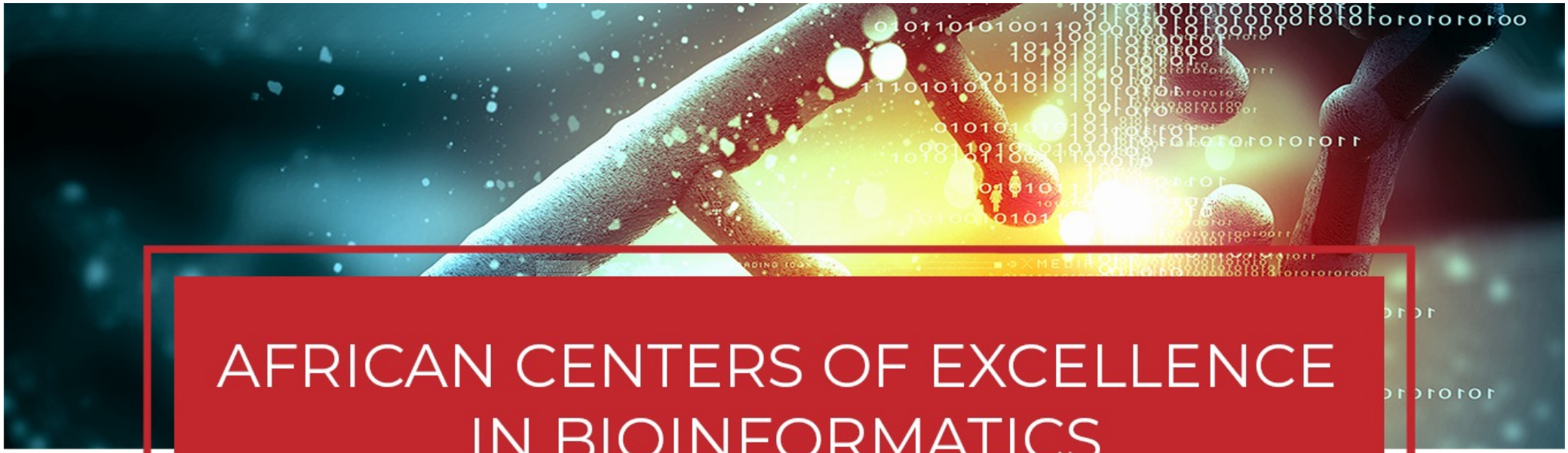


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

SUPPLEMENTAL TRAINING IN BIOINFORMATICS
MSB 7104: Online Bioinformatics and Sequence Database
Topic: Advanced Literature Search (**Sequence Read Archive**)
(Practical learning - April 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).
- Demonstrate methods for search and retrieval of sequence files from public databases
- Learn how to download files in batch from NCBI
- Overview of how to upload files to NCBI SRA (submission portal and METAGENOTE)

Potential reasons for downloading public data files

- To expand previous published analysis (for example searching for genomic elements that were not the original focus of the study)
- To reproduce a study, compare or integrate with additional data files
- For gaining experience with file manipulation and processing
- For benchmarking certain tools

Where are sequence files stored?

1. Sequence Read Archive (SRA)

   The SRA is NIH's primary archive of high-throughput sequencing data and is part of the International Nucleotide Sequence Database Collaboration (INSDC) that includes at the NCBI Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ). Data submitted to any of the three organizations are shared among them. https://www.ncbi.nlm.nih.gov/sra/docs/

2. Specialized Databases

   For example: MG-RAST

Which file types are stored at SRA?

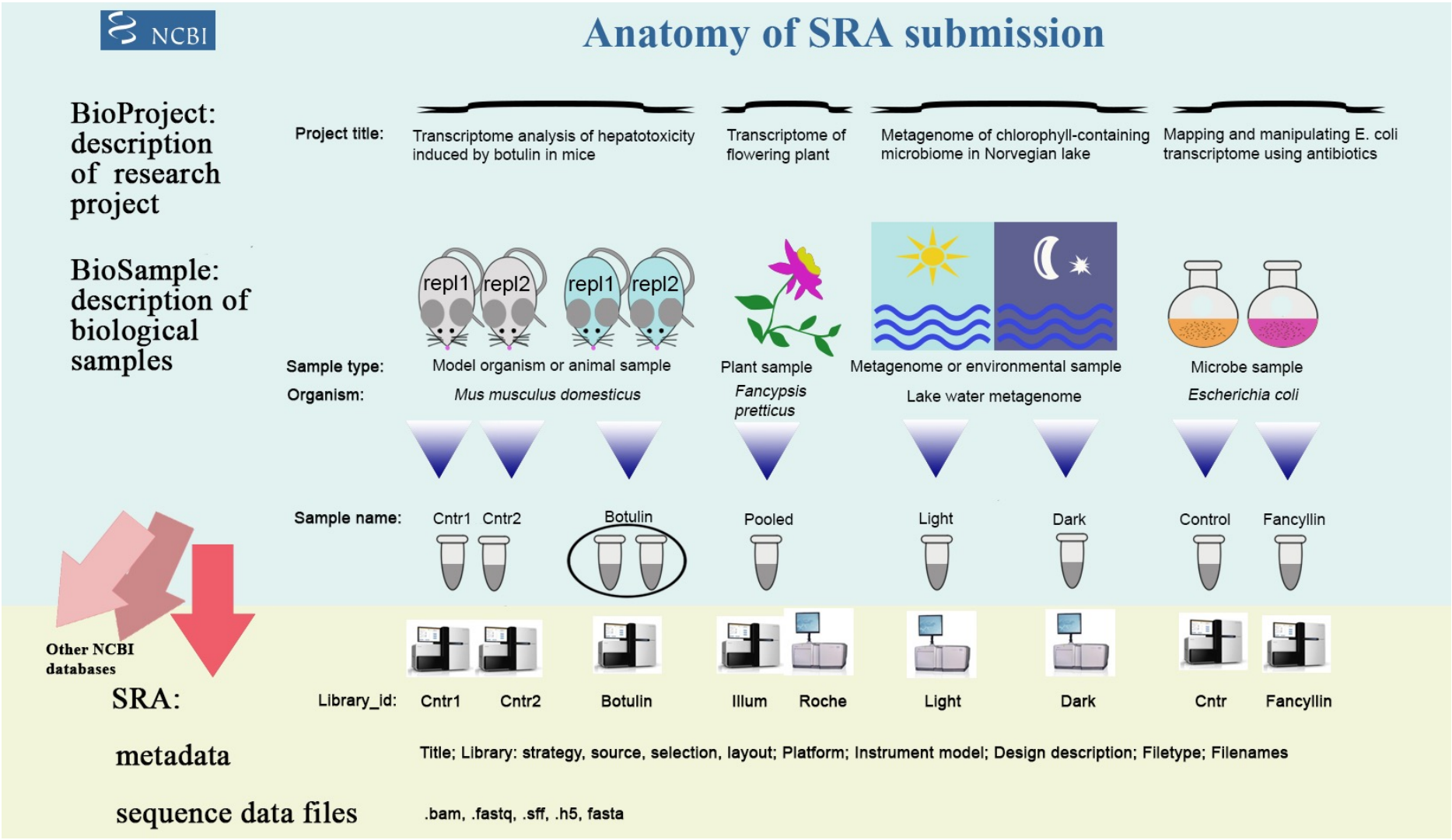
The SRA accepts **genetic data and the associated quality scores** produced by next generation sequencing technologies.

Typically FASTQ or BAM files

https://www.ncbi.nlm.nih.gov/sra/docs/submit/#accepted-data

File Format Guide

- Introduction
- BAM files
- CRAM files
- SFF files
- HDF5 files
  - PacBio
  - MinION Oxford Nanopore
  - HDF5 tools
- FASTQ files
  - Paired-end FASTQ
  - Platform specific FASTQ files
    - 454 fastq
    - Ion Torrent fastq
    - Recent Illumina fastq
    - Older Illumina fastq
    - QIIME de-multiplexed sequences in fastq
    - PacBio CCS (Circular Consensus Sequence) or RoI (Read of Insert) read
    - PacBio CCS subread
    - Helicos fastq with a fixed ASCII-based Phred value for quality
    - FASTA files
- FASTA with QUAL file pairs
- CSFASTA with QUAL Files
- Legacy Formats
  - SRF files
  - Native Illumina
  - QSEQ
- Machine Specific Information
  - Illumina
  - SOLiD
  - Roche 454 (formerly Life Sciences)
  - IonTorrent
  - PacBio
  - MinION Oxford Nanopore
  - Helicos
  - Capillary (Sanger)
  - CompleteGenomics



https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/

How to search SRA data?

1- Use basic SRA search box

2- Create a query using the Advanced Search Builder

3- Use SRA toolkit to query using command line

Alternatively, start by searching data associated to a BioProject using IDs listed on a publication of interest

https://www.ncbi.nlm.nih.gov/bioproject/

NCBI Resources How To quinonesm@niaid.nih.gov My NCBI Sign Out

SRA

SRA Search

Advanced Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

Practice 1: Let’s do a Basic and Advanced Search

1

SRA | SRA | uganda malaria treatment

Create alert Advanced

→ Explore the top result. View the Run record, explore the Analysis. Are the reads 100% plasmodium?

2

SRA Advanced Search Builder

((((plasmodium falciparum nf54[Organism]) AND illumina[Platform]) AND "paired"[Layout]) AND "rna seq"[Strategy])

Edit

Builder

	Organism	plasmodium falciparum nf54
AND	Platform	illumina
AND	Layout	"paired"[Layout]
AND	Strategy	"rna seq"[Strategy]

→ How many results?

Practice 2: Let's search using BioProject, then navigate to SRA

1

BioProject | BioProject | PRJNA493853

Advanced Browse by Project attributes

By Project attributes

2

BioProject > **BioProject list**

anopheles gambiae

Please note: Searches on this page are limited to the fields available in this table. For more information, see Help.

Filters

Choose Columns

#	Accession	Project Title	Organism	Or
1	PRJNA717536	Anopheles annulipes RefSeq Genome	Anopheles annulipes	Eukaryota; Ani
2	PRJNA716340	Anopheles coluzzii RefSeq Genome sequencing and assembly	Anopheles coluzzii	Eukaryota; Ani
3	PRJNA707074	Dissecting transcriptome intraspecific variation and sex-biased expression in Anopheles arabiensis	Multiple	
4	PRJEB35263	Transcriptomics of Anopheles gambiae insecticide resistant legs	not assigned	
5	PRJNA716336	Anopheles arabiensis RefSeq Genome sequencing and assembly	Anopheles arabiensis	Eukaryota; Ani
6	PRJEB35264	The role of miRNAs in insecticide resistance in Anopheles gambiae	not assigned	

BioProject > **BioProject list**

anopheles gambiae

Please note: Searches on this page are limited to the fields available in this table. For more information, s

Filters

Project Type	Primary submission (550) Umbrella project (6)
Data Type	Genome sequencing (92) Genome sequencing and assembly (43) Metagenom Targeted loci environmental (2) Targeted Locus (Loci) (4) more
Scope	Environment (13) Monoisolate (364) Multiisolate (113) Multispecies (59)
Property	has data (536) has publications (111)
Kingdom	Bacteria (2) Eukaryota (436) Metagenomes (10)
Group	Animals (435) FCB group (1) organismal metagenomes (10) Proteobacte
Subgroup	Apicomplexans (1) Bacteroidetes/Chlorobi group (1) Gammaproteobacteria (1)

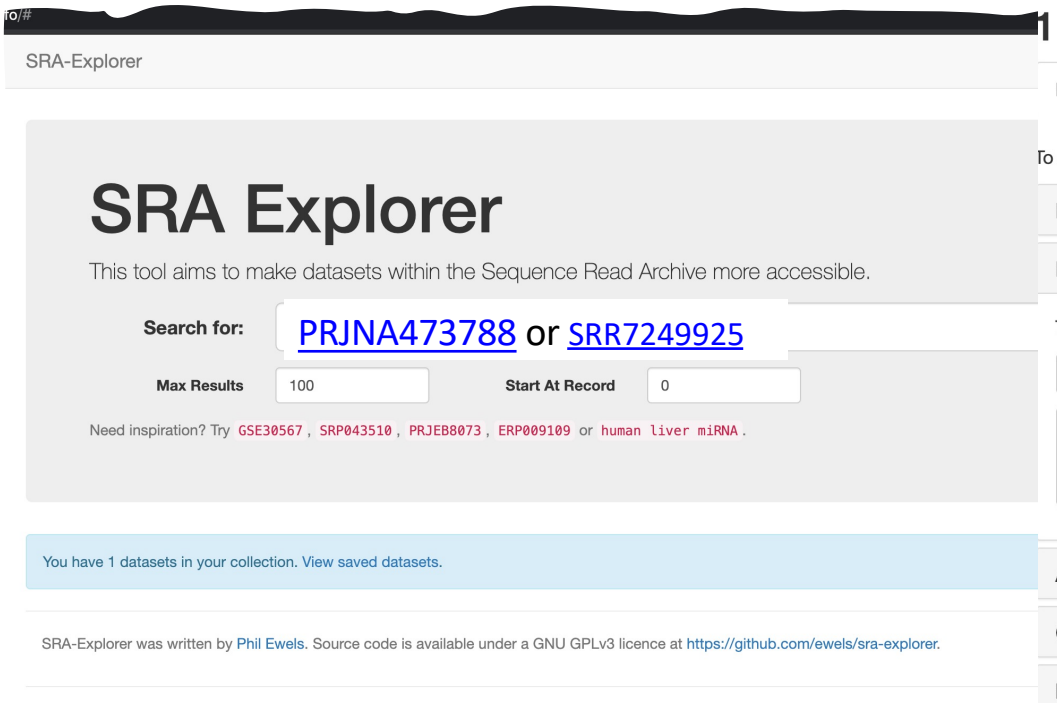
Practice 3: Let's practice downloading data associated to a publication

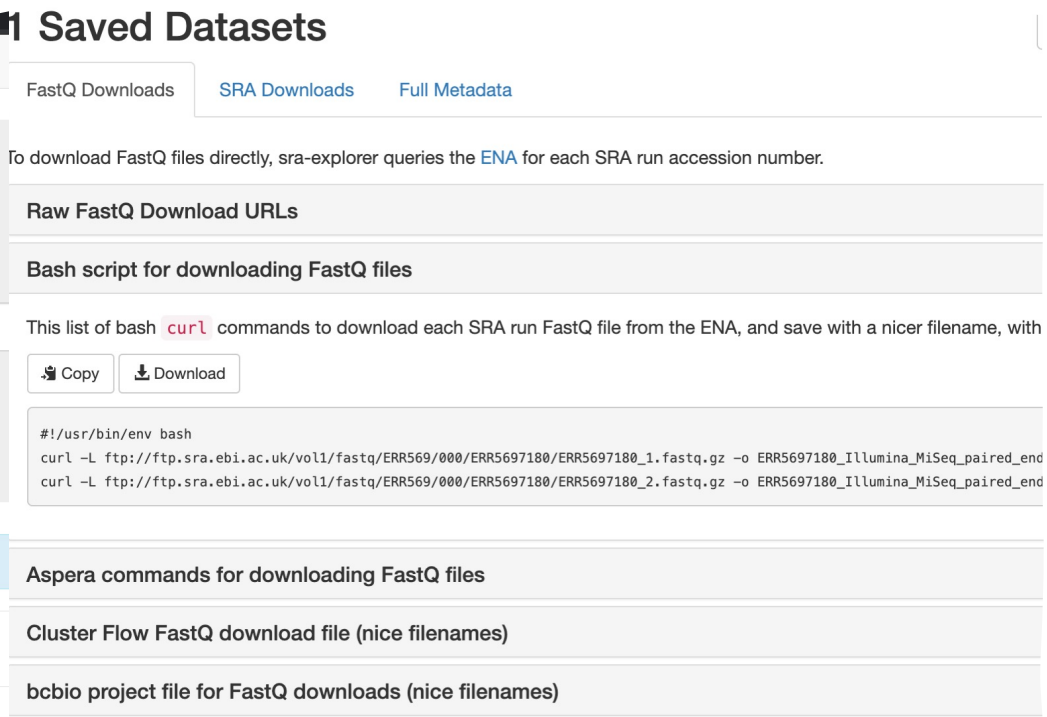
1- View this example publication: https://pubmed.ncbi.nlm.nih.gov/33665609/

2- Find BioProject ID listed on publication

3- Use NCBI Search tools to find URLs for download of FASTQ files of an adult buccal microbiome sample

Having trouble finding URLs to download? Try SRA Explorer https://sra-explorer.info/#





SRA Toolkit

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly

Download | Toolkit Documentation | XML Schema

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide

Protected Data Usage Guide

Frequently Used Tools:

fastq-dump: Convert SRA data into fastq format

prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data

sam-dump: Convert SRA data to sam format

sra-pileup: Generate pileup statistics on aligned SRA data

vdb-config: Display and modify VDB configuration information

vdb-decrypt: Decrypt non-SRA dbGaP data ("phenotype data")

Additional Tools:

abi-dump: Convert SRA data into ABI format (csfasta / qual)

illumina-dump: Convert SRA data into Illumina native formats (qseq, etc.)

sff-dump: Convert SRA data to sff format

sra-stat: Generate statistics about SRA data (quality distribution, etc.)

vdb-dump: Output the native VDB format of SRA data.

Tool: fastq-dump

Usage:

```
fastq-dump [options] <path/file> [<path/file> ...]
fastq-dump [options] <accession>
```

Use example to download paired end data:

```
fastq-dump -I --split-files SRR390728
```

Produces two fastq files (--split-files) containing ".1" and ".2" read suffices (-I) for paired-end data.

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

Practice 4a: Run a download with a simple command line in the server

```
mquinones@kla-ac-bio-03:~$ fastq-dump -I --split-files ERR5697180
```

```
mquinones@kla-ac-bio-03:~$ ls -lth
total 97M
-rw-rw-r-- 1 mquinones mquinones  49M Apr 21 03:55 ERR5697180_1.fastq
-rw-rw-r-- 1 mquinones mquinones  49M Apr 21 03:55 ERR5697180_2.fastq
drwxrwxr-x 3 mquinones mquinones 4.0K Apr 21 03:35 ncbi
```

Practice 4b: Find URLs in the Data access tab

trace.ncbi.nlm.nih.gov/Traces/sra/?run=ERR5697180

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

COVID-19 Information

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

Illumina MiSeq paired end sequencing; Raw reads: COV006440 (ERR5697180)

Change accession...

Metadata | Analysis | Reads | Data access

SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Egress	Access Type
run	15,379 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/era23/ERR/ERR5697/ERR5697180	worldwide	anonymous
		AWS	https://sra-pub-sars-cov2.s3.amazonaws.com/run/ERR5697180/ERR5697180	worldwide	anonymous
		GCP	gs://sra-pub-run-11/ERR5697180/ERR5697180.1	gs.US	gcp identity

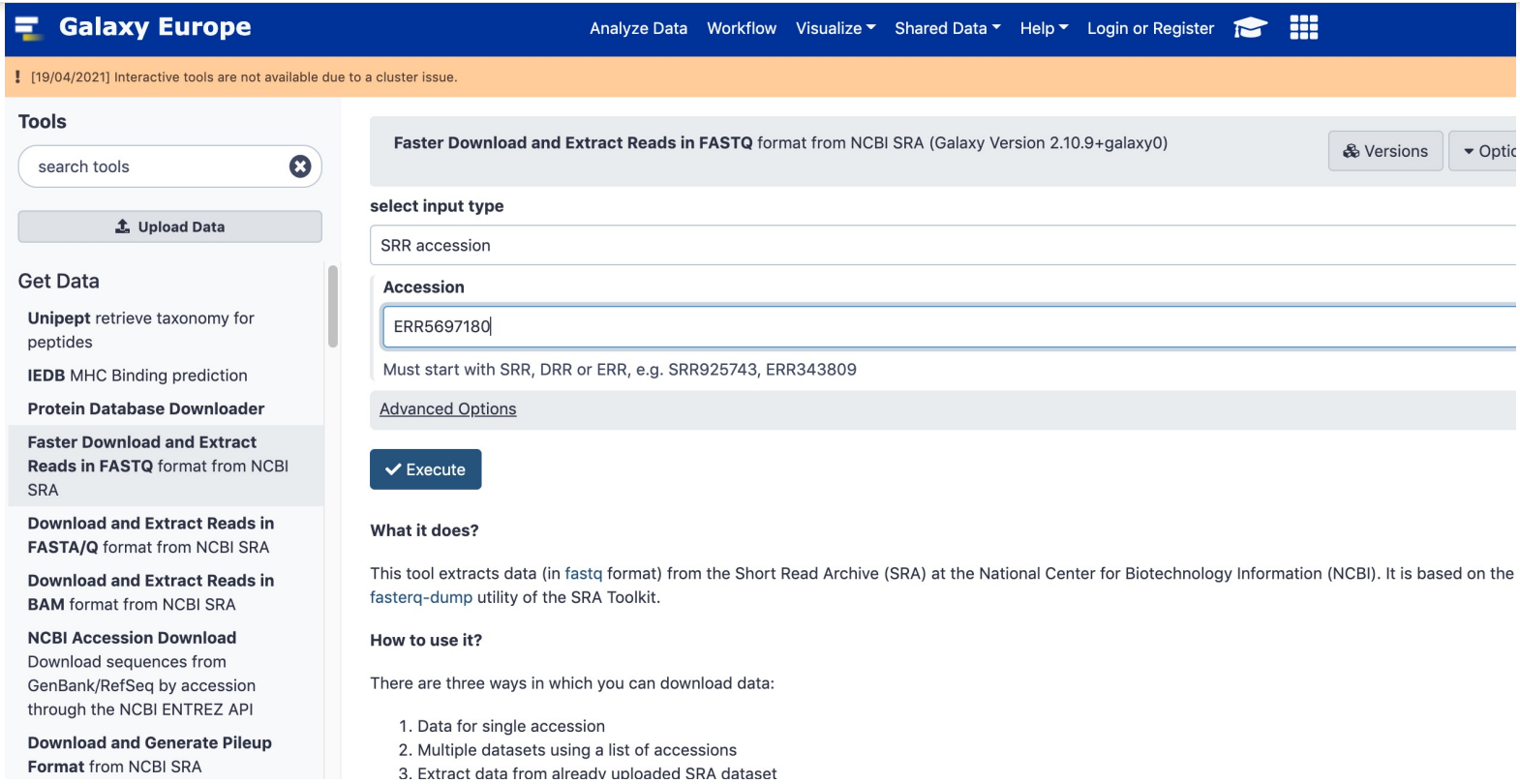
Original format

The original files submitted to SRA. These files may require specific software to open, read and interpret data.

Type	Size	Location	Name	Free Egress	Access Type
fastq	9,447 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/era23/ERZ/005697/ERR5697180/COV006440.R1.humanfilt.fastq.gz	worldwide	anonymous
fastq	10,009 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/era23/ERZ/005697/ERR5697180/COV006440.R2.humanfilt.fastq.gz	worldwide	anonymous

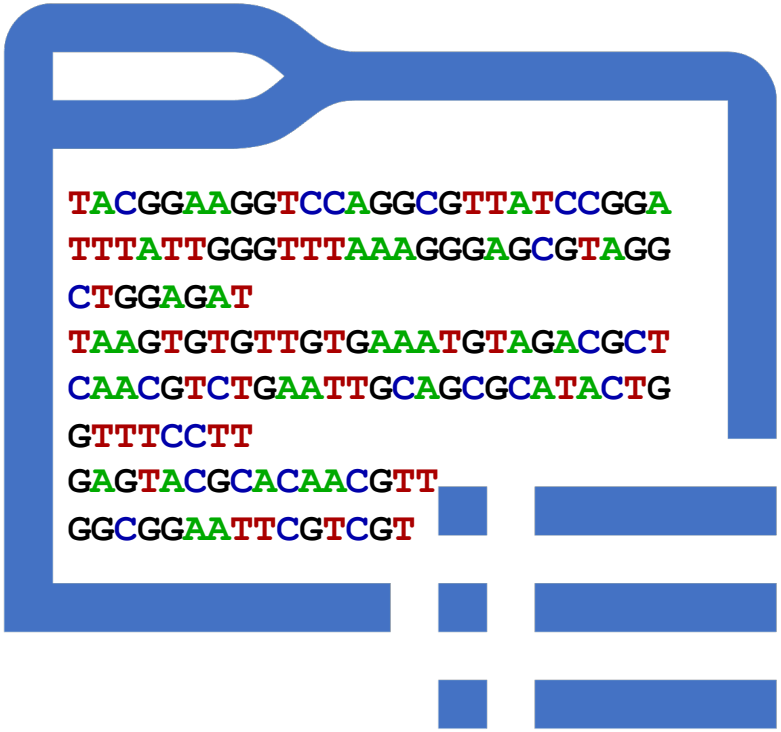
https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=ERR5697180

Practice 4c: Download directly into Galaxy <usegalaxy.org> or <usegalaxy.eu>



BioProject, BioSample and Runs contain important details

A raw file without metadata is not very useful



What do we mean by **metadata** of a Genomic Sample?
The "**What**", **How**", "**Where**", "**When**" of biological sample and processing

BioSample

- Host: *Homo sapiens*
- Organism: Severe acute respiratory syndrome coronavirus 2
- Collection date: 01-Feb-2020
- Host disease: COVID-19 DOID:0080600
- Sample collection device: Nasopharyngeal swab

SRA

- Sequencing methods: Illumina HiSeq 1000
- Library source: Viral RNA

Why is it important to annotate and publish sample metadata?

To facilitate:

- ✓ Reproducible research
- ✓ Reuse of data
- ✓ Integration of multiple studies & meta-analyses

How is metadata organized? Using models “packages” or “checklists”

NCBI Example 1: *Sars-CoV-2 from South Africa*

Assay Type:	AMPLICON
AvgSpotLen:	418
BioProject:	PRJNA636748
BioSample:	SAMN15082663
BioSampleModel:	Pathogen.cl
Center Name:	KWAZULU-NATAL RESEARCH INNOVATION AND SEQUENCING PLATFORM (UKZN
Consent:	public
Experiment:	SRX8454220
InsertSize:	0
Instrument:	Illumina MiSeq
LibraryLayout:	SINGLE
LibrarySelection:	RT-PCR
LibrarySource:	VIRAL RNA
Library Name:	KRISP_0101
LoadDate:	2020-06-02
MBases:	300
MBytes:	160
Organism:	Severe acute respiratory syndrome coronavirus 2
Platform:	ILLUMINA
ReleaseDate:	2020-06-02
Run:	SRR11907531
SRA Sample:	SRS6755790
SRA Study:	SRP265610
Sample Name:	KPCOVID_0101
collected by:	Molecular Diagnostics Services, KZN, South Africa
collection date:	2020-04-30
geo loc name:	South Africa: KZN
host:	Homo sapiens
host disease:	COVID-19
isolate:	KRISP_0101
isolation source:	nasopharyngeal swabs
lat lon:	missing
library ID:	KRISP_0101
strain:	SARS-CoV-2

BioSample Model “Package”

- Provides template
- Enforces a minimum set of attributes

*Other Models include those developed by
Genomics Standards Consortium
(e.g. MIMS, MIMARKS, MIGS)

Example 2: Sars-CoV-2 from US patients

BioProject

Severe acute respiratory syndrome coronavirus 2 Accession: PI

Severe acute respiratory syndrome coronavirus 2 Raw sequence reads

Sequence reads from US cases of COVID-19 / SARS-CoV-2

Accession	PRJNA610248
Data Type	Raw sequence reads
Scope	Multiisolate
Organism	**Severe acute respiratory syndrome coronavirus 2** [Taxonomy ID: 2697049] Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus; Severe acute respiratory syndrome coronavirus 2
Submission	Registration date: 4-Mar-2020 **CDC Pathogen Discovery Team**
Relevance	COVID-19 outbreak sequence activity

SRA Study

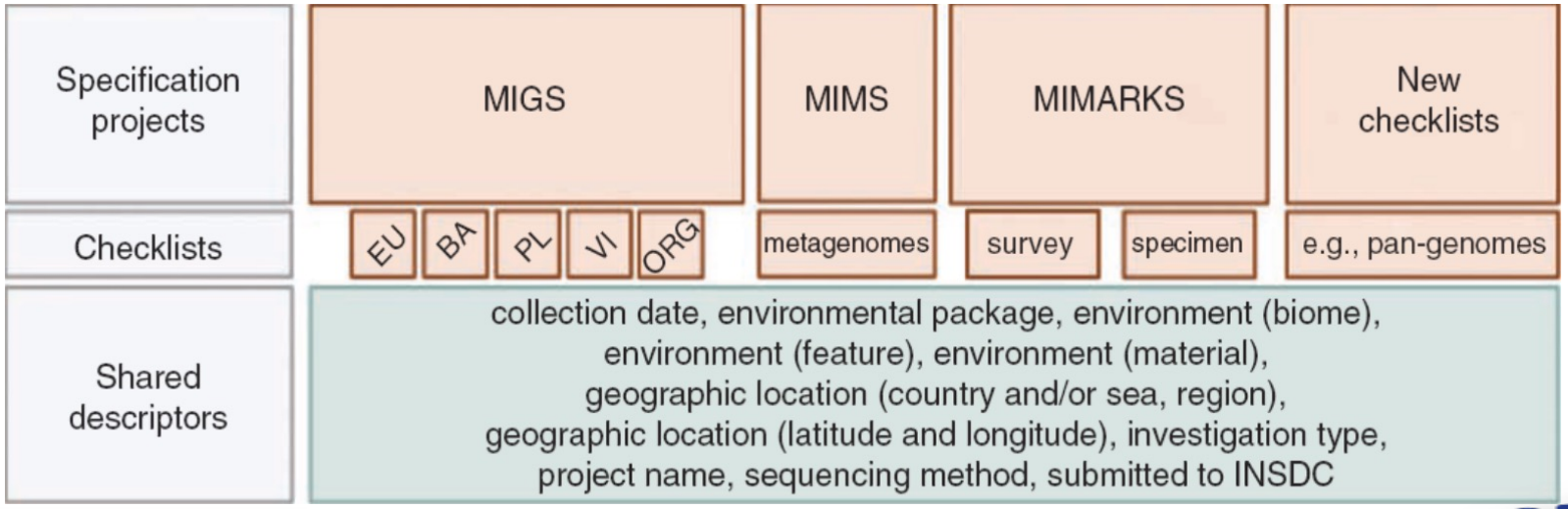
Assay Type:	AMPLICON
BioProject:	PRJNA610248
BioSampleModel:	Pathogen.cl
Center Name:	CDC-PDD
Consent:	public
InsertSize:	0
LibraryLayout:	SINGLE
LibrarySelection:	RT-PCR
LibrarySource:	VIRAL RNA
LoadDate:	2020-04-28
Organism:	Severe acute respiratory syndrome coronavirus 2
Platform:	OXFORD_NANOPORE
ReleaseDate:	2020-04-28
SRA Study:	SRP258998
host:	Homo sapiens
host disease:	COVID-19
isolation source:	human

BioSample metadata

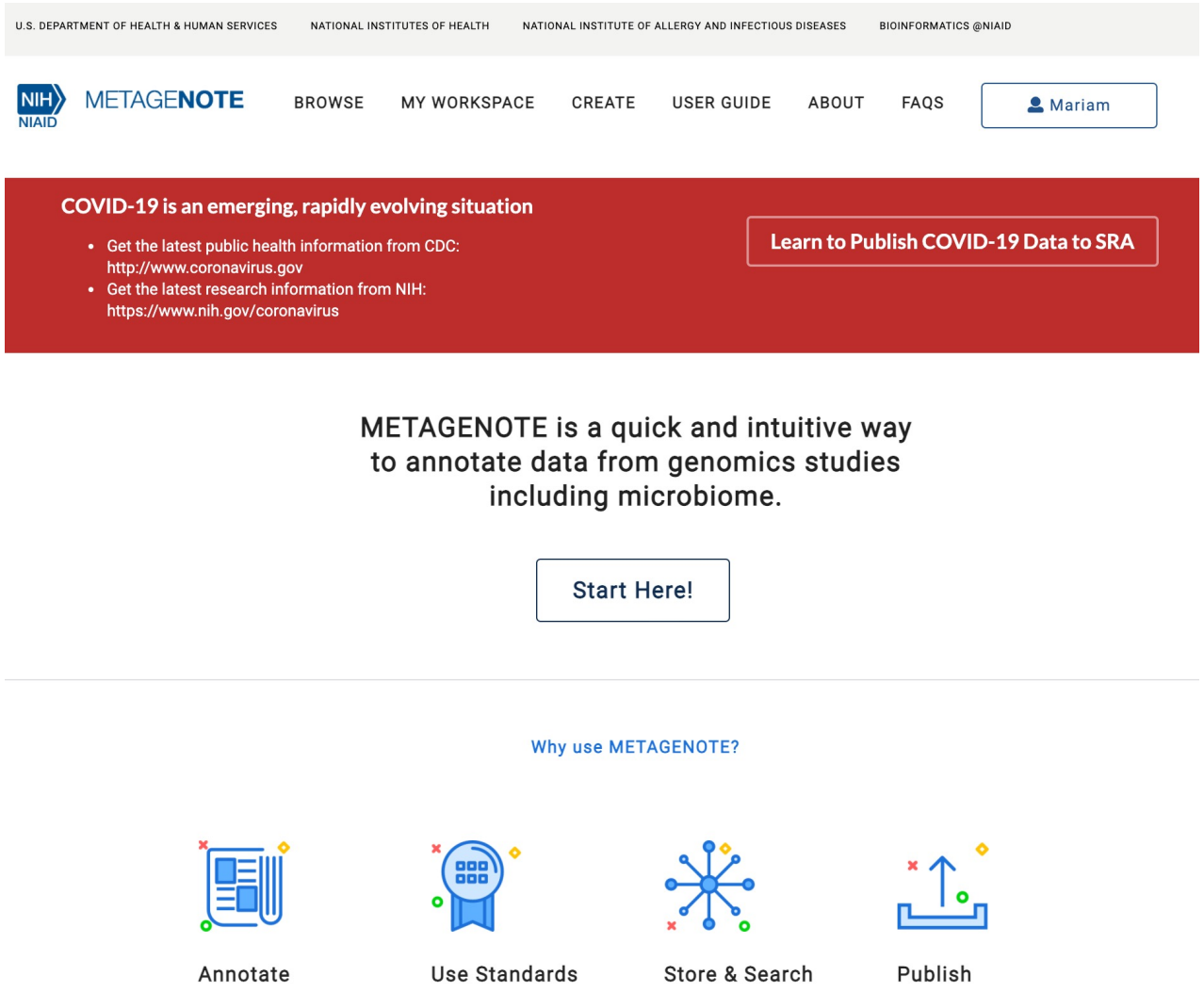
Experiment	Instrument	MBases	MBytes	collected by	collection date	geo loc name	host tissue sampled	isolate	lat lon
SRX8201037	MinION	99	92	Not available	2020-02-24	USA	oropharynx	2019-nCoV/USA-CruiseA-26/2020	39.82 N 98.57 W
SRX8201038	MinION	93	87	WA State Department of Health	2020-03-13	USA: Washington	nasopharynx, oropharynx	2019-nCoV/USA/WA-NH2/2020	47.75 N 120.74 W
SRX8201039	MinION	101	94	WA State Department of Health	2020-03-13	USA: Washington	nasopharynx, oropharynx	2019-nCoV/USA/WA-NH11/2020	47.75 N 120.74 W
SRX8201040	GridION	132	121	IA State Hygienic Laboratory	2020-03-07	USA: Iowa	oropharynx	2019-nCoV/USA-IA_6391/2020	41.87 N 93.09 W
SRX8201041	MinION	129	120	WA State Department of Health	2020-03-13	USA: Washington	nasopharynx, oropharynx	2019-nCoV/USA/WA-NH19/2020	47.75 N 120.74 W
SRX8201042	MinION	24	22	CDPH, Viral and Rickettsial Disease Laboratory	2020-01-26	USA: California	nasopharynx	2019-nCoV/USA/CDC-unassigned_TD6874	36.77 N 119.41 W

GSC: WIDELY USED STANDARD VOCABULARY FOR MICROBIOME SEQUENCE DATA

- MIMARKS - Minimum Information about a MARKer gene Sequence Project
- MIMS - metagenome
- MIGS - genome



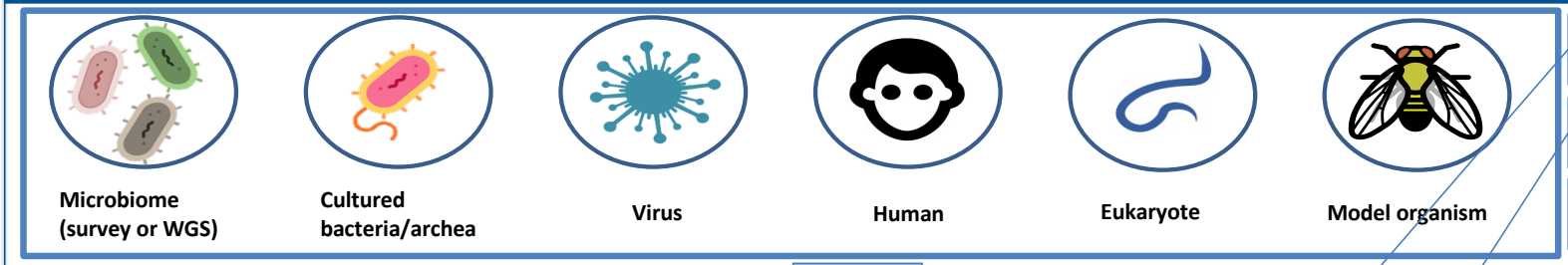
https://metagenote.niaid.nih.gov/



- ✓ Tool to annotate files with rich metadata
- ✓ Free to use
- ✓ Facilitates file transfer to SRA through a drag-and-drop function
- ✓ Provides access to ontologies
- ✓ Automates submissions to SRA

METAGENOTE's Notebook provide templates for various model types

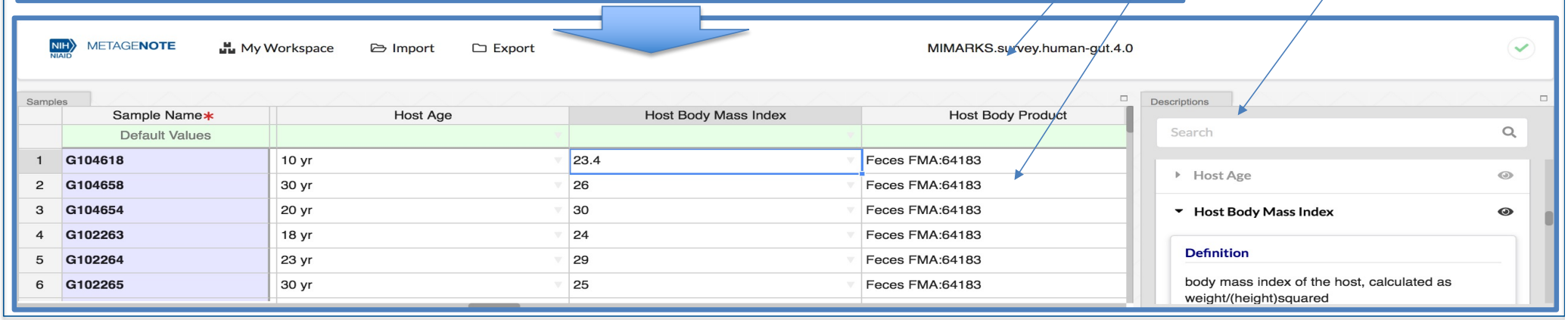
Select a template appropriate for your sample group, then annotate



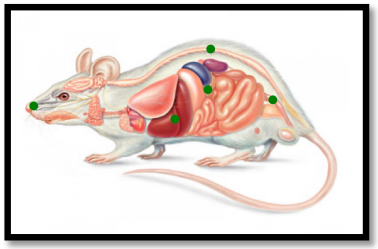
Model type

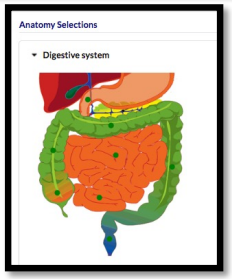
Ontologies

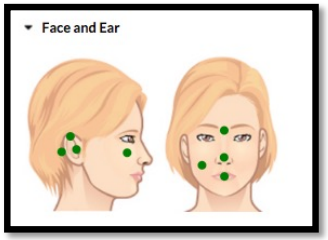
Description



"Host tissue sampled" has an anatomy diagram



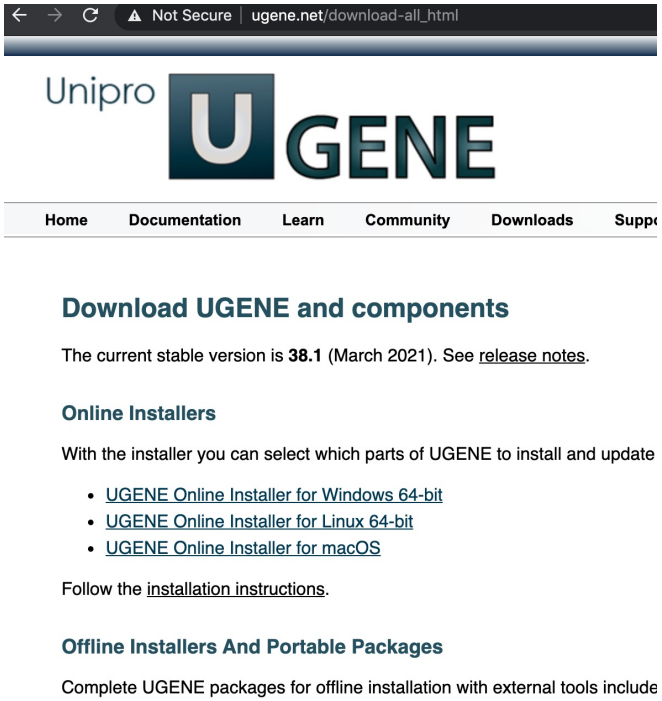




Next session:
MSB 7101: Molecular Biology for Bioinformatics
Topic: Nucleic Acid Techniques and Tools

If you are joining remotely,
please consider installing UGENE

http://ugene.net/download-all_html



Questions?

Email us: bioinformatics@niaid.nih.gov

mariam.quinones@nih.gov