Les databases

Pratique #1

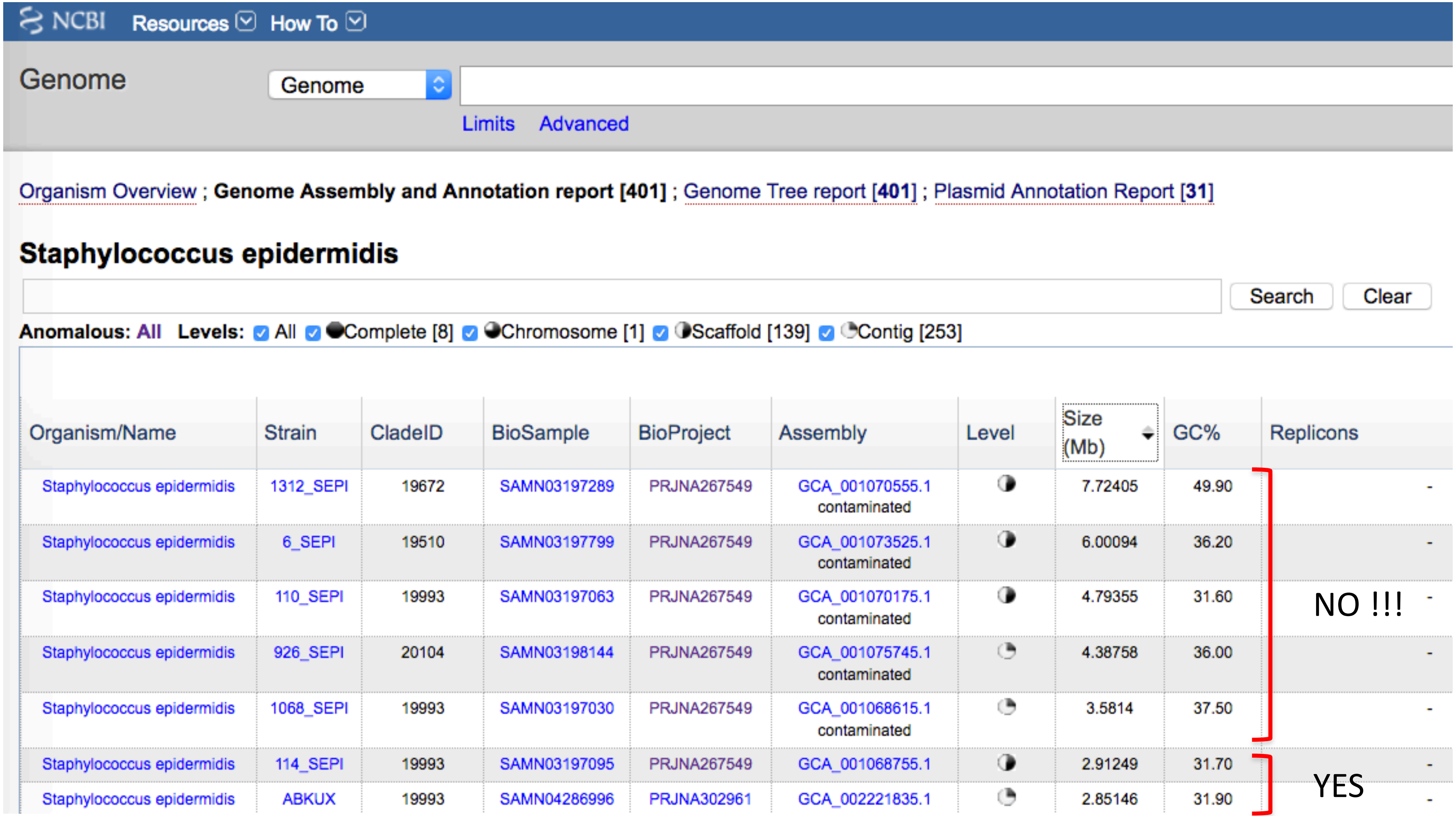
Genome DataBases

- All published genome sequence have to be available in a public DB
  - Consortia made by 3 big DBs
    - EMBL (European Molecular Biology Laboratory nucleotide sequence database at EBI, Hinxton, UK)
    - GenBank (at National Center for Biotechnology information, NCBI, Bethesda, MD, USA)
    - DDBJ (DNA Data Bank Japan at CIB , Mishima, Japan)
    - Ces grandes banques généralistes s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (The DDBJ/EMBL/GenBank Feature Table Definition).
  - GOLD ("Genomes OnLine Database") : base de données qui recense les milliers de génomes séquencés ou en voie de séquençage.
  - PATRIC ("the Pathosystems Resource Integration Center")
  - MAGE (Genoscope)

Retrieve genomes

- “Assemblies” from NCBI /!\

Species in repositories of WGS data are often mislabelled! **Trust no one!**



Organism Overview ; **Genome Assembly and Annotation report [401]** ; Genome Tree report [401] ; Plasmid Annotation Report [31]

Staphylococcus epidermidis

Anomalous: All **Levels:** ☑ All ☑ ●Complete [8] ☑ ●Chromosome [1] ☑ ◑Scaffold [139] ☑ ◔Contig [253]

Organism/Name	Strain	CladeID	BioSample	BioProject	Assembly	Level	Size (Mb)	GC%	Replicons
Staphylococcus epidermidis	1312_SEPI	19672	SAMN03197289	PRJNA267549	GCA_001070555.1 contaminated	◑	7.72405	49.90	-
Staphylococcus epidermidis	6_SEPI	19510	SAMN03197799	PRJNA267549	GCA_001073525.1 contaminated	◑	6.00094	36.20	-
Staphylococcus epidermidis	110_SEPI	19993	SAMN03197063	PRJNA267549	GCA_001070175.1 contaminated	◑	4.79355	31.60	-
Staphylococcus epidermidis	926_SEPI	20104	SAMN03198144	PRJNA267549	GCA_001075745.1 contaminated	◔	4.38758	36.00	-
Staphylococcus epidermidis	1068_SEPI	19993	SAMN03197030	PRJNA267549	GCA_001068615.1 contaminated	◔	3.5814	37.50	-
Staphylococcus epidermidis	114_SEPI	19993	SAMN03197095	PRJNA267549	GCA_001068755.1	◑	2.91249	31.70	-
Staphylococcus epidermidis	ABKUX	19993	SAMN04286996	PRJNA302961	GCA_002221835.1	◔	2.85146	31.90	-

NO !!!

YES

- **Contigs** = first **level**
- **Scaffolds** (supercontigs) = place several contigs in the correct order and orientation and represent sequencing gaps between the contigs with series of NNN’s
- **Chromosome** = generally 1record for each chromosome (with N’s)
- **Complete =** assemblies without sequencing gaps

GenBank Format

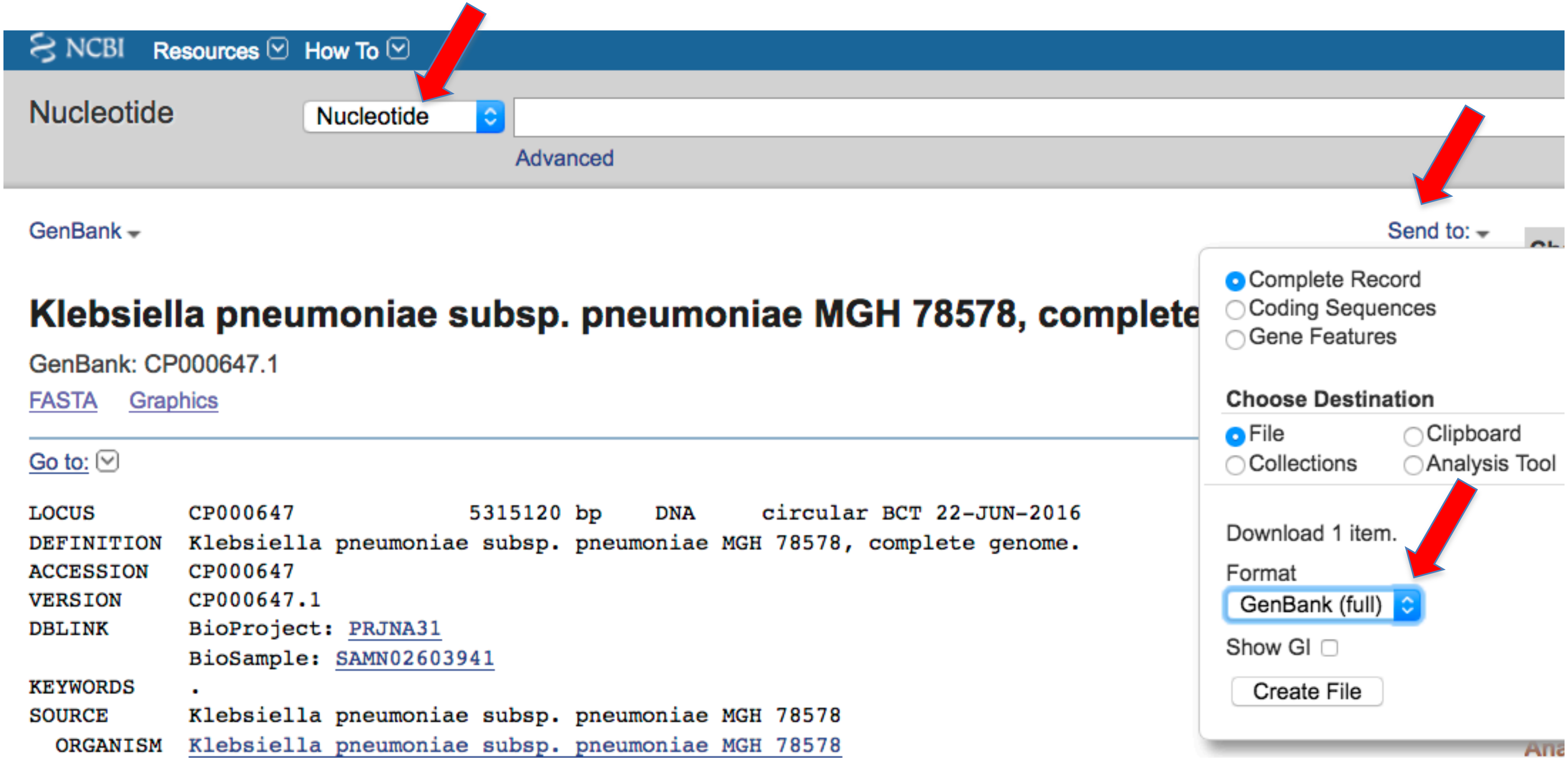
```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required
for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a
novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University,
New
            Haven, CT, USA
FEATURES             Location/Qualifiers
 source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
```

suite

```
gene            687..3158
                     /gene="AXL2"
 CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML
BASE COUNT     1510 a   1074 c    835 g   1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg

         ...
//
```

Download genbank @NCBI nucleotide



Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete

GenBank: CP000647.1

FASTA Graphics

Go to:

```
LOCUS       CP000647            5315120 bp    DNA     circular BCT 22-JUN-2016
DEFINITION  Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete genome.
ACCESSION   CP000647
VERSION     CP000647.1
DBLINK      BioProject: PRJNA31
            BioSample: SAMN02603941
KEYWORDS    .
SOURCE      Klebsiella pneumoniae subsp. pneumoniae MGH 78578
  ORGANISM  Klebsiella pneumoniae subsp. pneumoniae MGH 78578
```

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

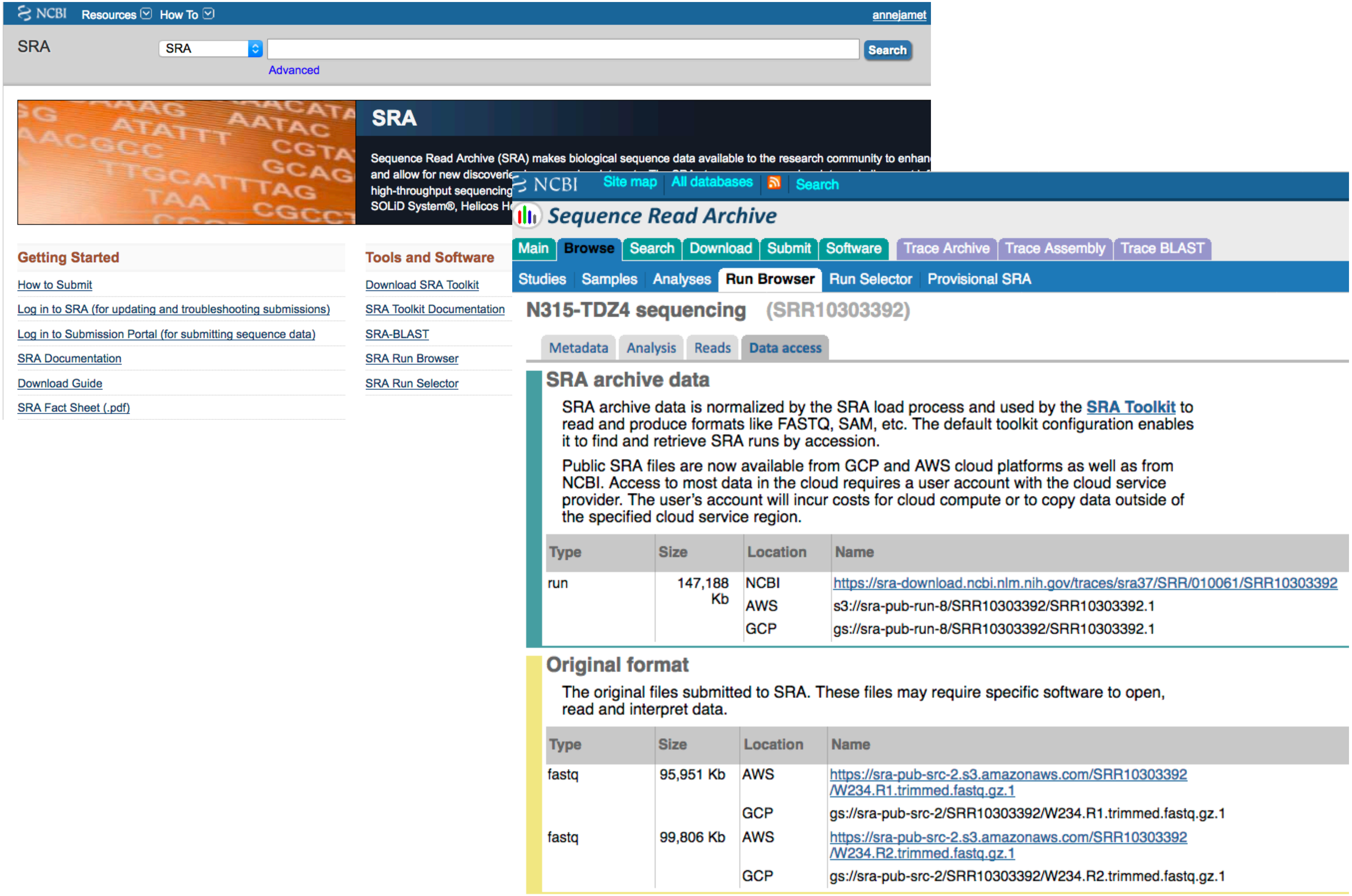
Download 1 item.

Format: GenBank (full)

Show GI □

Create File

Sequences brutes @NCBI



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhan
and allow for new discoverie
high-throughput sequencing
SOLiD System®, Helicos He

Getting Started

- How to Submit
- Log in to SRA (for updating and troubleshooting submissions)
- Log in to Submission Portal (for submitting sequence data)
- SRA Documentation
- Download Guide
- SRA Fact Sheet (.pdf)

Tools and Software

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

N315-TDZ4 sequencing (SRR10303392)

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.

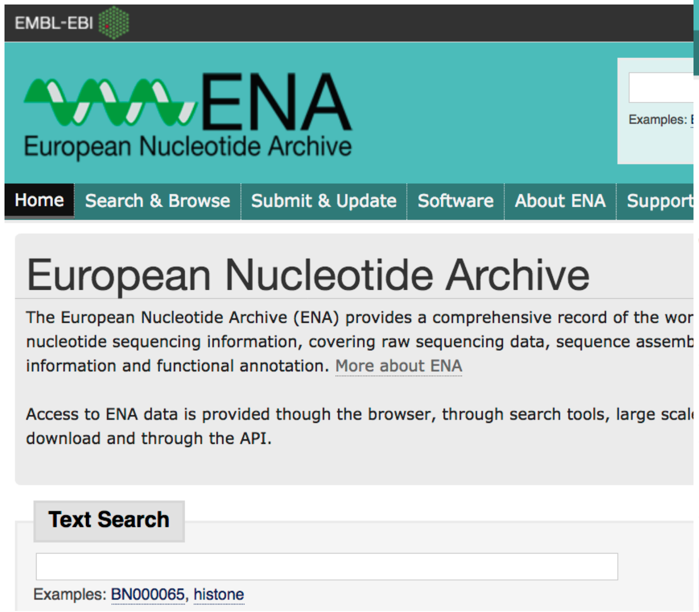
Type	Size	Location	Name
run	147,188 Kb	NCBI AWS GCP	https://sra-download.ncbi.nlm.nih.gov/traces/sra37/SRR/010061/SRR10303392 s3://sra-pub-run-8/SRR10303392/SRR10303392.1 gs://sra-pub-run-8/SRR10303392/SRR10303392.1

Original format

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

Type	Size	Location	Name
fastq	95,951 Kb	AWS GCP	https://sra-pub-src-2.s3.amazonaws.com/SRR10303392/W234.R1.trimmed.fastq.gz.1 gs://sra-pub-src-2/SRR10303392/W234.R1.trimmed.fastq.gz.1
fastq	99,806 Kb	AWS GCP	https://sra-pub-src-2.s3.amazonaws.com/SRR10303392/W234.R2.trimmed.fastq.gz.1 gs://sra-pub-src-2/SRR10303392/W234.R2.trimmed.fastq.gz.1

Sequences brutes @ENA



EMBL-EBI

ENA European Nucleotide Archive

Examples:

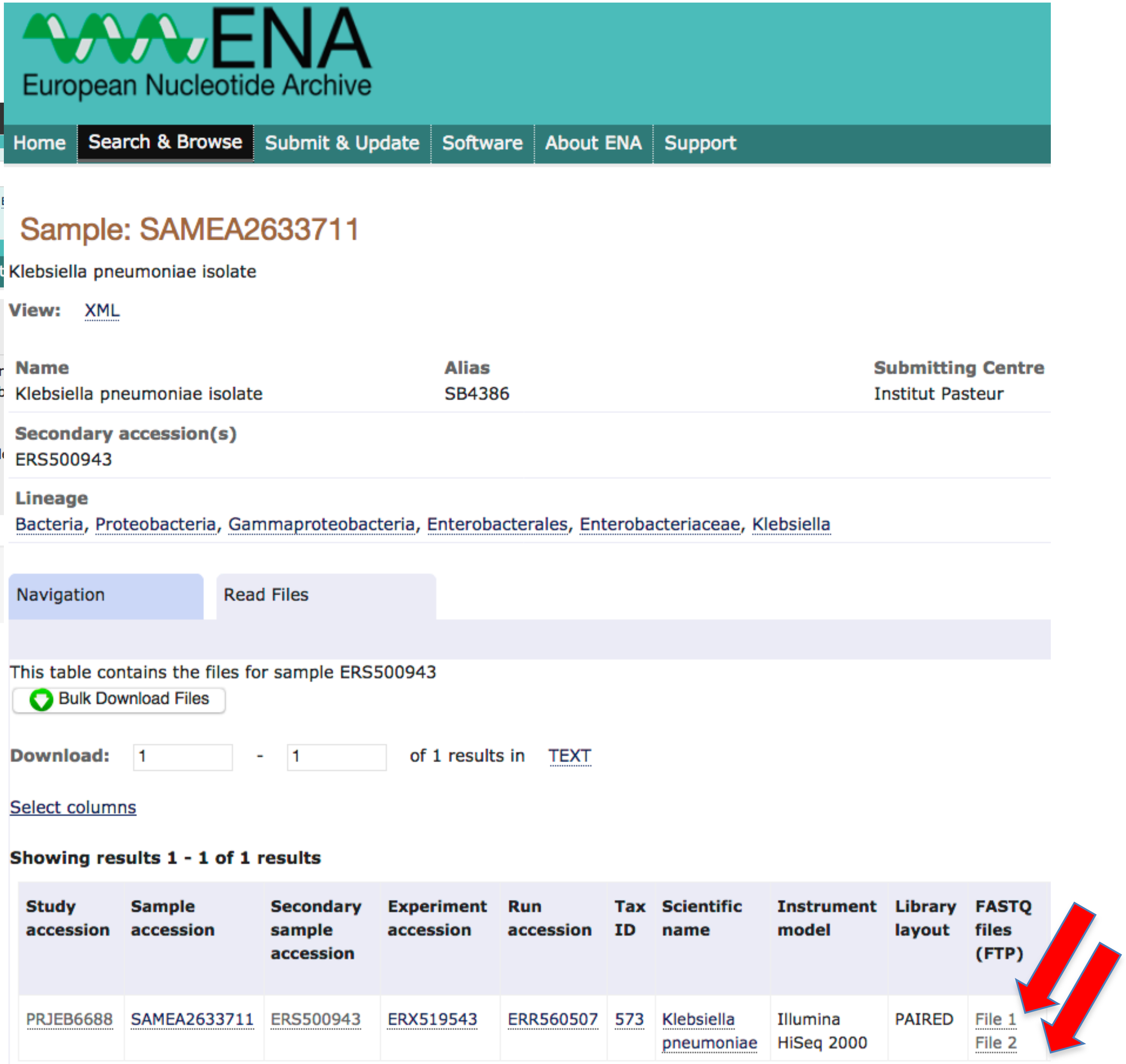
European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the wor nucleotide sequencing information, covering raw sequencing data, sequence assemb information and functional annotation. More about ENA

Access to ENA data is provided though the browser, through search tools, large scal download and through the API.

Text Search

Examples: BN000065, histone



ENA European Nucleotide Archive

Sample: SAMEA2633711

Klebsiella pneumoniae isolate

View: XML

Name	**Alias**	**Submitting Centre**
Klebsiella pneumoniae isolate	SB4386	Institut Pasteur

Secondary accession(s)

ERS500943

Lineage

Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacterales, Enterobacteriaceae, Klebsiella

This table contains the files for sample ERS500943

Download: 1 - 1 of 1 results in TEXT

Select columns

Showing results 1 - 1 of 1 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)
PRJEB6688	SAMEA2633711	ERS500943	ERX519543	ERR560507	573	Klebsiella pneumoniae	Illumina HiSeq 2000	PAIRED	File 1 File 2

Protein and domain DataBases

- Uniprot
  - ("Universal Protein Resource") : c'est la base de données des protéines : ExPASy Proteomics Server. Consortium [EBI - SIB - PIR].
- PDB ("Protein Data Bank”)
- PFAM et INTERPRO