Microbial sequence data

What’s out there?

Levels of sequence data

Raw data

- Amplicon
  - PCR product, usually Sanger sequence (.ab1, .fasta)
- Locus
  - Multiple overlapping amplicons assembled (.fasta)
- Genome
  - Whole genome sequencing reads (.fastq.gz)
- Transcriptome
  - RNA (cDNA) sequencing reads (.fastq.gz)

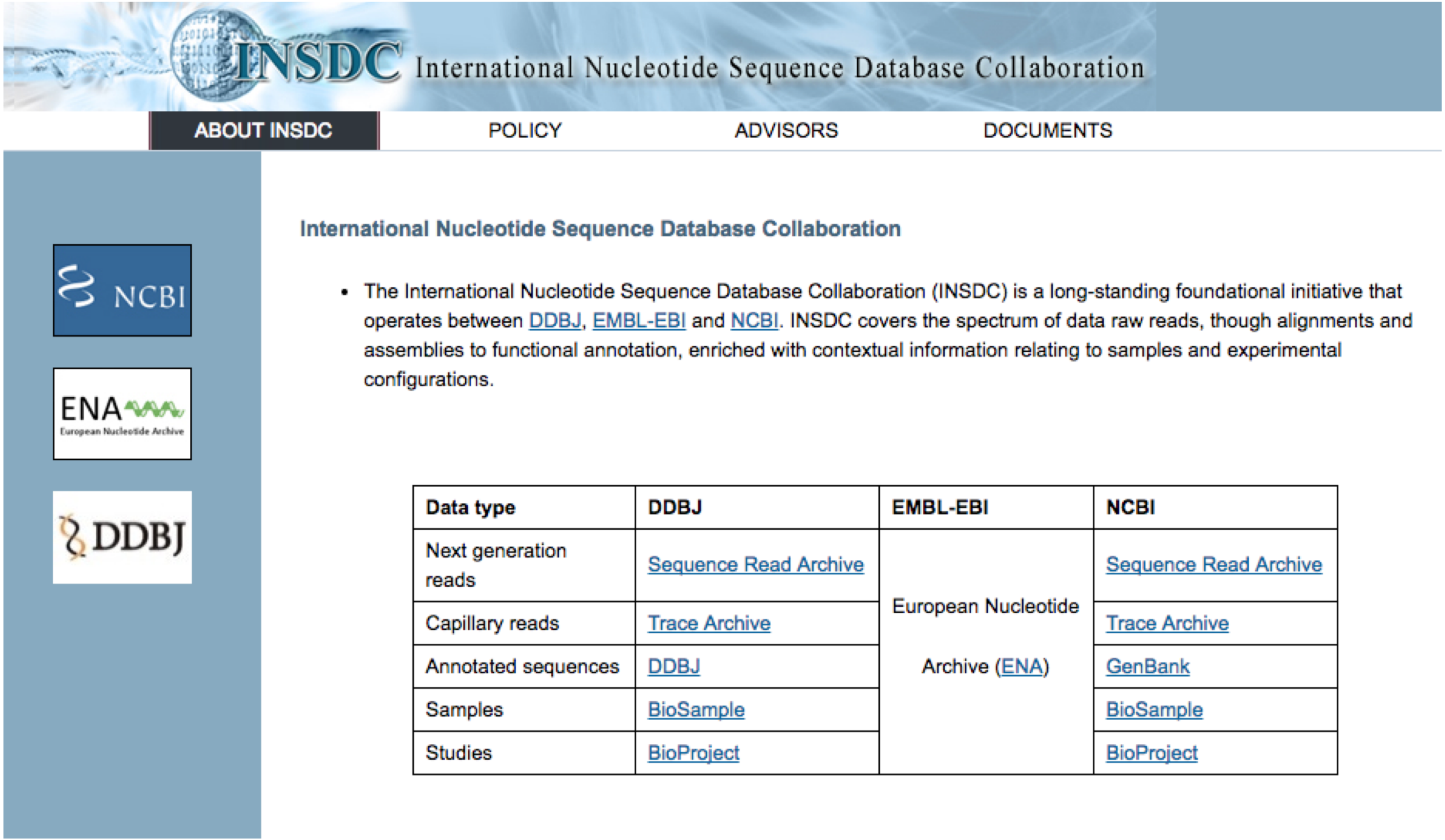
Derived data

- Assembled genome (.fasta)
  - Draft - multiple contigs
  - Complete - one contig per replicon
- Annotated genome (.gbk or .gff)
  - Genomic features labelled *e.g.* genes
- Protein sequences
  - Translated from predicted genes
  - Find in assembled transcripts

Curated data

- “Curated” means
  - Assessed for quality
  - Usually some human contribution
- Orthologs
  - Protein families
- Sequence “profiles”
  - Alignments of orthologous sequences
  - DNA or Protein

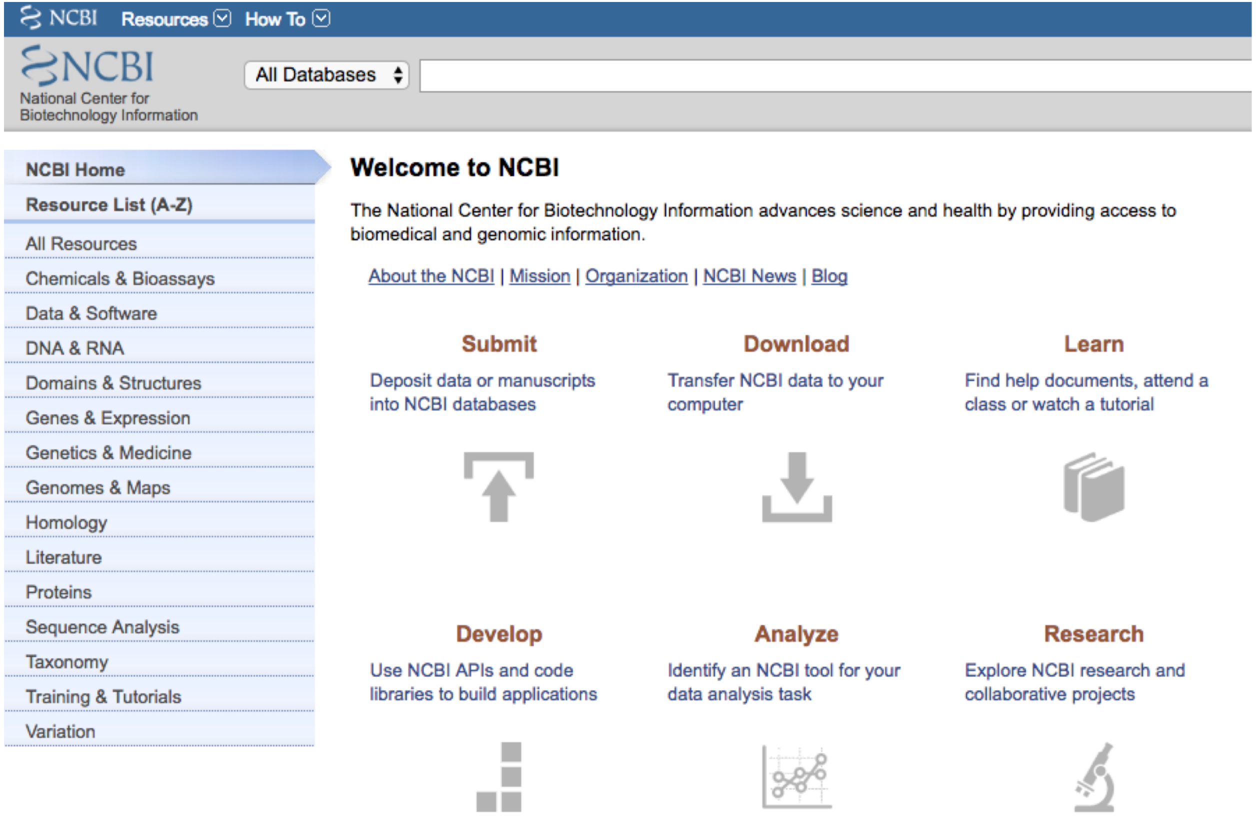
The INDSC


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

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, though 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



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

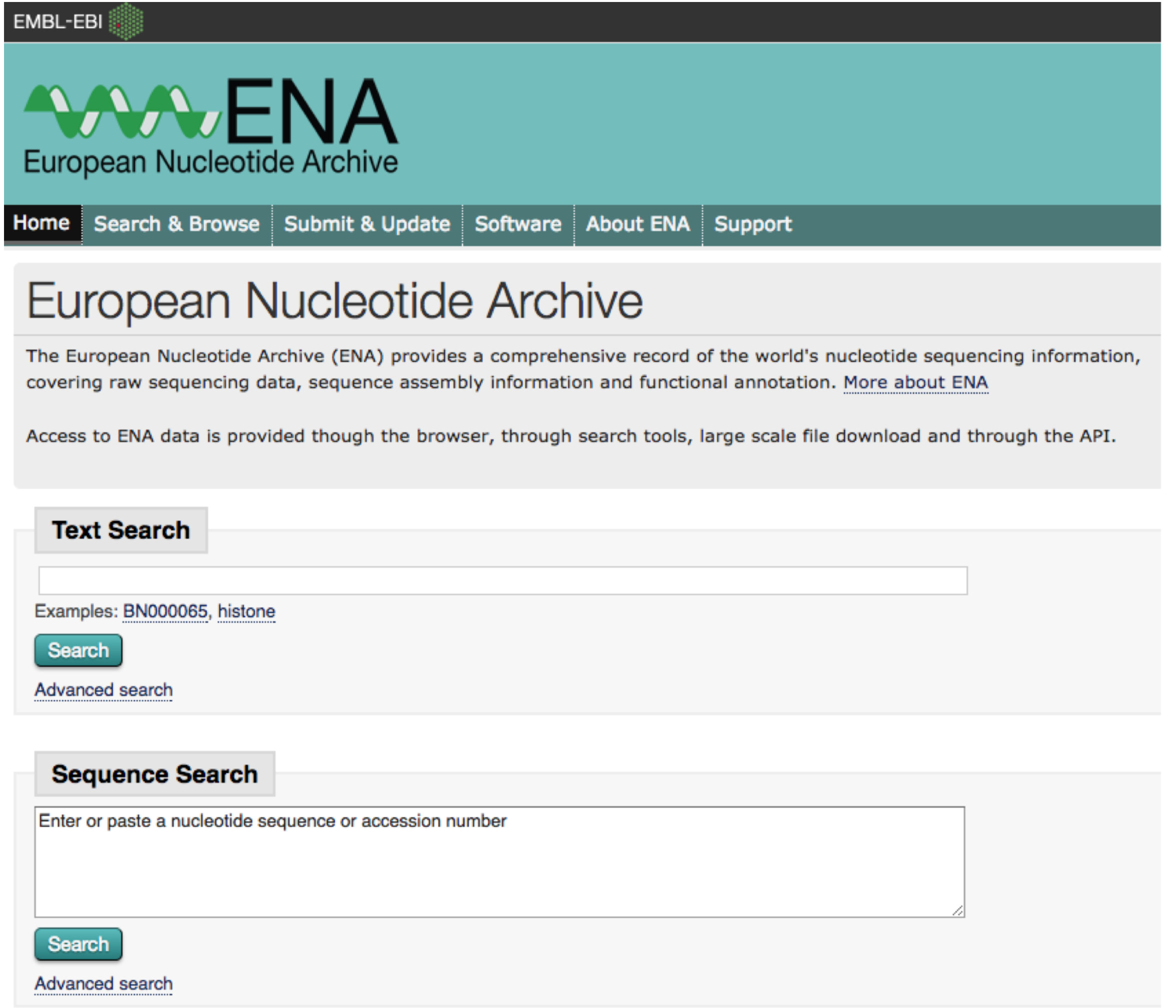
Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Hosted at NCBI in Washington, USA



European Nucleotide Archive

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

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

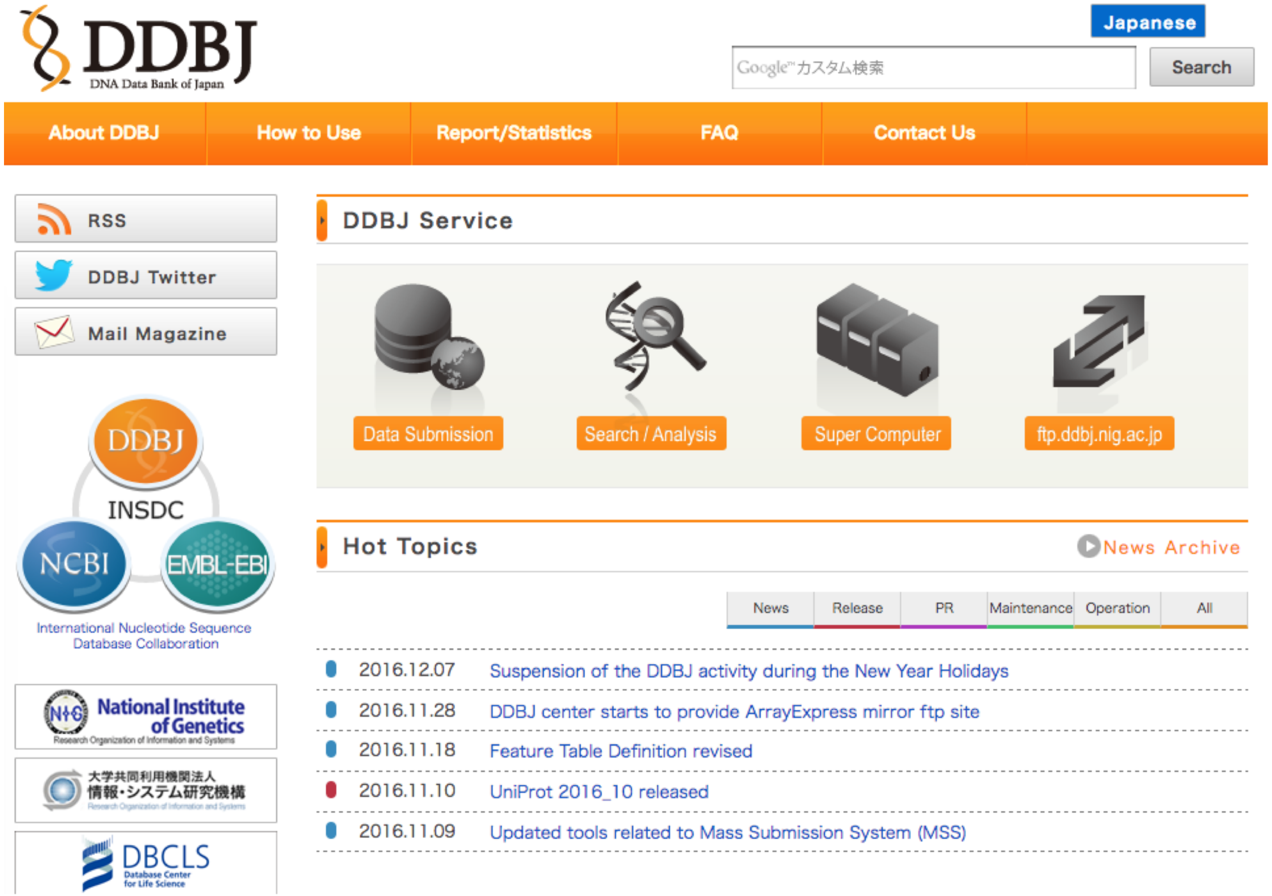
Text Search

Examples: BN000065, histone

Sequence Search

Enter or paste a nucleotide sequence or accession number

Hosted at EBI in Cambridge UK



DDBJ
DNA Data Bank of Japan

Japanese

Google™カスタム検索 Search

About DDBJ | How to Use | Report/Statistics | FAQ | Contact Us

- RSS
- DDBJ Twitter
- Mail Magazine

International Nucleotide Sequence Database Collaboration

National Institute of Genetics
Research Organization of Information and Systems

大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

DBCLS
Database Center for Life Science

DDBJ Service

Data Submission | Search / Analysis | Super Computer | ftp.ddbj.nig.ac.jp

Hot Topics

News Archive

News | Release | PR | Maintenance | Operation | All

2016.12.07	Suspension of the DDBJ activity during the New Year Holidays
2016.11.28	DDBJ center starts to provide ArrayExpress mirror ftp site
2016.11.18	Feature Table Definition revised
2016.11.10	UniProt 2016_10 released
2016.11.09	Updated tools related to Mass Submission System (MSS)

Hosted at NIG in Mishima, Japan

Sequence Read Archive (SRA)

The FASTQ dumping ground

- NCBI → “SRA”
  Download “.sra” files
  Convert with “fastq-dump”

- ENA → “ENA”
  Download “.fastq.gz” files

- DDBJ → “DRA”
  ???

Sample 1

Sample 2

Study

Sample 3

Sample 1

Experiment 1

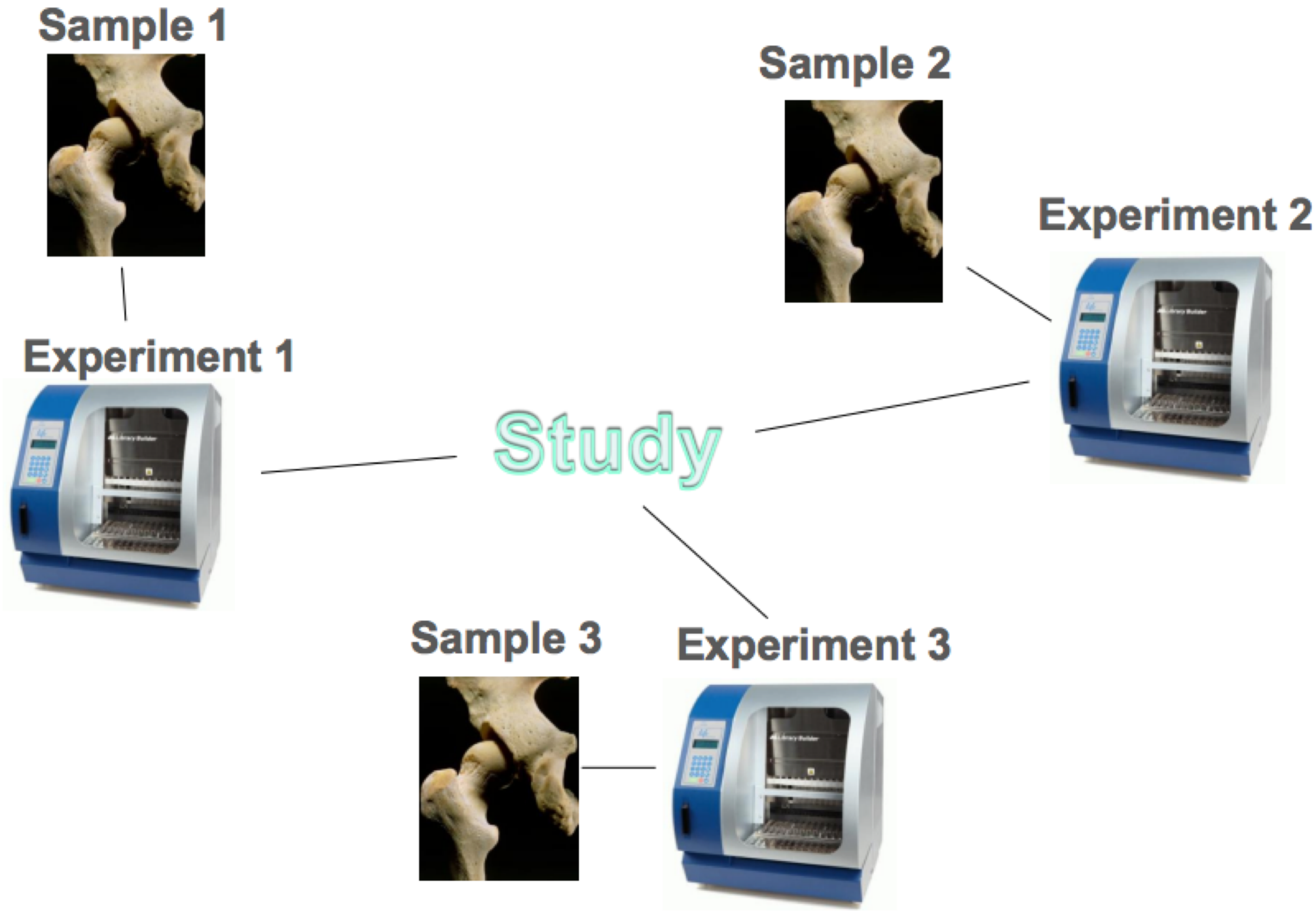
Study

Sample 2

Experiment 2

Sample 3

Experiment 3



Sample 1

Experiment 1

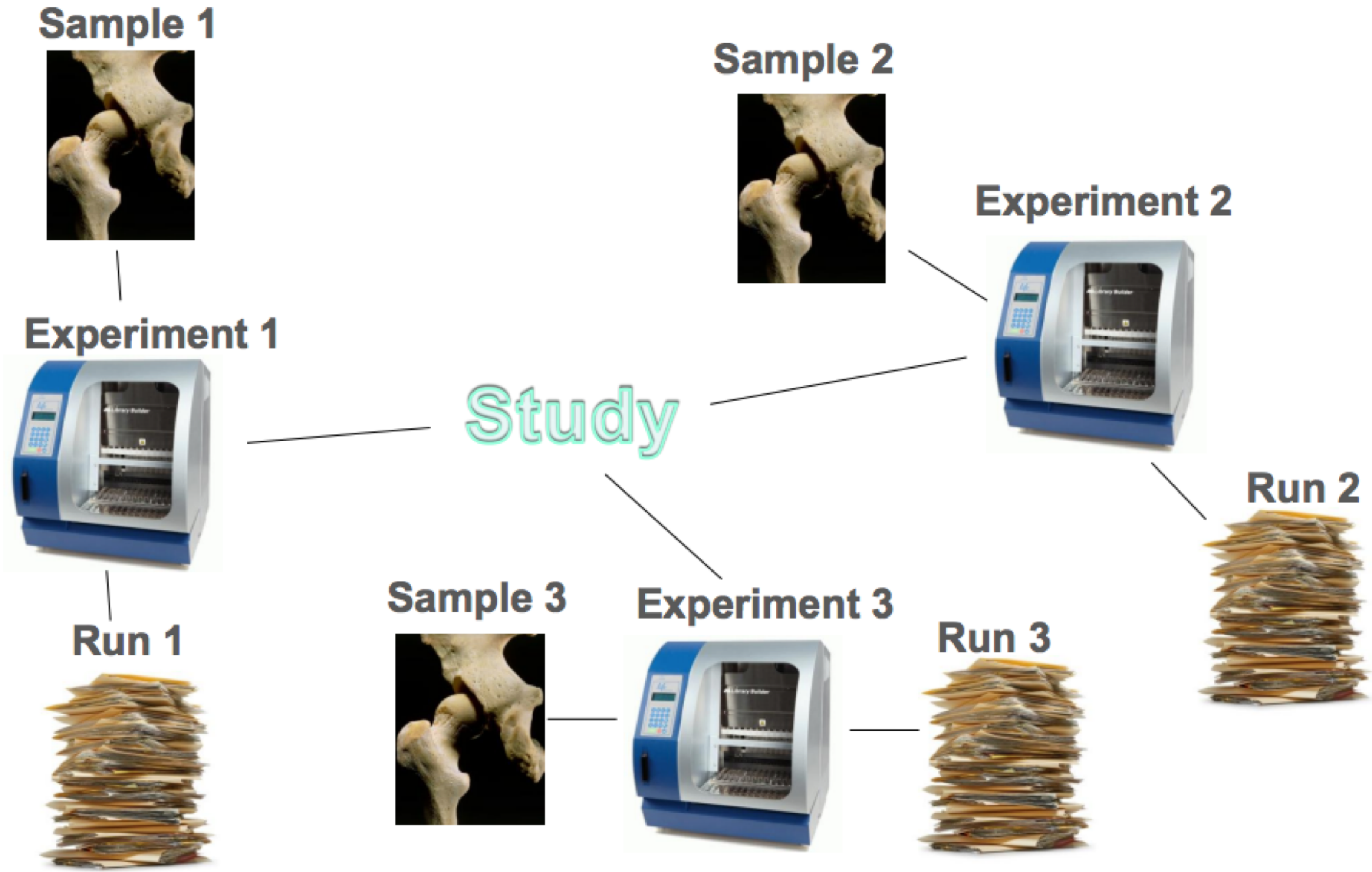
Study

Sample 2

Experiment 2

Sample 3

Experiment 3



Sample 1

Experiment 1

Run 1

Study

Sample 2

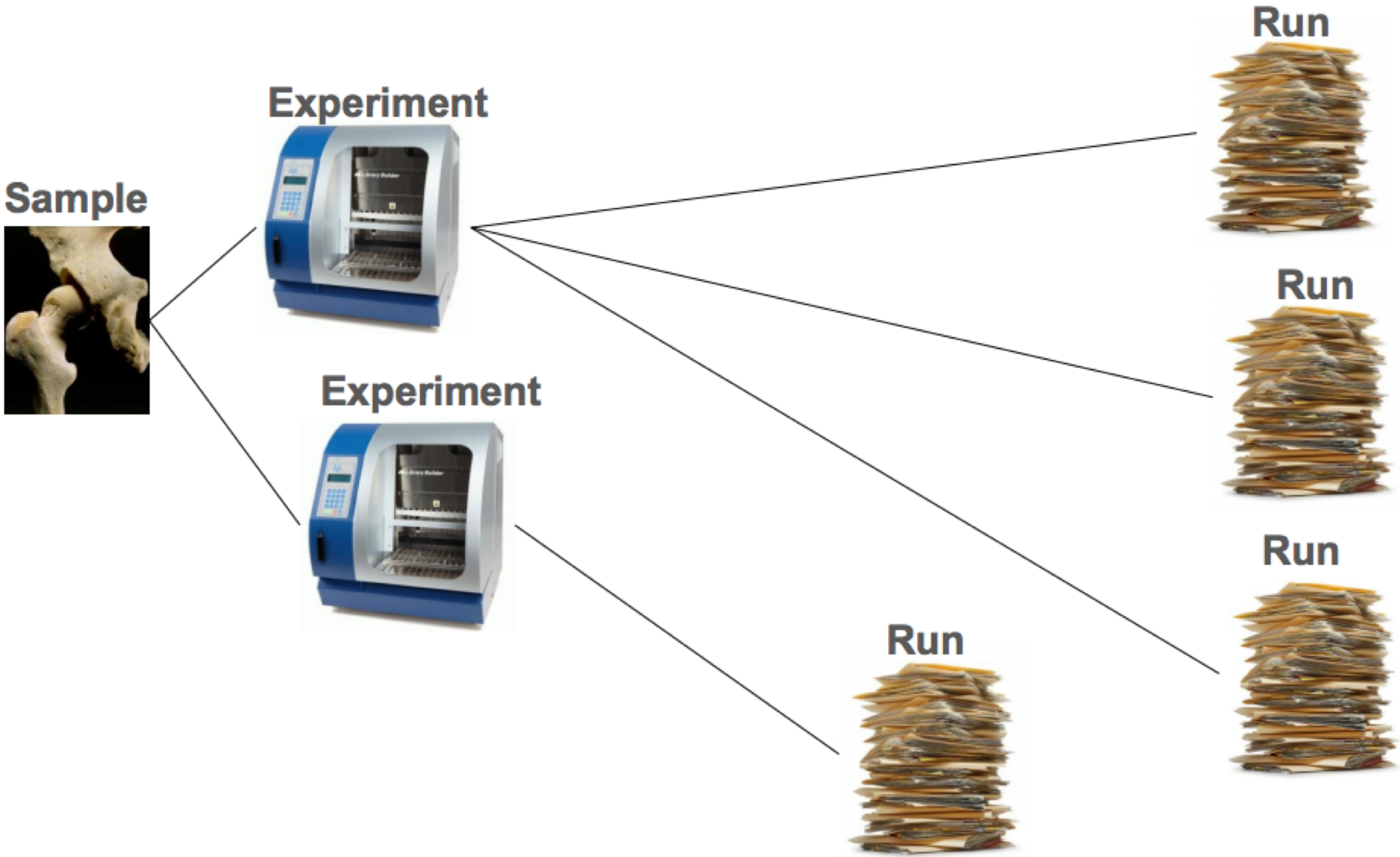
Experiment 2

Run 2

Sample 3

Experiment 3

Run 3



Sample

Experiment

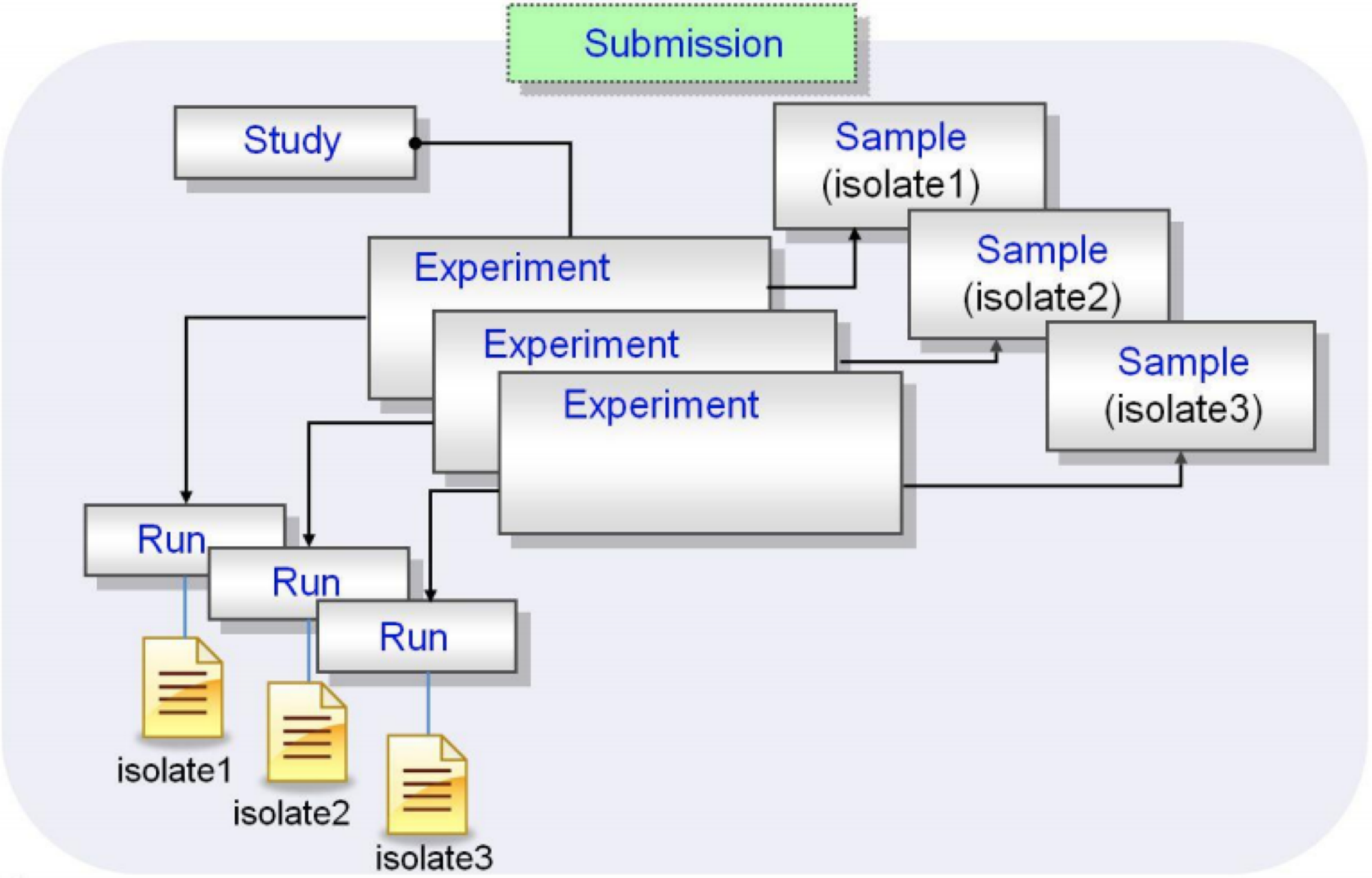
Experiment

Run

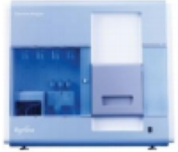
Run

Run

Run


Submission
Study
Sample
(isolate1)
Sample
(isolate2)
Sample
(isolate3)
Experiment
Experiment
Experiment
Run
Run
Run
isolate1
isolate2
isolate3

one ← many

Sample ←

Experiment

Experiment ←

Run

Study ←

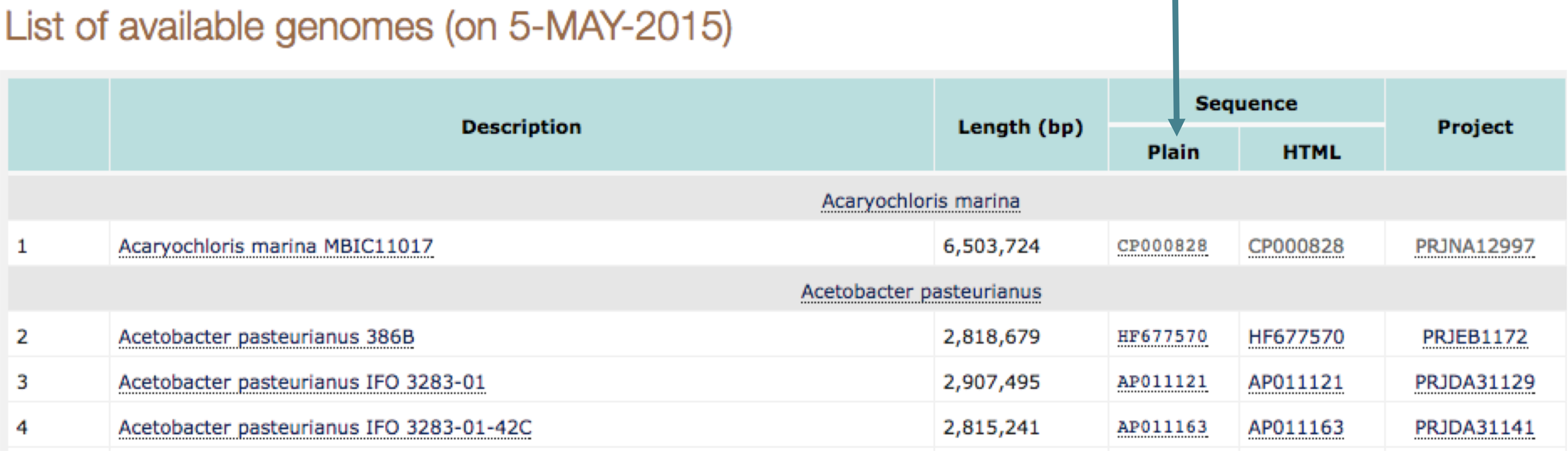
Experiment

Reference genomes

Reference genomes (ENA)

http://www.ebi.ac.uk/genomes/bacteria.html

When downloading use *“Sequence - Plain”*

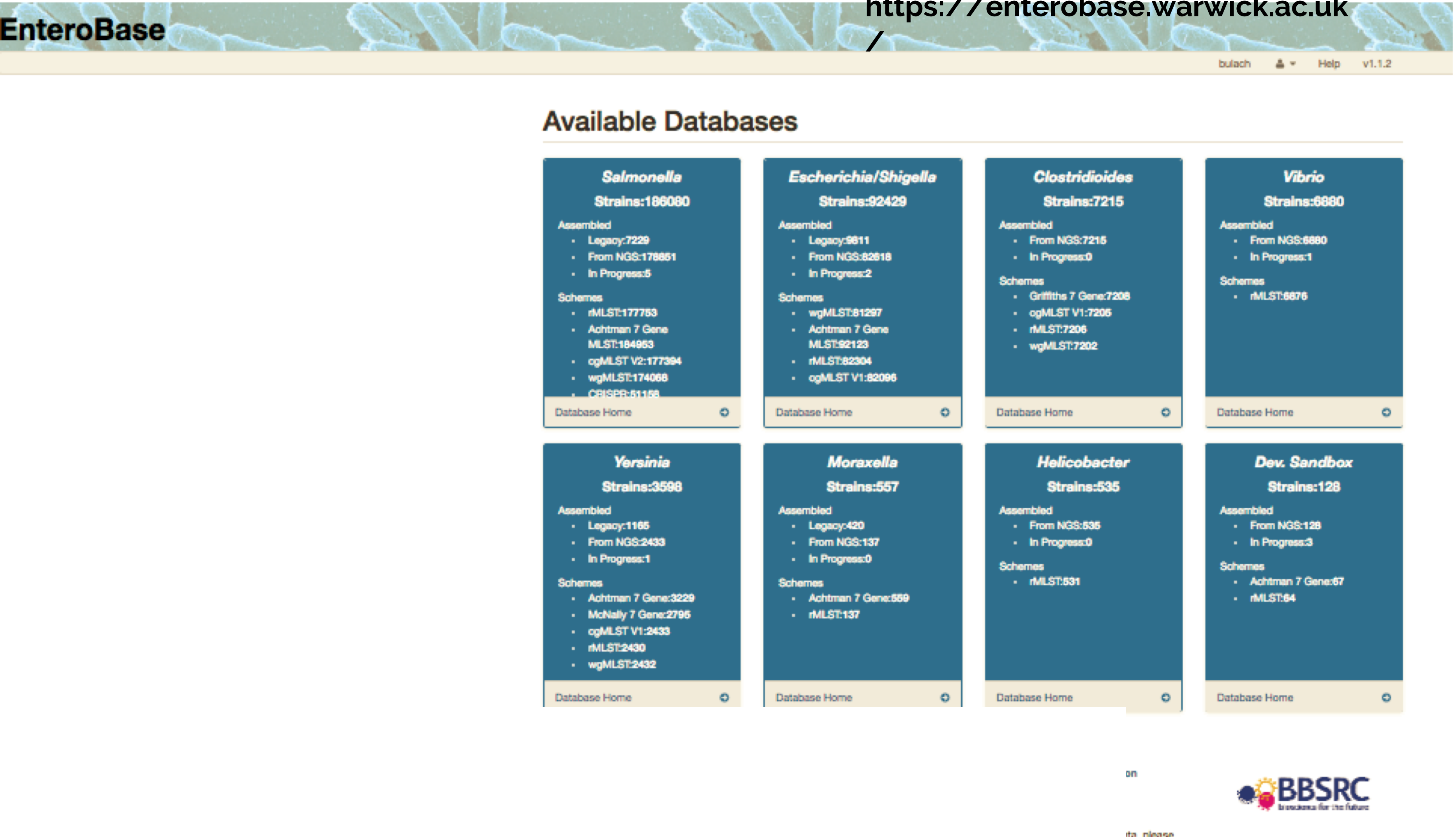


List of available genomes (on 5-MAY-2015)

	Description	Length (bp)	Sequence		Project
			Plain	HTML	
	Acaryochloris marina				
1	Acaryochloris marina MBIC11017	6,503,724	CP000828	CP000828	PRJNA12997
	Acetobacter pasteurianus				
2	Acetobacter pasteurianus 386B	2,818,679	HF677570	HF677570	PRJEB1172
3	Acetobacter pasteurianus IFO 3283-01	2,907,495	AP011121	AP011121	PRJDA31129
4	Acetobacter pasteurianus IFO 3283-01-42C	2,815,241	AP011163	AP011163	PRJDA31141

Curated databases

https://enterobase.warwick.ac.uk/



Available Databases

Salmonella
Strains:186080
Assembled
- Legacy:7229
- From NGS:178851
- In Progress:5

Schemes
- rMLST:177753
- Achtman 7 Gene MLST:184953
- cgMLST V2:177394
- wgMLST:174068
- CRISPR:51158

Database Home

Escherichia/Shigella
Strains:92429
Assembled
- Legacy:9811
- From NGS:82618
- In Progress:2

Schemes
- wgMLST:81297
- Achtman 7 Gene MLST:92123
- rMLST:82304
- cgMLST V1:82096

Database Home

Clostridioides
Strains:7215
Assembled
- From NGS:7215
- In Progress:0

Schemes
- Griffiths 7 Gene:7208
- cgMLST V1:7206
- rMLST:7206
- wgMLST:7202

Database Home

Vibrio
Strains:6880
Assembled
- From NGS:6880
- In Progress:1

Schemes
- rMLST:6876

Database Home

Yersinia
Strains:3598
Assembled
- Legacy:1165
- From NGS:2433
- In Progress:1

Schemes
- Achtman 7 Gene:3229
- McNally 7 Gene:2796
- cgMLST V1:2433
- rMLST:2430
- wgMLST:2432

Database Home

Moraxella
Strains:557
Assembled
- Legacy:420
- From NGS:137
- In Progress:0

Schemes
- Achtman 7 Gene:559
- rMLST:137

Database Home

Helicobacter
Strains:535
Assembled
- From NGS:535
- In Progress:0

Schemes
- rMLST:531

Database Home

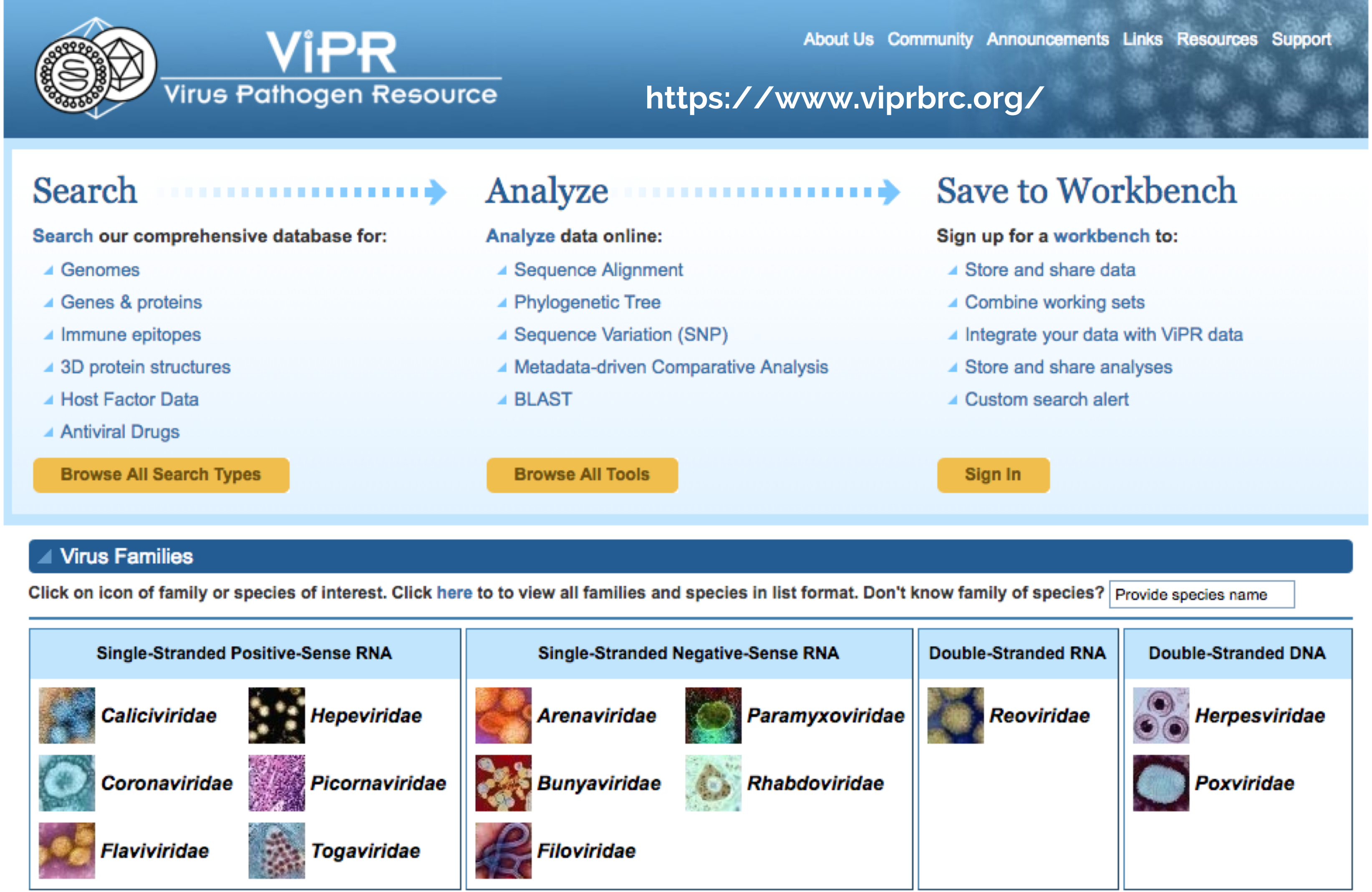
Dev. Sandbox
Strains:128
Assembled
- From NGS:128
- In Progress:3

Schemes
- Achtman 7 Gene:67
- rMLST:64

Database Home

on

ta. please



ViPR
Virus Pathogen Resource

About Us Community Announcements Links Resources Support

https://www.viprbrc.org/

Search

Search our comprehensive database for:

- Genomes
- Genes & proteins
- Immune epitopes
- 3D protein structures
- Host Factor Data
- Antiviral Drugs

Browse All Search Types

Analyze

Analyze data online:

- Sequence Alignment
- Phylogenetic Tree
- Sequence Variation (SNP)
- Metadata-driven Comparative Analysis
- BLAST

Browse All Tools

Save to Workbench

Sign up for a workbench to:

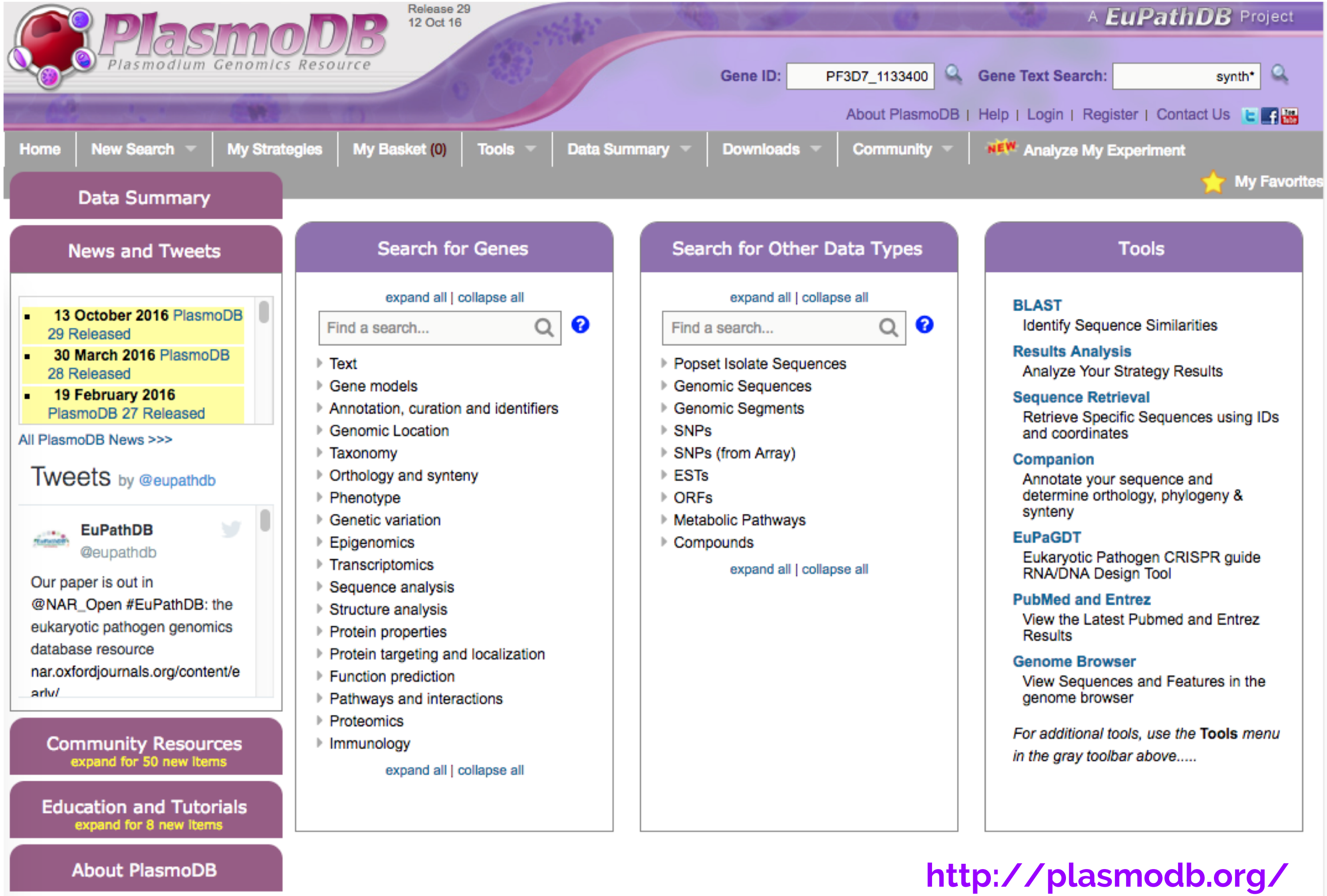
- Store and share data
- Combine working sets
- Integrate your data with ViPR data
- Store and share analyses
- Custom search alert

Sign In

Virus Families

Click on icon of family or species of interest. Click here to to view all families and species in list format. Don't know family of species? Provide species name

Single-Stranded Positive-Sense RNA		Single-Stranded Negative-Sense RNA		Double-Stranded RNA	Double-Stranded DNA
Caliciviridae	***Hepeviridae***	***Arenaviridae***	***Paramyxoviridae***	***Reoviridae***	***Herpesviridae***
Coronaviridae	***Picornaviridae***	***Bunyaviridae***	***Rhabdoviridae***		***Poxviridae***
Flaviviridae	***Togaviridae***	***Filoviridae***			



Data Summary

News and Tweets

- 13 October 2016 PlasmoDB 29 Released
- 30 March 2016 PlasmoDB 28 Released
- 19 February 2016 PlasmoDB 27 Released

All PlasmoDB News >>>

Tweets by @eupathdb

EuPathDB
@eupathdb

Our paper is out in @NAR_Open #EuPathDB: the eukaryotic pathogen genomics database resource nar.oxfordjournals.org/content/e arly/

Community Resources
expand for 50 new items

Education and Tutorials
expand for 8 new items

About PlasmoDB

Search for Genes

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
- Epigenomics
- Transcriptomics
- Sequence analysis
- Structure analysis
- Protein properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

expand all | collapse all

Search for Other Data Types

expand all | collapse all

Find a search...

- Popset Isolate Sequences
- Genomic Sequences
- Genomic Segments
- SNPs
- SNPs (from Array)
- ESTs
- ORFs
- Metabolic Pathways
- Compounds

expand all | collapse all

Tools

BLAST
Identify Sequence Similarities

Results Analysis
Analyze Your Strategy Results

Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates

Companion
Annotate your sequence and determine orthology, phylogeny & synteny

EuPaGDT
Eukaryotic Pathogen CRISPR guide RNA/DNA Design Tool

PubMed and Entrez
View the Latest Pubmed and Entrez Results

Genome Browser
View Sequences and Features in the genome browser

For additional tools, use the **Tools** *menu in the gray toolbar above.....*

http://plasmodb.org/

Conclusions

Conclusions

- Be skeptical!
  - ENA + Genbank accept anything
  - Garbage in, Garbage out

- Curated data
  - Refseq vs GenBank
  - Specialised sites (better QC)
  - Be wary of draft assemblies
  - Go back to primary reads

I'D TURN BACK
IF I WERE YOU!

Further reading

https://www.ncbi.nlm.nih.gov/core/assets/sra/files/Factsheet_SRA.pdf

https://www.ddbj.nig.ac.jp/dra/index-e.html

https://p.ddbj.nig.ac.jp/pipeline/Login.do