical significance of the matches.

The alignment score is computed by assigning a value to each aligned pair of letters and then summing these values over the length of the alignment. For protein sequence alignments, scores for every possible amino acid letter pair are given in a “substitution matrix” where likely substitutions have positive values and unlikely substitutions have negative values. By default, BLAST uses the “blosum62” matrix

Example: BLAST - Pho4p (*S. cerevisiae*)


Results (output) of BLAST
Bit-score
E-value
Identity (%)
Similarity (%)
Positive score in the substitution matrix
Gaps (%)

```
 Score = 83.6 bits (205),  Expect = 3e-14, Method: Compositional matrix adjust.
 Identities = 61/136 (44%), Positives = 73/136 (53%), Gaps = 18/136 (13%)

Query  184  KPKPKQYPKVILPSNSTRRISPVTAKTSSSAEGVVVASESPVIAPHGSSHSRSLSKRRSS  243
            KP P   P+ ILPSN+ +R  P     S     V+ AS+SPVI P+ +         RS
Sbjct  269  KPAPG-LPRFILPSNNPQRQLPPPPSDS-----VIHASQSPVIKPNYAGKPPGFVSARSV  322
```

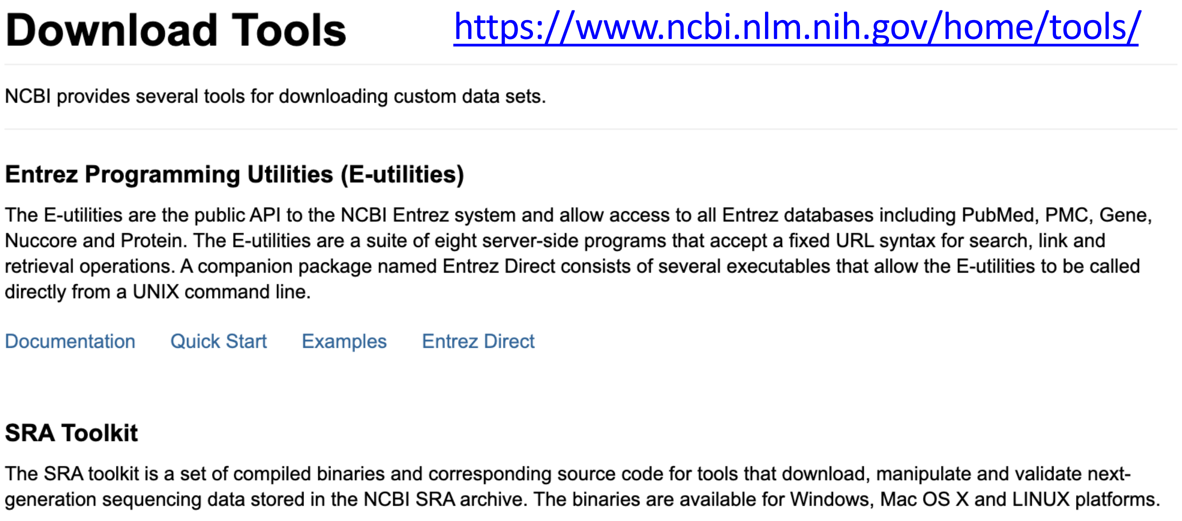
The “**Expect Value**” is the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the size of the database searched. “Expect Values” in the range of 0.001 to 0.0000001 are commonly used to restrict the alignments shown to those of high quality.

How to download data in batch?

- For each database of interest, read the documentation and search for APIs and utilities developed to facilitate data download.

- For example, in NCBI, use E-utilities (Entrez databases) or SRA Toolkit (raw reads)

Note: We will practice download of sequences towards the end of the course.



https://www.ncbi.nlm.nih.gov/home/tools/

Download Tools

NCBI provides several tools for downloading custom data sets.

Entrez Programming Utilities (E-utilities)

The E-utilities are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene, Nuccore and Protein. The E-utilities are a suite of eight server-side programs that accept a fixed URL syntax for search, link and retrieval operations. A companion package named Entrez Direct consists of several executables that allow the E-utilities to be called directly from a UNIX command line.

SRA Toolkit

The SRA toolkit is a set of compiled binaries and corresponding source code for tools that download, manipulate and validate next-generation sequencing data stored in the NCBI SRA archive. The binaries are available for Windows, Mac OS X and LINUX platforms.

Exercises

Part 1 – Practical exercises

a) NCBI databases: Gene and Protein sequences – [Respond to Quiz here]

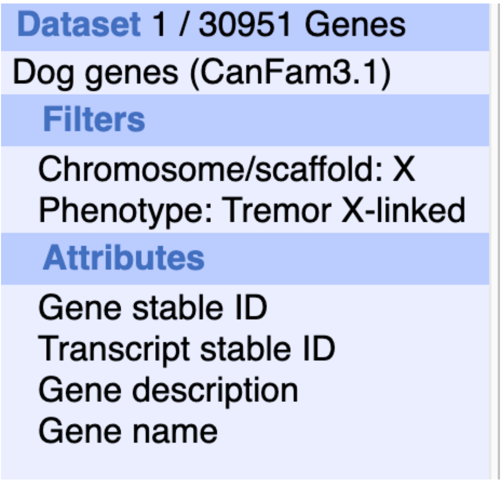
b) EMBL-EBI resources – explore https://www.ebi.ac.uk/ and compare to NCBI

c) e!Ensembl databases: http://useast.ensembl.org/index.html

  a) BioMart – view demo of generating table with all dog genes with phenotype

  b) Perform the search described here and provide the:

  Gene name ________



Dataset 1 / 30951 Genes
Dog genes (CanFam3.1)

Filters
Chromosome/scaffold: X
Phenotype: Tremor X-linked

Attributes
Gene stable ID
Transcript stable ID
Gene description
Gene name

Training Resources for NCBI and EBI

1. EBI Course: Bioinformatics for the terrified

   Guided Exercises:

   a. Search EBI

   b. Comparing sequences

2. NCBI Courses

   a. Types of Databases (old course but useful)

   b. Webinars

Questions?

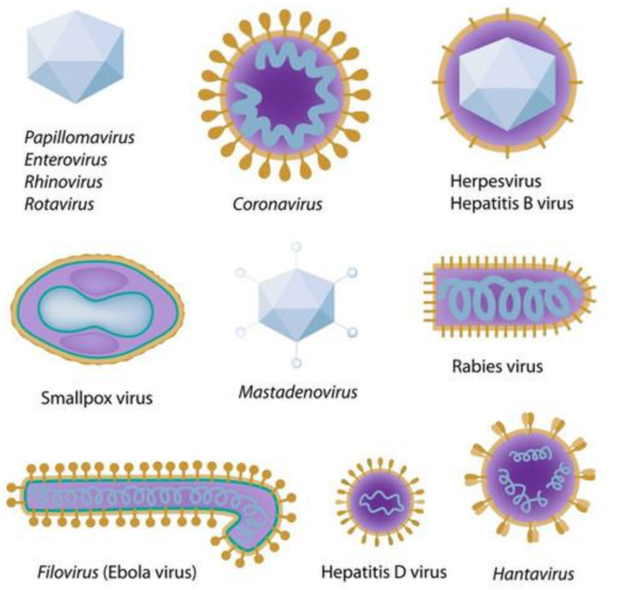
Email us: bioinformatics@niaid.nih.gov

mariam.quinones@nih.gov

Part II: Specialized Sequence Databases

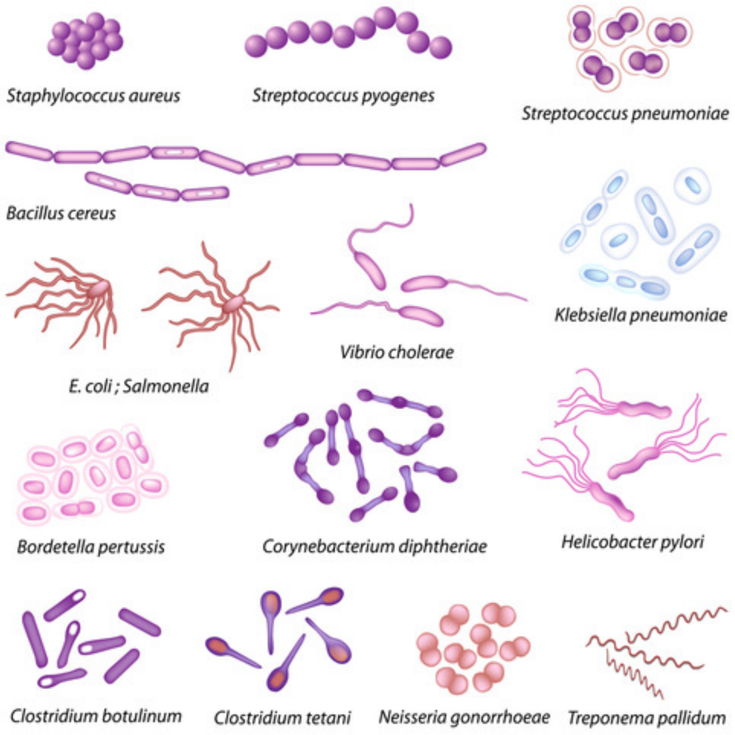
1- Genomes from prokaryotic organisms (e.g. archea, bacteria, microbiome, uncultivated virus, viral genomes, bacteriophages)

PATRIC



Morphology of common viruses that infect humans.

https://www.ck12.org/book/ck-12-biology-advanced-concepts/section/11.24/

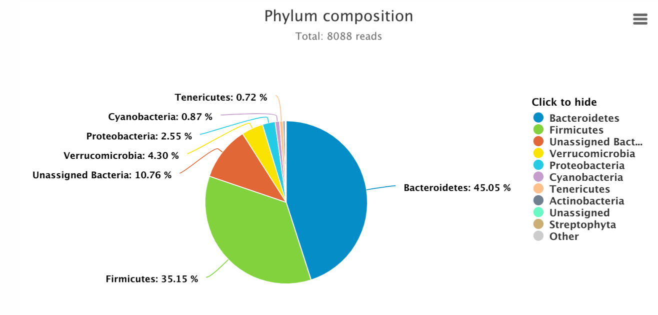


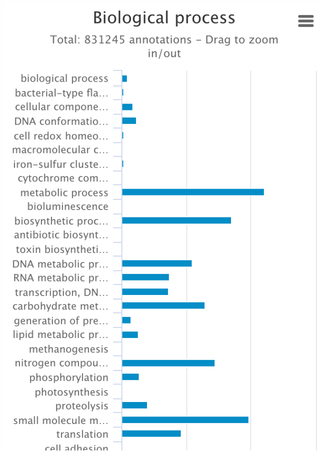
Which types of sequences are stored?

- Chromosomes
- Bacterial Plasmids
- Viral RNA genomes
- Metagenomes
- Protein / proteome

Part II: Specialized Sequence Databases

2- Metagenomes from an animal host or environment



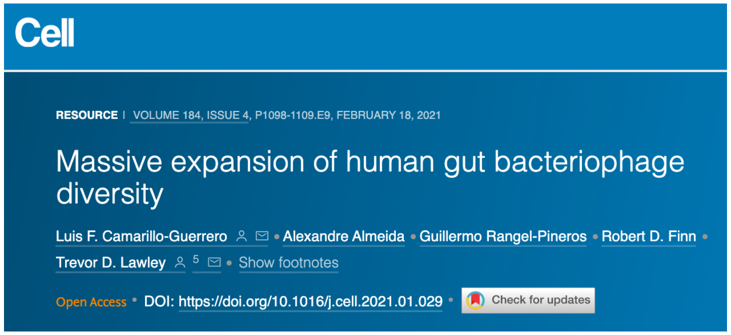


Example: Human stool microbiome sample
https://www.ebi.ac.uk/metagenomics/analyses/MGYA00581604

https://www.ck12.org/book/ck-12-biology-advanced-concepts/section/11.24/

The great diversity of viral and bacterial species requires a great effort of curation

New bacteriophage database! – Feb 2021



Cell

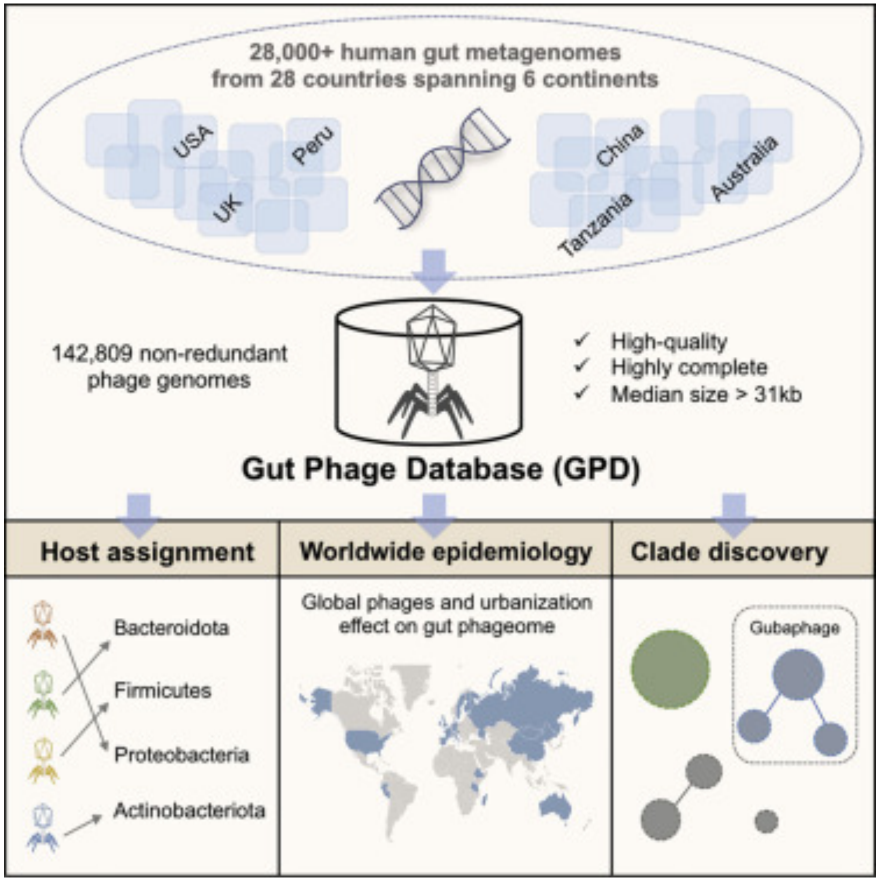
RESOURCE | VOLUME 184, ISSUE 4, P1098-1109.E9, FEBRUARY 18, 2021

Massive expansion of human gut bacteriophage diversity

Luis F. Camarillo-Guerrero • Alexandre Almeida • Guillermo Rangel-Pineros • Robert D. Finn • Trevor D. Lawley 5 • Show footnotes

Open Access • DOI: https://doi.org/10.1016/j.cell.2021.01.029 • Check for updates

Gut Phage Database, a collection of ~142,000 non-redundant viral genomes (>10 kb) obtained by mining a dataset of 28,060 globally distributed human gut metagenomes and 2,898 reference genomes of cultured gut bacteria.



DOI:https://doi.org/10.1016/j.cell.2021.01.029

Part II: Specialized Sequence Databases

3- Genomes from eukaryotic pathogens (parasites and fungi) – see VEuPathDB.org

FungiDB
Fungal & Oomycete Informatics Resources

AmoebaDB
Amoeba Informatics Resources

MicrosporidiaDB
Microsporidia Informatics Resources

GiardiaDB
Giardia Informatics Resources

PlasmoDB
Plasmodium Informatics Resources

CryptoDB
Cryptosporidium Informatics Resources

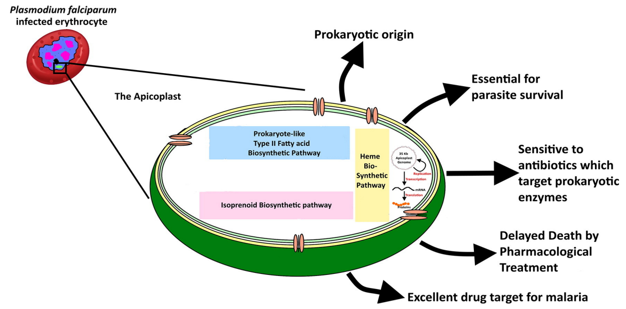
TriTrypDB
Kinetoplastid Informatics Resources

TrichDB
Trichomonas Informatics Resources

Which types sequences are stored?

- Chromosomes
- Yeast plasmids
- Apicoplast genomes

Parasites belonging to the apicomplexa which infect animals and humans include *Toxoplasma* and *Plasmodium*, and the genera *Eimeria*, *Isospora*, *Cyclospora*, *Babesia*, *Cryptosporidium*, *Theileria*, and *Sarcocystis*.


Plasmodium falciparum infected erythrocyte
The Apicoplast
Prokaryotic origin
Essential for parasite survival
Prokaryote-like Type II Fatty acid Biosynthetic Pathway
Heme Bio-Synthetic Pathway
35 Kb Apicoplast Genome
Replication
Transcription
mRNA
Translation
Proteins
Isoprenoid Biosynthetic pathway
Sensitive to antibiotics which target prokaryotic enzymes
Delayed Death by Pharmacological Treatment
Excellent drug target for malaria

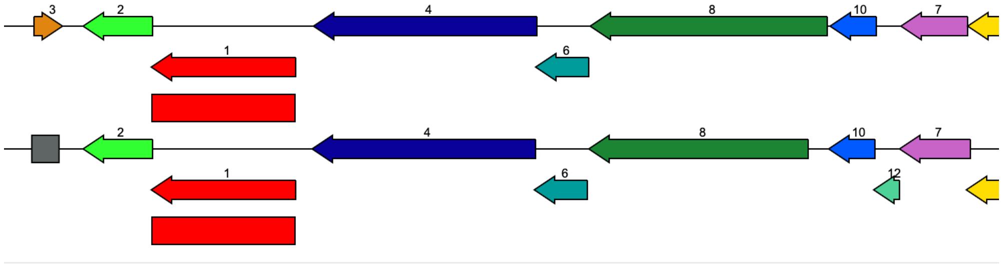
https://www.sciencedirect.com/science/article/abs/pii/S0024320516303861

Exercises

Part 2 – Practical exercises

a) Explore Bacterial genomes using PATRIC

1. Search for *Vibrio cholerae*, str. N16961
2. How many chromosomes? ____ Any plasmids? ____
3. Are there bacteriophages genes? (Select “Features” and use keyword “phage”)
4. Are there any genes encoding a toxin? (Select “Features” and use keyword “toxin”), type Enterotoxin, A, then click on F FEATURE to explore the Feature View page.
5. From the Feature View page, go to tab “Compare Region Viewer” and see which other *V. choleare* strain has a similar arrangement of toxin genes.

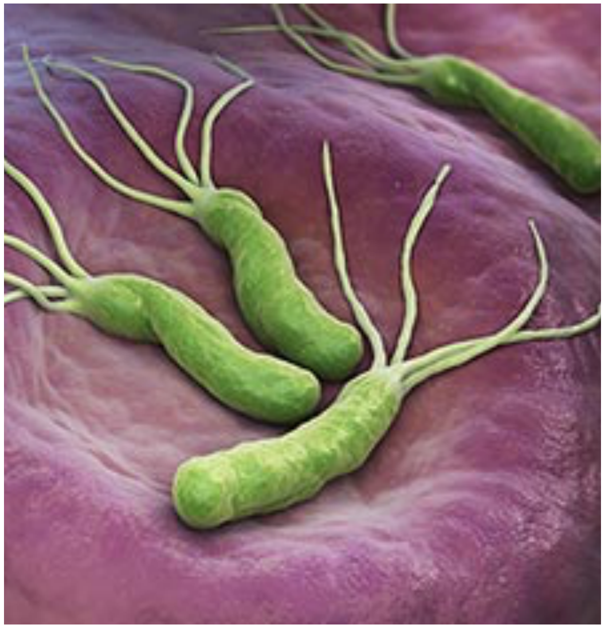


Exercises

- Part 2 – Practical exercises

  b) Explore Bacterial genomes using PATRIC

  - Search for *Helicobacter pylori, click on Genomes.*
  - Add a column for Plasmids. Which strain isolated from Mexico has 3 plasmids? ___________
  - How long are the plasmids in kbp (go to Sequences tab)? ______, ____, ____
  - Which potential virulence factors are present in the longest plasmid?



Exercises

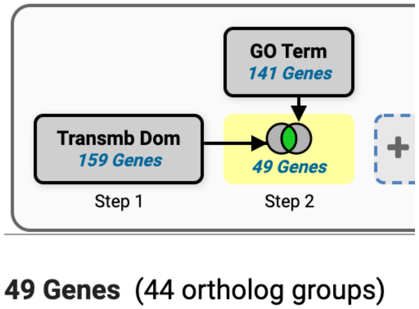
Part 2 – Practical exercises

c) Explore PlasmoDB's functionality

1. Search for *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) gene from strain 3D7 (quick look: https://plasmodb.org/plasmo/app/record/gene/PF3D7_0100100)
2. What is the gene symbol _________
3. Explore the gene page. How many exons? ______How long is the transcript? ______

d) Use Search Strategies in PlasmoDB to filter for genes of interest

1. Find genes in *P. falciparum* with transmembrane domains using search menu and keyword "transmembrane". Filter for minimum of 8 transmembrane domains
2. Add a step and select Function prediction, limit for transporter activity



49 Genes (44 ortholog groups)

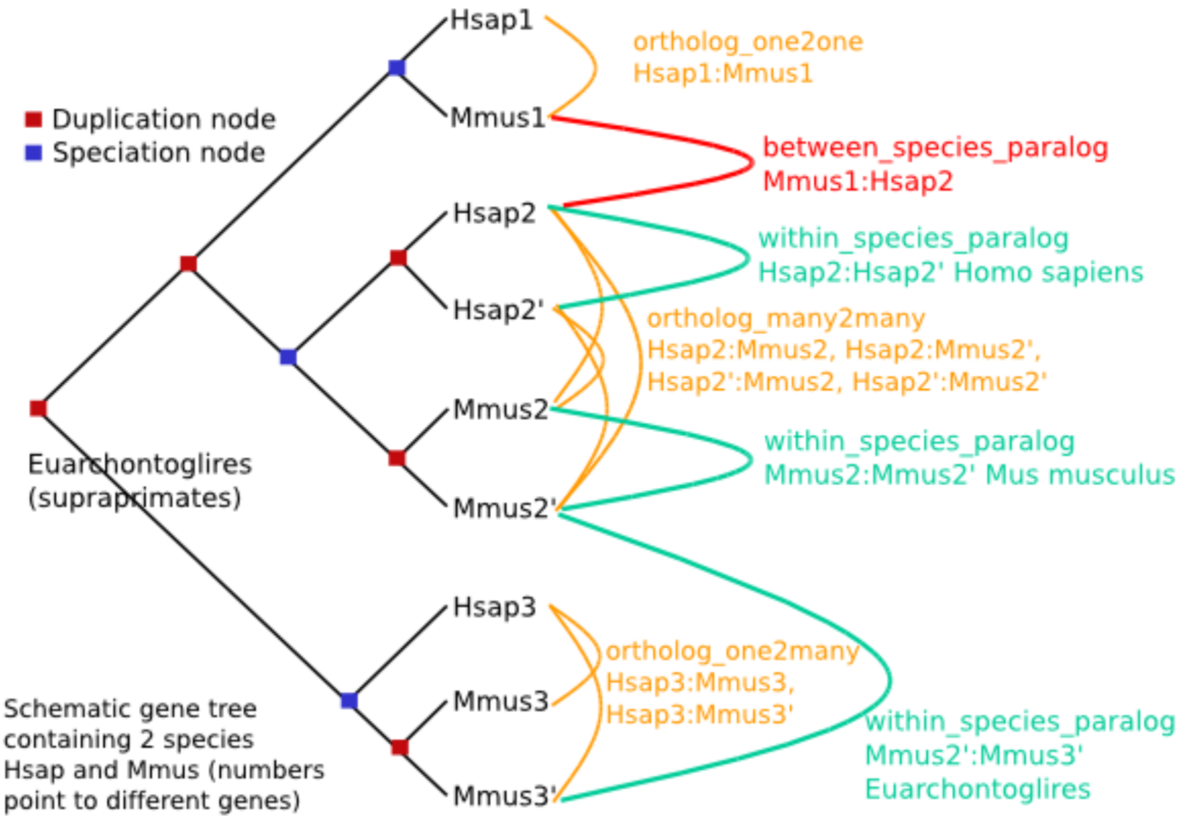
Reference: https://static-content.veupathdb.org/documents/SearchStrategies.pdf

Reference material

Homology types

Orthologs are genes in different species evolved from a common ancestral gene.

Paralogs are gene copies created by a duplication event within the same genome.



Schematic gene tree containing 2 species Hsap and Mmus (numbers point to different genes)

ensembl.org