Judit Szarvas, postdoc

Infectious disease bioinformatics, April 2022

Data acquisition from public databases for infectious disease studies

Infectious disease often requires context

- Studies frequently require a more global outlook
- Global supply chains make foodborne disease travel long distances

- Initiatives for increased reproducibility and open science prompted scientific publishers and journals to require sequencing data to be openly available at manuscript submission

- Dedicated repositories for sequencing data AND the related metadata

INSDC

- International Nucleotide Sequence Database Collaboration:
  - DDBJ (Japan) DRX/DRR0000001
  - EMBL-EBI (Europe) ERX/ERR0000001
  - NCBI (United States) SRX/SRR0000001
- Data deposited in either system are available from all (after few days)

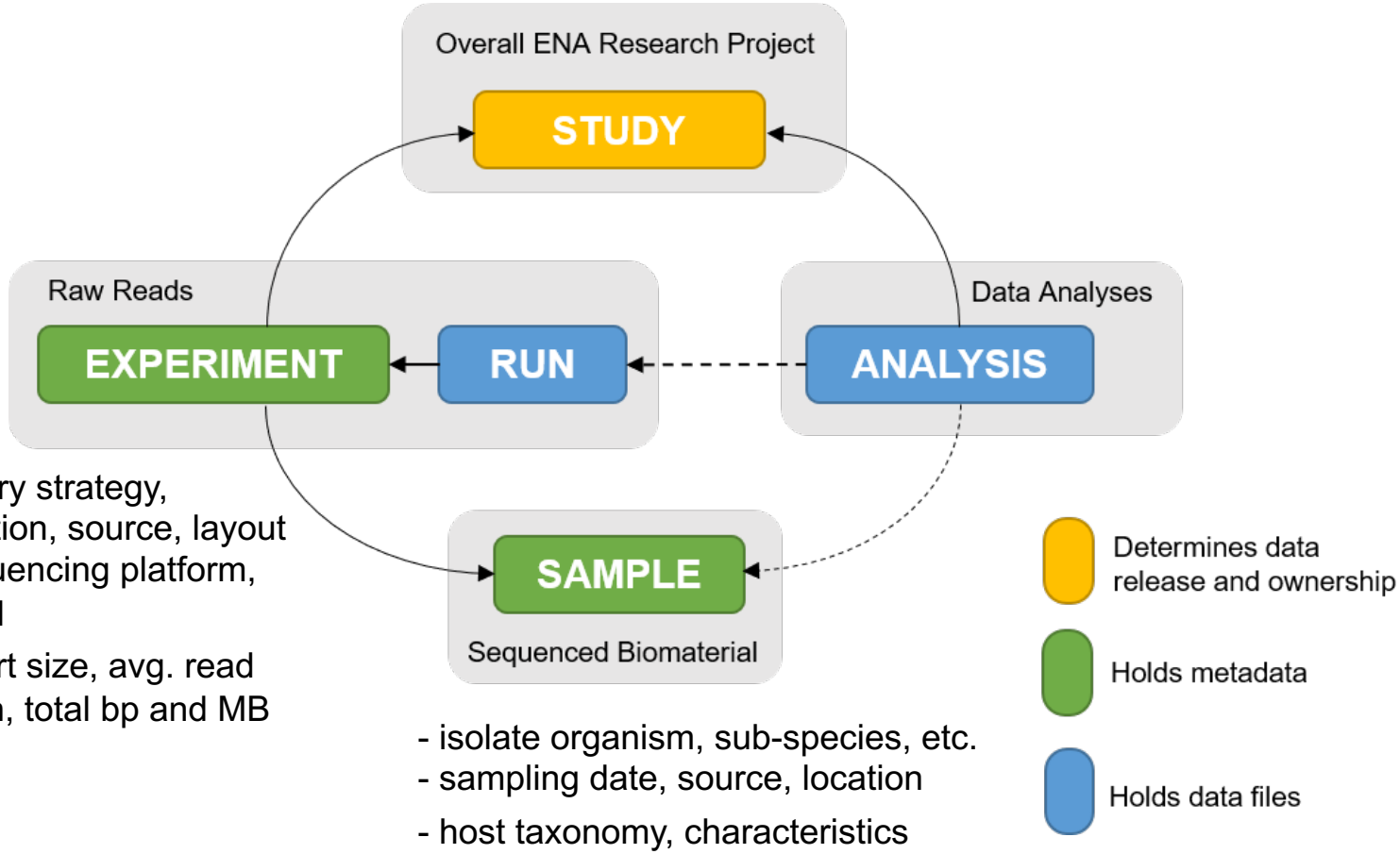
Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive (SRA)
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Common data structure

ENA	NCBI	Description
Study	BioProject	the encompassing research project fx. Citrobacter freundii samples collected from microbiology laboratories of Genovia in 2015
Sample	BioSample	a biological sample fx. swab
Experiment	Experiment	sequencing experiment for a sample and its technical replicates, seq. library and platform/instrument fx. Illumina HiSeq 2000, paired layout, random selection
Run	Run	sequencing run

https://www.ncbi.nlm.nih.gov/sra/docs/submitbio/
https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/
https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

ENA metadata model



- library strategy, selection, source, layout
- sequencing platform, model
- insert size, avg. read length, total bp and MB

- isolate organism, sub-species, etc.
- sampling date, source, location
- host taxonomy, characteristics
- data provider

How to cite specific data?

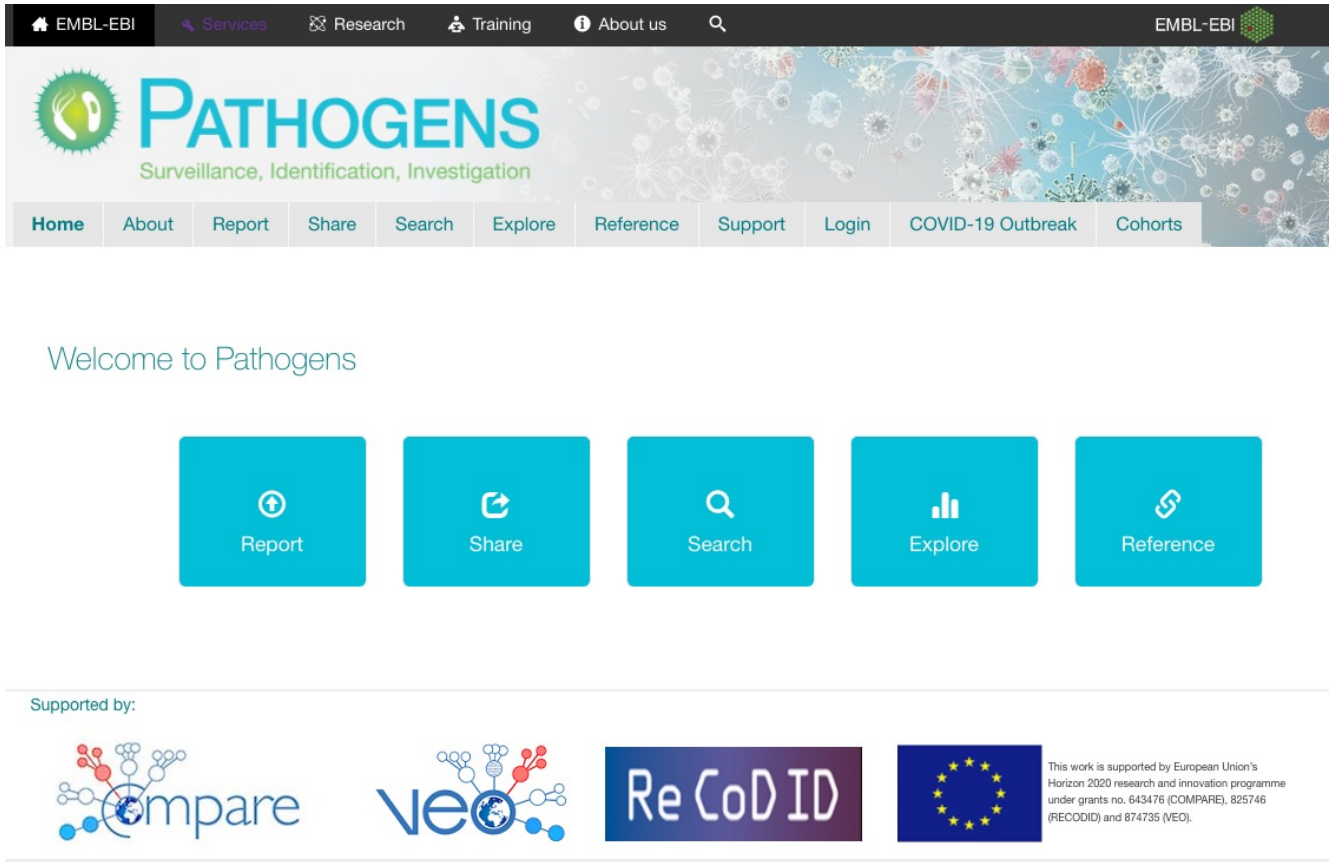
Known identifier*	Cite
Manuscript DOI and Study/Project accession	Manuscript DOI and Study/Project accession
Study/Project accession	Study/Project accession
Record accession	Study/Project accession and individual record accession

* Do your best to find a manuscript doi

https://www.ebi.ac.uk/ena/browser/about/citing-ena

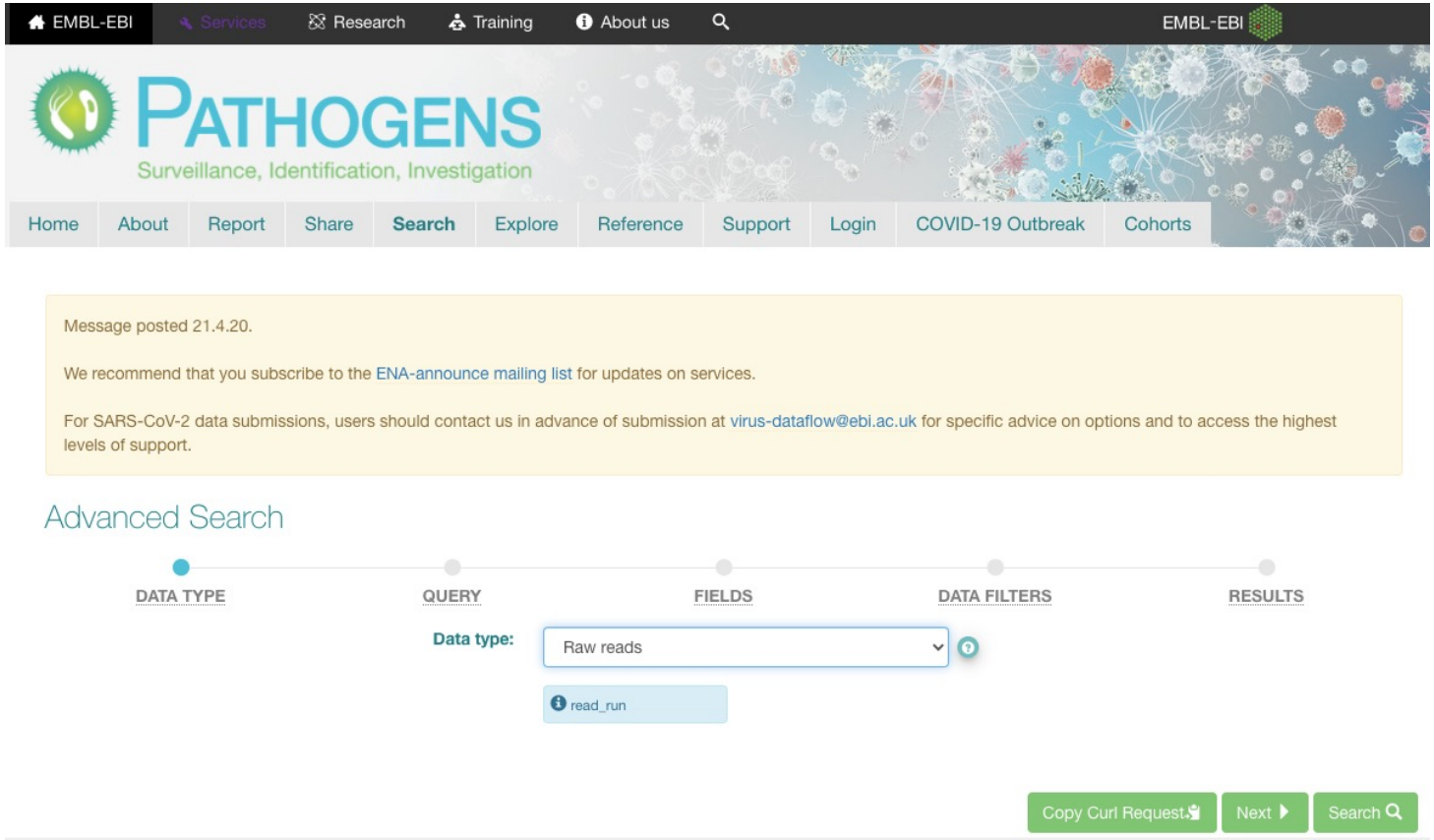
EMBL-EBI ENA - Pathogens

- EMBL-EBI Pathogens is the infectious disease sub-portal of ENA
- Advanced search has indexed sample attributes: https://www.ebi.ac.uk/ena/pathogens/search

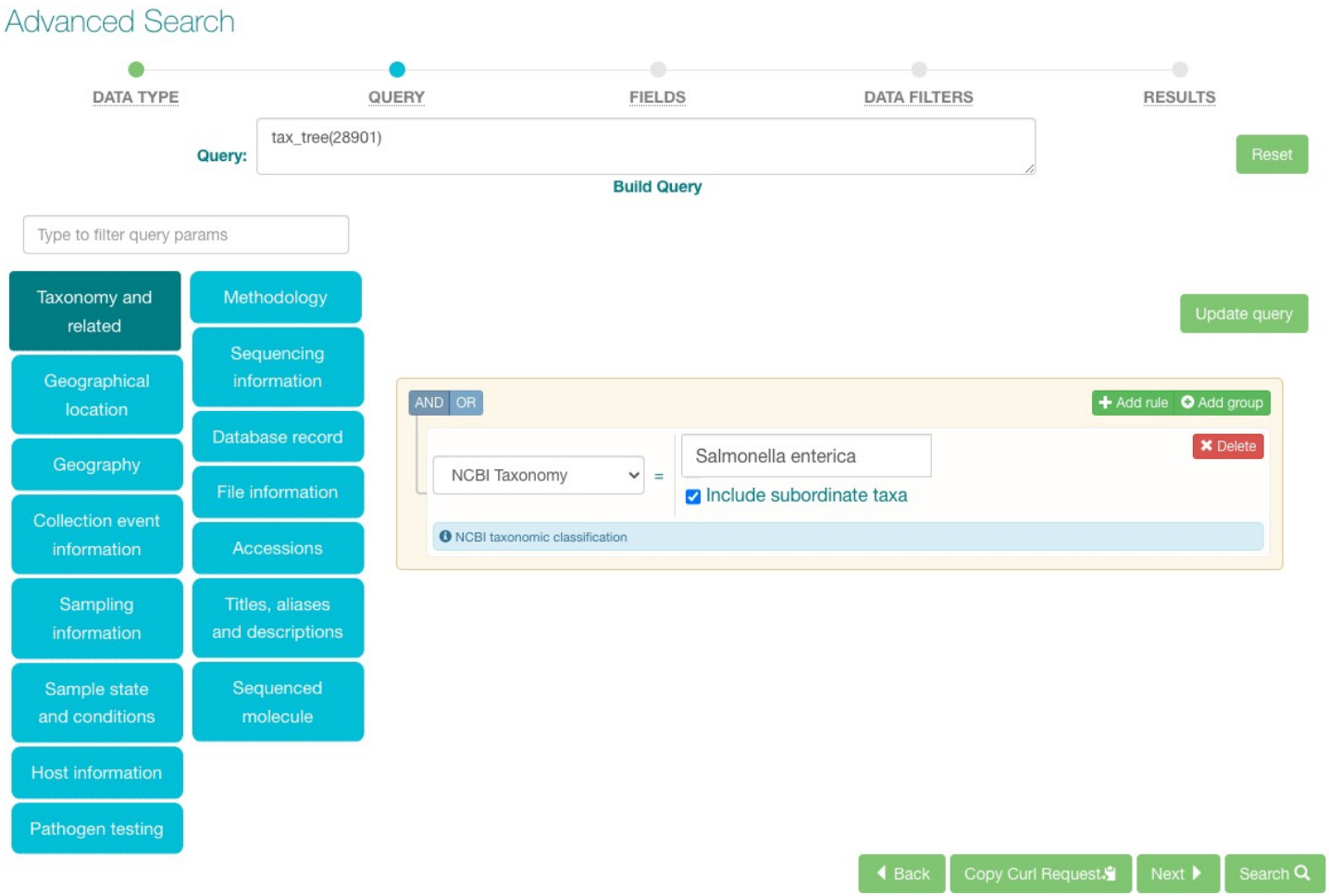


Pathogens - Advanced search

- Data type: the type of the data record in search results
- The metadata for the linked samples and experiments are jointly searched!
- Let's search for Salmonella enterica WGS PE data isolated from tomatoes



Pathogens – Isolate taxonomy



Pathogens – Isolate source

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Query: tax_tree(28901) AND isolation_source="*tomato*"

Reset

Build Query

Type to filter query params

Taxonomy and related
Geographical location
Geography
Collection event information
Sampling information
Sample state and conditions
Host information
Pathogen testing
Methodology
Sequencing information
Database record
File information
Accessions
Titles, aliases and descriptions
Sequenced molecule

Update query

AND OR

Add rule | Add group

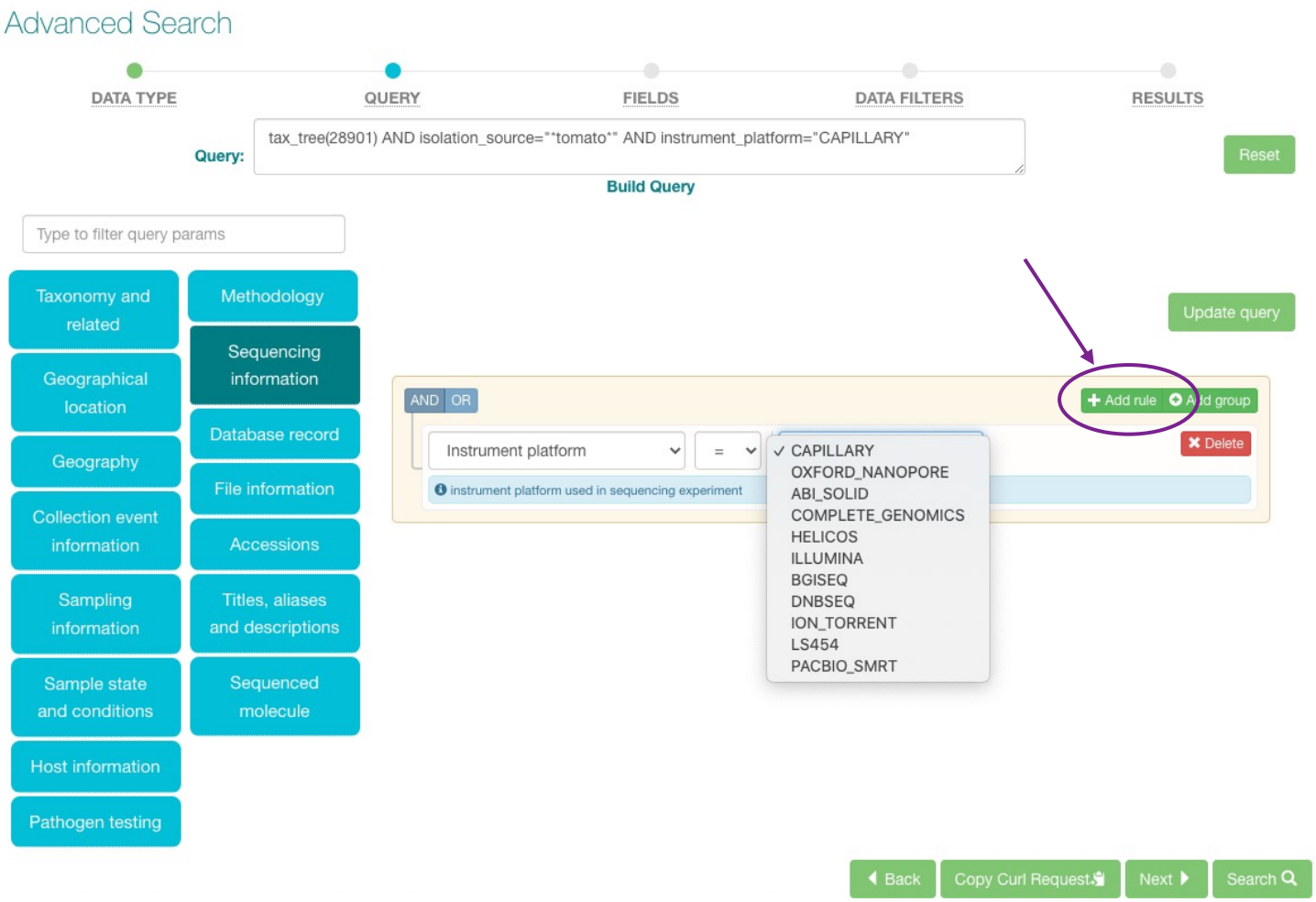
Isolation source | = | *tomato*

Delete

describes the physical, environmental and/or local geographical source of the sample

Back | Copy Curl Request | Next | Search

Pathogens – Sequencing experiment setup



Pathogens – Sequencing experiment setup

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Query: tax_tree(28901) AND isolation_source="*tomato*" AND instrument_platform="ILLUMINA" AND library_layout="PAIRED"

Reset

Build Query

Type to filter query params

Taxonomy and related
Geographical location
Geography
Collection event information
Sampling information
Sample state and conditions
Host information
Pathogen testing
Methodology
Sequencing information
Database record
File information
Accessions
Titles, aliases and descriptions
Sequenced molecule

Update query

AND OR

Add rule | Add group

Instrument platform = ILLUMINA Delete

instrument platform used in sequencing experiment

Library layout = SINGLE / ✓ PAIRED Delete

sequencing library layout

Back | Copy Curl Request | Next | Search

Pathogens – Sequencing experiment setup

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Query: tax_tree(28901) AND isolation_source="*tomato*" AND instrument_platform="ILLUMINA" AND library_layout="PAIRED" AND library_selection="RANDOM"

Reset

Build Query

Type to filter query params

Taxonomy and related
Geographical location
Geography
Collection event information
Sampling information
Sample state and conditions
Host information
Pathogen testing
Methodology
Sequencing information
Database record
File information
Accessions
Titles, aliases and descriptions
Sequenced molecule

Update query

AND OR

Add rule | Add group

Instrument platform = ILLUMINA — Delete
instrument platform used in sequencing experiment

Library layout = PAIRED — Delete
sequencing library layout

Library selection = RANDOM — Delete
method used to select or enrich the material being sequenced

Back | Copy Curl Request | Next | Search

Pathogens – Sequencing experiment setup

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Query: tax_tree(28901) AND isolation_source="*tomato*" AND instrument_platform="ILLUMINA" AND library_layout="PAIRED" AND library_selection="RANDOM" AND library_strategy="ChIA-PET"

Build Query

Reset

Type to filter query params

Taxonomy and related
Geographical location
Geography
Collection event information
Sampling information
Sample state and conditions
Host information
Pathogen testing
Methodology
Sequencing information
Database record
File information
Accessions
Titles, aliases and descriptions
Sequenced molecule

Update query

AND OR

Add group

Instrument platform = Delete
instrument platform used in sequencing experiment

Library layout = Delete
sequencing library layout

Library selection = Delete
method used to select or enrich the material being sequ

Library strategy = Delete
sequencing technique intended for the library

- ✓ ChIA-PET
- ChIP-Seq
- EST
- FL-cDNA
- Hi-C
- MRE-Seq
- POOLCLONE
- RAD-Seq
- Tn-Seq
- ssRNA-seq
- ATAC-seq
- Bisulfite-Seq
- CLONE
- CLONEEND
- DNase-Hypersensitivity
- GBS
- MBD-Seq
- MNase-Seq
- OTHER
- SELEX
- Synthetic-Long-Read
- Tethered Chromatin Conformation Capture
- WCS
- WGA
- WGS
- ncRNA-Seq
- AMPLICON
- CTS
- FAIRE-seq
- FINISHING
- MeDIP-Seq
- RIP-Seq
- RNA-Seq
- Targeted-Capture
- VALIDATION
- WXS
- miRNA-Seq

Next Search

Pathogens – Sequencing experiment setup

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Query: tax_tree(28901) AND isolation_source=""tomato"" AND instrument_platform="ILLUMINA" AND library_layout="PAIRED" AND library_selection="RANDOM" AND library_strategy="WGS" AND

Build Query

Reset

Type to filter query params

Taxonomy and related | Methodology
Geographical location | Sequencing information
Geography | Database record
Collection event information | File information
Sampling information | Accessions
Sample state and conditions | Titles, aliases and descriptions
Host information | Sequenced molecule
Pathogen testing

Update query

AND OR | Add rule | Add group

Instrument platform = ILLUMINA | Delete
instrument platform used in sequencing experiment

Library layout = PAIRED | Delete
sequencing library layout

Library selection = RANDOM | Delete
method used to select or enrich the material being sequenced

Library strategy = | Delete
sequencing technique intended for the library

Library source = | Delete
source material being sequenced

SYNTHETIC
TRANSCRIPTOMIC
GENOMIC SINGLE CELL
TRANSCRIPTOMIC SINGLE CELL
VIRAL RNA
✓ GENOMIC
METAGENOMIC
METATRANSCRIPTOMIC
OTHER

Back | Copy Curl Request | Next | Search

Pathogens – Result attributes/fields

Advanced Search

DATA TYPE | QUERY | FIELDS | DATA FILTERS | RESULTS

Default fields | Manually select fields

Fields:

fastq_aspera,fastq_bytes,fastq_md5,serovar,fastq_ftp,isolation_source,sample_accession,run_accession,location,instrument_model,first_created,country,collection_date,base_count,collected_by

Reset

Field sets: Query Fields | FASTQ Files | SRA Files | Submitted Files

Select and order Fields

Available Fields

broker_name
checklist
collecting_institute
lon
lat
center_name
cram_index_aspera
cram_index_ftp
cram_index_galaxy
dev_stage
environmental_sample

Selected Fields

fastq_ftp
isolation_source
sample_accession
run_accession
location
instrument_model
first_created
country
collection_date
base_count
collected_by

Back | Copy Curl Request | Next | Search

Pathogens – Results

DATA TYPE · QUERY · FIELDS · DATA FILTERS · RESULTS

Type to filter results

Download Files as ZIP | Download selected files

Download report: JSON TSV

FASTQ Aspera	FASTQ Bytes	Serovar	FASTQ FTP	Isolation Source	FASTQ MD5
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_2.fastq.gz	124MB 158MB		SRR10024071_1.fastq.gz SRR10024071_2.fastq.gz	tomato	5a0fe65963361d3 de7dd5289d4e46a
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_2.fastq.gz	132MB 153MB	Newport	SRR1175828_1.fastq.gz SRR1175828_2.fastq.gz	vine ripe red tomato	ba76c983c6ba1f4 1b713bd0d2a8d66
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_2.fastq.gz	129MB 142MB	Oranienburg	SRR1198902_1.fastq.gz SRR1198902_2.fastq.gz	grape tomato	ecde5d655e3c702 1535daa3b2ba445
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR120/057/SRR12017757/SRR12017757_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR120/057/SRR12017757/SRR12017757_2.fastq.gz	130MB 160MB	Litchfield	SRR12017757_1.fastq.gz SRR12017757_2.fastq.gz	cherry tomato	20216f16a47252e cd3a3dbd45e0ad6
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/000/SRR1249000/SRR1249000_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/000/SRR1249000/SRR1249000_2.fastq.gz	228MB 276MB	Abaetetuba	SRR1249000_1.fastq.gz SRR1249000_2.fastq.gz	roma tomato	69bf35d907b9033 f2e2a8f32697912a
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/001/SRR1249081/SRR1249081_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/001/SRR1249081/SRR1249081_2.fastq.gz	176MB 196MB		SRR1249081_1.fastq.gz SRR1249081_2.fastq.gz	cherry tomato	c88eac54e5dcf0cf bd825a6edd839eb
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR126/004/SRR1263424/SRR1263424_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR126/004/SRR1263424/SRR1263424_2.fastq.gz	185MB 211MB	Bareilly	SRR1263424_1.fastq.gz SRR1263424_2.fastq.gz	green house tomatoes	9e5ed3c51b9c357 f5d5deb06b7b6d2
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR127/008/SRR1272528/SRR1272528_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR127/008/SRR1272528/SRR1272528_2.fastq.gz	67MB 81MB	Typhimurium	SRR1272528_1.fastq.gz SRR1272528_2.fastq.gz	tomato	e43286d7148139 0a55dc621bd58a2

New Search | Back | Copy Curl Request

Pathogens – Results via curl

```
curl  -X POST -H "Content-Type: application/x-www-form-urlencoded" -d
"dataPortal=pathogen&result=read_run&query=tax_tree(28901)%20AND%20is
olation_source%3D%22*tomato*%22%20AND%20instrument_platform%3D%22ILLU
MINA%22%20AND%20library_layout%3D%22PAIRED%22%20AND%20library_selecti
on%3D%22RANDOM%22%20AND%20library_strategy%3D%22WGS%22%20AND%20librar
y_source%3D%22GENOMIC%22&fields=fastq_aspera%2Cfastq_bytes%2Cserovar%
2Cfastq_ftp%2Cisolation_source%2Cfastq_md5%2Csample_accession%2Crun_a
ccession%2Clocation%2Clat%2Cinstrument_model%2Cfirst_created%2Ccountr
y%2Ccollection_date%2Clon%2Cbase_count%2Ccollected_by&limit=0&format=
tsv" https://www.ebi.ac.uk/ena/portal/api/search > search_results.tsv
```

url encoding for special characters:

%22 = “

%2C = ,

%3D = :

%20 = space

Pathogens – Raw data download via browser

DATA TYPE — QUERY — FIELDS — DATA FILTERS — RESULTS

Type to filter results

Download Files as ZIP | Download selected files | Download report: JSON TSV

FASTQ Aspera	FASTQ Bytes	Serovar	FASTQ FTP	Isolation Source	FASTQ MD5
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_2.fastq.gz	124MB 158MB		SRR10024071_1.fastq.gz SRR10024071_2.fastq.gz	tomato	5a0fe65963361d3 de7dd5289d4e46a
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_2.fastq.gz	132MB 153MB	Newport	SRR1175828_1.fastq.gz SRR1175828_2.fastq.gz	vine ripe red tomato	ba76c983c6ba1f4 1b713bd0d2a8d66
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_2.fastq.gz	129MB 142MB	Oranienburg	SRR1198902_1.fastq.gz SRR1198902_2.fastq.gz	grape tomato	ecde5d655e3c702 1535daa3b2ba445
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR120/057/SRR12017757/SRR12017757_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR120/057/SRR12017757/SRR12017757_2.fastq.gz	130MB 160MB	Litchfield	SRR12017757_1.fastq.gz SRR12017757_2.fastq.gz	cherry tomato	20216f16a47252e cd3a3dbd45e0ad6
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/000/SRR1249000/SRR1249000_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/000/SRR1249000/SRR1249000_2.fastq.gz	228MB 276MB	Abaetetuba	SRR1249000_1.fastq.gz SRR1249000_2.fastq.gz	roma tomato	69bf35d907b9033 f2e2a8f32697912a
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/001/SRR1249081/SRR1249081_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR124/001/SRR1249081/SRR1249081_2.fastq.gz	176MB 196MB		SRR1249081_1.fastq.gz SRR1249081_2.fastq.gz	cherry tomato	c88eac54e5dcf0cf bd825a6edd839eb
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR126/004/SRR1263424/SRR1263424_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR126/004/SRR1263424/SRR1263424_2.fastq.gz	185MB 211MB	Bareilly	SRR1263424_1.fastq.gz SRR1263424_2.fastq.gz	green house tomatoes	9e5ed3c51b9c357 f5d5deb06b7b6d2
fasp.sra.ebi.ac.uk:/vol1/fastq/SRR127/008/SRR1272528/SRR1272528_1.fastq.gz fasp.sra.ebi.ac.uk:/vol1/fastq/SRR127/008/SRR1272528/SRR1272528_2.fastq.gz	67MB 81MB	Typhimurium	SRR1272528_1.fastq.gz SRR1272528_2.fastq.gz	tomato	e43286d7148139 0a55dc621bd58a2

New Search | Back | Copy Curl Request

Pathogens – Raw data download via tsv

astq_aspera fastq_bytes serovar fastq_ftp isolation_source fastq_md5 sample_accession
run_accession location lat instrument_model first_created country
collection_date lon base_count collected_by

fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_1.fastq.gz;fasp.sra.ebi.ac.uk:/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_2.fastq.gz 129552490;165286840
ftp.sra.ebi.ac.uk/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_1.fastq.gz;ftp.sra.ebi.ac.uk/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_2.fastq.gz tomato
5a0fe65963361d3c933359ca4d5772fa;de7dd5289d4e46a01c23b079f4c3a14e SAMN12662374
SRR10024071 Illumina MiSeq 2019-09-02 USA:NY
2019-08-09 402393561 NYSDOH

fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_1.fastq.gz;fasp.sra.ebi.ac.uk:/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_2.fastq.gz 138525092;160096679 Newport
ftp.sra.ebi.ac.uk/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_1.fastq.gz;ftp.sra.ebi.ac.uk/vol1/fastq/SRR117/008/SRR1175828/SRR1175828_2.fastq.gz vine ripe red tomato
ba76c983c6ba1f41716133d279856aae;1b713bd0d2a8d6663e1b0f22532ebf7b SAMN02345427
SRR1175828 Illumina MiSeq 2014-03-03 USA:IN 2012-01-03
394193056 FDA

fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_1.fastq.gz;fasp.sra.ebi.ac.uk:/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_2.fastq.gz 135077054;149106613 Oranienburg
ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_1.fastq.gz;ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/002/SRR1198902/SRR1198902_2.fastq.gz grape tomato
ecde5d655e3c702c80b0f21de5d95a68;1535daa3b2ba44593cd02d6001aa3ad2 SAMN02345583
SRR1198902 Illumina MiSeq 2014-03-20 USA:TX 2012-06-07
367572158 FMA

Pathogens – Raw data download with wget

```
wget --quiet --directory-prefix=<download_dir_path> --timeout=30
--tries=3 --waitretry=2
ftp.sra.ebi.ac.uk/vol1/fastq/SRR100/071/SRR10024071/SRR10024071_
1.fastq.gz
```

--timeout: connection is kept alive until X sec, even if no data is received

--tries: in case the connection breaks, it attempts to re-connect and continue downloading X times

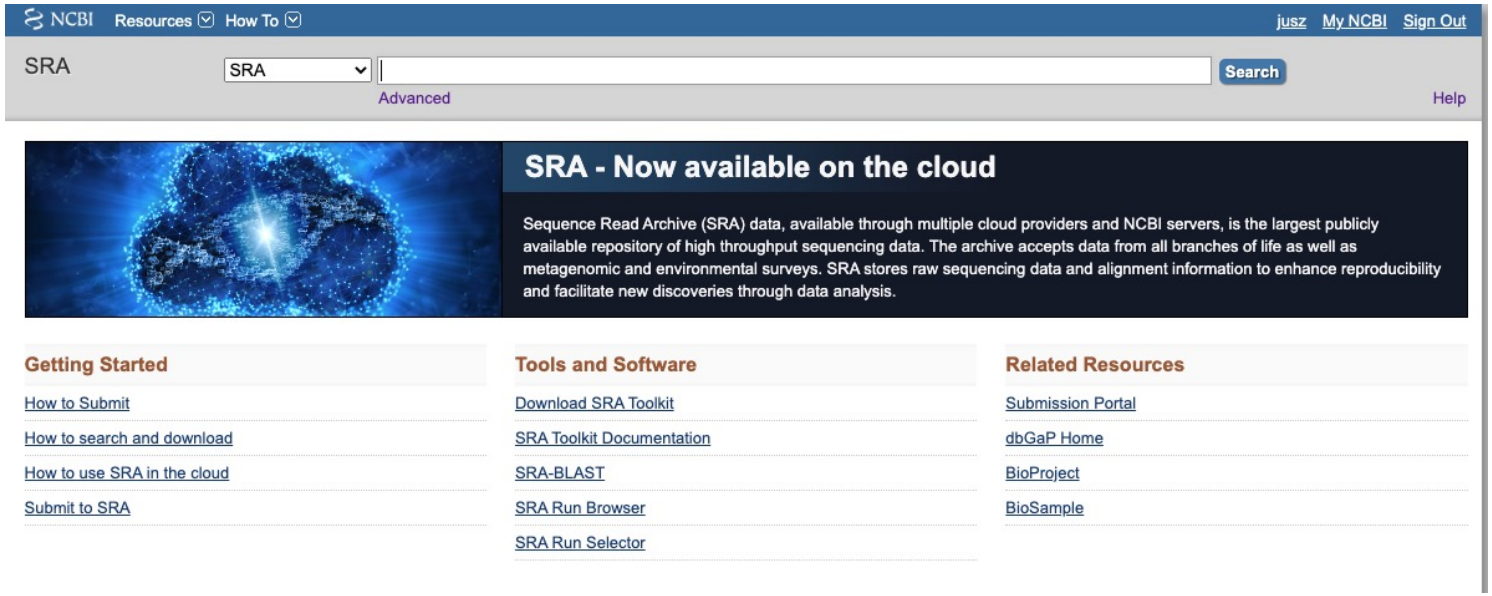
--waitretry: X sec of sleep between re-connecting

Pathogens – Raw data download in a loop

```
cut -f 4 search_results.tsv | cut -d";" -f 1 > fastq_urls.lst

while read FWD_URL;
do
    wget --quiet --directory-prefix=<download_dir_path> --timeout=30
--tries=3 --waitretry=2 ${FWD_URL};
    wget --quiet --directory-prefix=<download_dir_path> --timeout=30
--tries=3 --waitretry=2 ${FWD_URL/_1.fastq.gz/_2.fastq.gz};
done<fastq_urls.lst
```

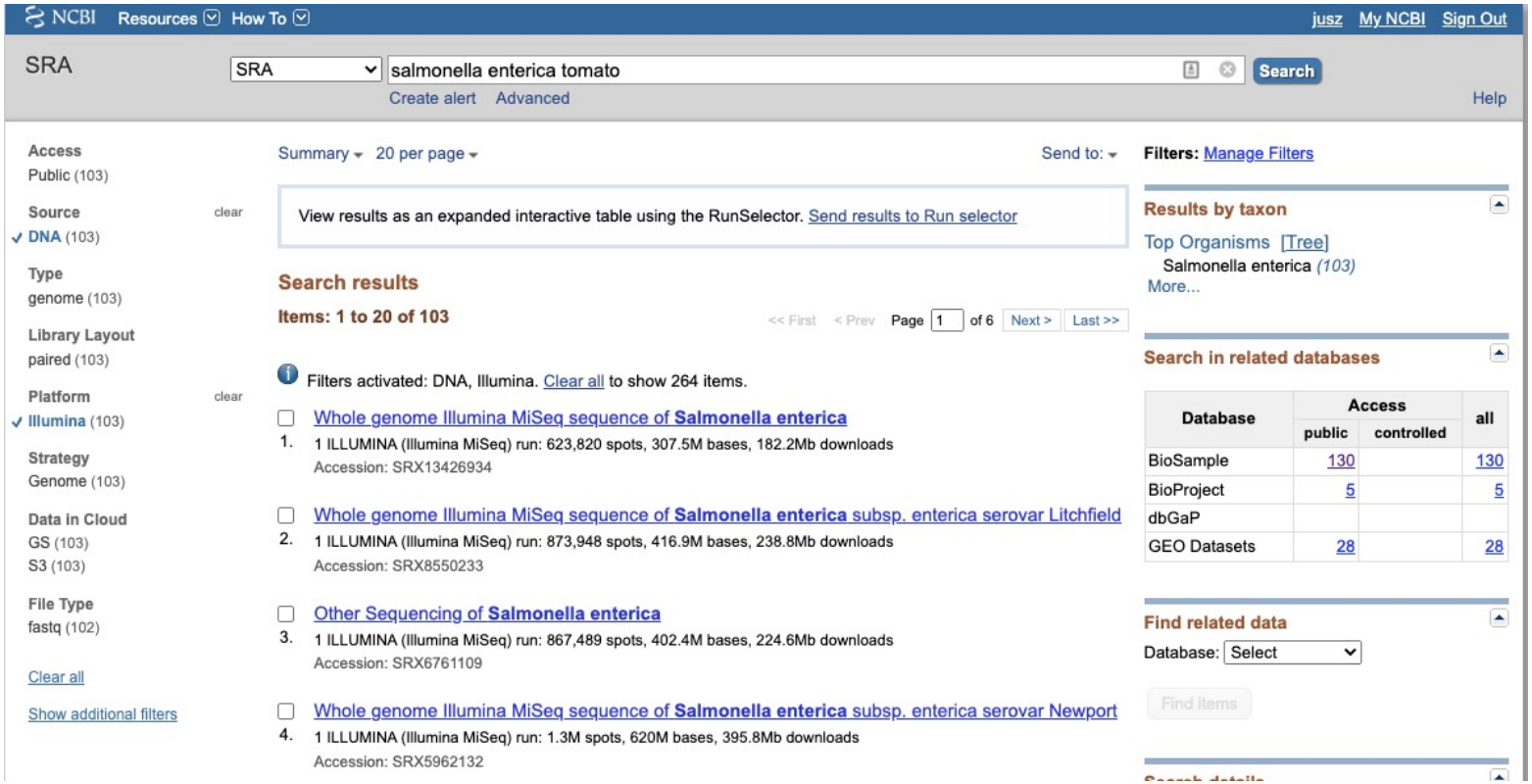
NCBI Short Read Archive



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

NCBI SRA search

- A possible workflow:
  - construct the search query by refining the filters via the SRA website https://www.ncbi.nlm.nih.gov/sra



NCBI SRA runinfo

– use the Trace SRA API access to download a runinfo csv with run records for this query https://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?save=efetch&db=sra&rettype=runinfo&term=query_string

```
Run,ReleaseDate,LoadDate,spots,bases,spots_with_mates,avgLength,size_MB,AssemblyName,down
load_path,Experiment,LibraryName,LibraryStrategy,LibrarySelection,LibrarySource,LibraryLa
yout,InsertSize,InsertDev,Platform,Model,SRAStudy,BioProject,Study_Pubmed_id,ProjectID,Sa
mple,BioSample,SampleType,TaxID,ScientificName,SampleName,g1k_pop_code,source,g1k_analysi
s_group,Subject_ID,Sex,Disease,Tumor,Affection_Status,Analyte_Type,Histological_Type,Body
_Site,CenterName,Submission,dbgap_study_accession,Consent,RunHash,ReadHash
SRR3476842,2016-05-10 00:00:00,2016-05-05
04:37:42,2278432,455686400,2278432,200,164,,https://sra-downloadb.st-
va.ncbi.nlm.nih.gov/sos2/sra-pub-run-
3/SRR3476842/SRR3476842.1,SRX1741527,63_E4743,WGS,RANDOM,GENOMIC,PAIRED,0,0,ILLUMINA,Illu
mina HiSeq
1000,SRP074336,PRJNA320449,,320449,SRS1421088,SAMN04939497,simple,90371,Salmonella
enterica subsp. enterica serovar Typhimurium,63_E4743,,,,,,,no,,,,,"INSTITUTE OF CLINICAL
PATHOLOGY AND MEDICAL RESEARCH, WESTMEAD
HOSPITAL",SRA423252,,public,E536B20BF1274C27823BBC050C1FF31B,E9FF0E1D4F3163C3D7369089344C
5B0C
SRR7723982,2018-08-20 16:22:11,2018-08-20
16:15:57,362046,172455250,362046,476,119,,https://sra-downloadb.be-
md.ncbi.nlm.nih.gov/sos3/sra-pub-run-20/SRR7723982/SRR7723982.1,SRX4580494,Nextera XT
library SEQ000077814,WGS,RANDOM,GENOMIC,PAIRED,500,0,ILLUMINA,Illumina
MiSeq,SRP018785,PRJNA186035,2,186035,SRS3694824,SAMN09865679,simple,108619,Salmonella
enterica subsp. enterica serovar
Newport,CFSAN001892,,,,,,,no,,,,,CFSAN,SRA761059,,public,FC70D7DAD3ED37217A12F4F3925509FF
,4CE6BFBF4C84E4E484EBA60C35032196
```

NCBI SRA sample metadata

– use SRA API to download biosample metadata for individual runs
https://www.ncbi.nlm.nih.gov/sra/?term=SRR6666614&format=text

```
Accession: SRX3643475
Title: Whole genome Illumina MiSeq sequence of Salmonella enterica
Experiment Design: MiSeq deep shotgun sequencing of cultured isolate.
Submission: CFSAN
Study accession: SRP098999
Study Title: GenomeTrakr Project: Michigan Department of Agriculture and Rural Development
Study Abstract: Whole genome sequencing of cultured Salmonella enterica as part of the US Food and Drug
Administration's WGS surveillance effort for the rapid traceback of foodborne pathogens.
Study Center:
Study Center Project: Salmonella enterica (ID=)
Project name: Salmonella enterica
Sample Accession: SRS2908290
Sample Description:
Sample Common Name: , (TaxonId=28901)
Sample Attributes: collection_date=2018; isolation source=cherry tomato; source type=Food; IFSAC+
Category=seeded vegetables (solanaceous); ontological term=miniature tomato (whole,
raw):FOODON_03311800; attribute_package=environmental/food/other; strain=18FMFA000068-10;
collected_by=Michigan Department of Agriculture and Rural Development; lat_lon=missing;
geo_loc_name=USA:MI; isolate_name_alias=CFSAN074863; PublicAccession=CFSAN074863;
ProjectAccession=PRJNA368989; Species=enterica; Genus=Salmonella; BioSampleModel=Pathogen.env
Sample Links: =
Library Name: Nextera XT library SEQ000068723
Library Strategy: WGS
Library Source: GENOMIC
Library Selection: RANDOM
Library Layout: PAIRED, Orientation: , Nominal length: 500, Nominal Std Dev:
Platform Name: ILLUMINA
Spot descriptor: 1) Application Read, Forward; 2) Application Read, Reverse
Total: 1 run, 2M spots, 694.3M bases
Run #1: SRR6666614, 2022210 spots, 694252924 bases
```

NCBI SRA run download

- use NCBI SRA Toolkit prefetch and fasterq-dump to download raw reads using their run accession
  - installed on computerome: sratoolkit/3.0.0
  - for downloading run (fastq) files https://github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump
  - it downloads the .sra file, which is a “compressed” file format, and fasterq-dump is needed to unpack it (uses CPU time!):

```
prefetch SRR3476842;
fasterq-dump SRR3476842;
```

NCBI SRA programmatically

- NCBI Entrez utilities
  - for programmatic access https://www.ncbi.nlm.nih.gov/books/NBK25499/
  - recommended to use a parser, for example biopython's Entrez module: https://biopython.org/docs/latest/api/Bio.Entrez.html

Exercise

Exercise

Search in ENA Pathogens for *Listeria monocytogenes* isolates that were in connection with sprouts (any kind), sampled after 2015-01-01, and shotgun whole-genome sequenced on an Illumina platform.

Gather ftp download paths and metadata regarding collection date, country and source, together with accession numbers for the run, sample, and study.

Download one (1) sample from the most recent ones using either wget (you can run on the login node) or sratoolkit (create and submit a job for it)

DTU