Food and Agriculture Organization of the United Nations

# Módulo 2

# Intercambio de datos (NCBI y bases de datos de cepas concretas)

**Ayorinde O. Afolayan**

*Universidad de Ibadán, Estado de Oyo (Nigeria)*

## Descargo de responsabilidad

*La traducción de esta presentación del idioma inglés al francés fue realizada por la FAO en colaboración con el Instituto Nacional de Alimentos de la Universidad Técnica de Dinamarca (DTU), que actúa como Centro de Referencia de la FAO para la Resistencia a los Antimicrobianos en Dinamarca. Este trabajo se llevó a cabo en el marco del Proyecto de Seguridad Sanitaria Global de la FAO - Vigilancia de Alerta Temprana de Enfermedades Infecciosas, financiado por la United States Agency for International Development (USAID) (Agencia de los Estados Unidos para el Desarrollo Internacional).*

*La versión original en inglés de esta presentación fue desarrollada por la DTU como parte de la subvención regional SeqAfrica del Fondo Fleming, financiada por el Departamento de Salud y Asistencia Social del Reino Unido.*

# Esquema

- Visión general de GenBank y el NCBI
- Envío de datos de secuencias a GenBank
- Visión general de las bases de datos de cepas concretas
- Envío de datos a la base de datos del complejo de especies *Klebsiella*

# GenBank

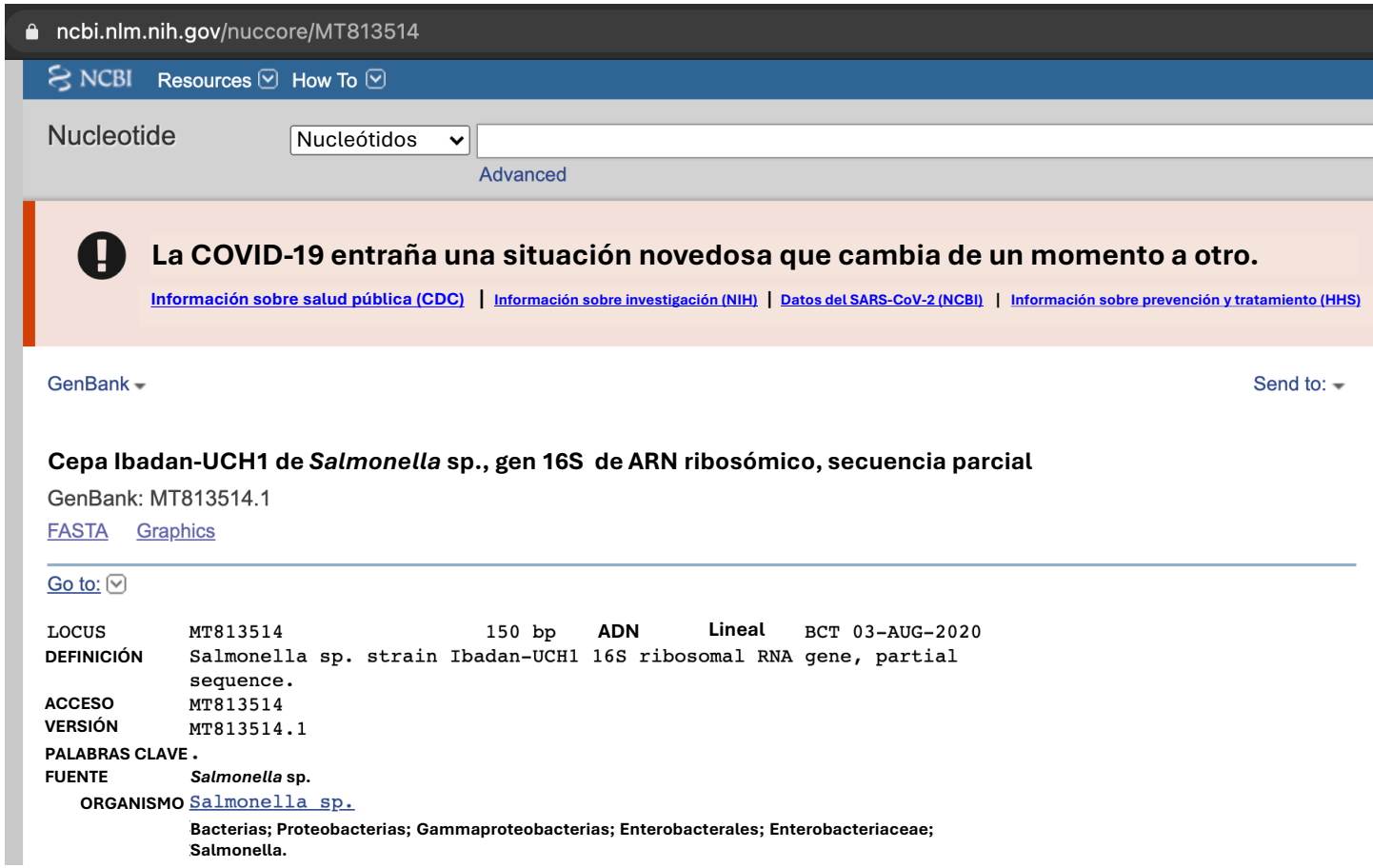
- Base de datos de secuencias que contiene secuencias de ADN (y anotaciones) de acceso público enviadas por la comunidad científica de todo el mundo (https://www.ncbi.nlm.nih.gov/genbank/).
- Colaboración internacional de bases de datos de secuencias nucleotídicas:

  - Banco de datos de ADN de Japón (DDBJ),
  - Archivo Europeo de Nucleótidos (ENA) y
  - GenBank del NCBI

- Estas tres organizaciones intercambian datos a diario.

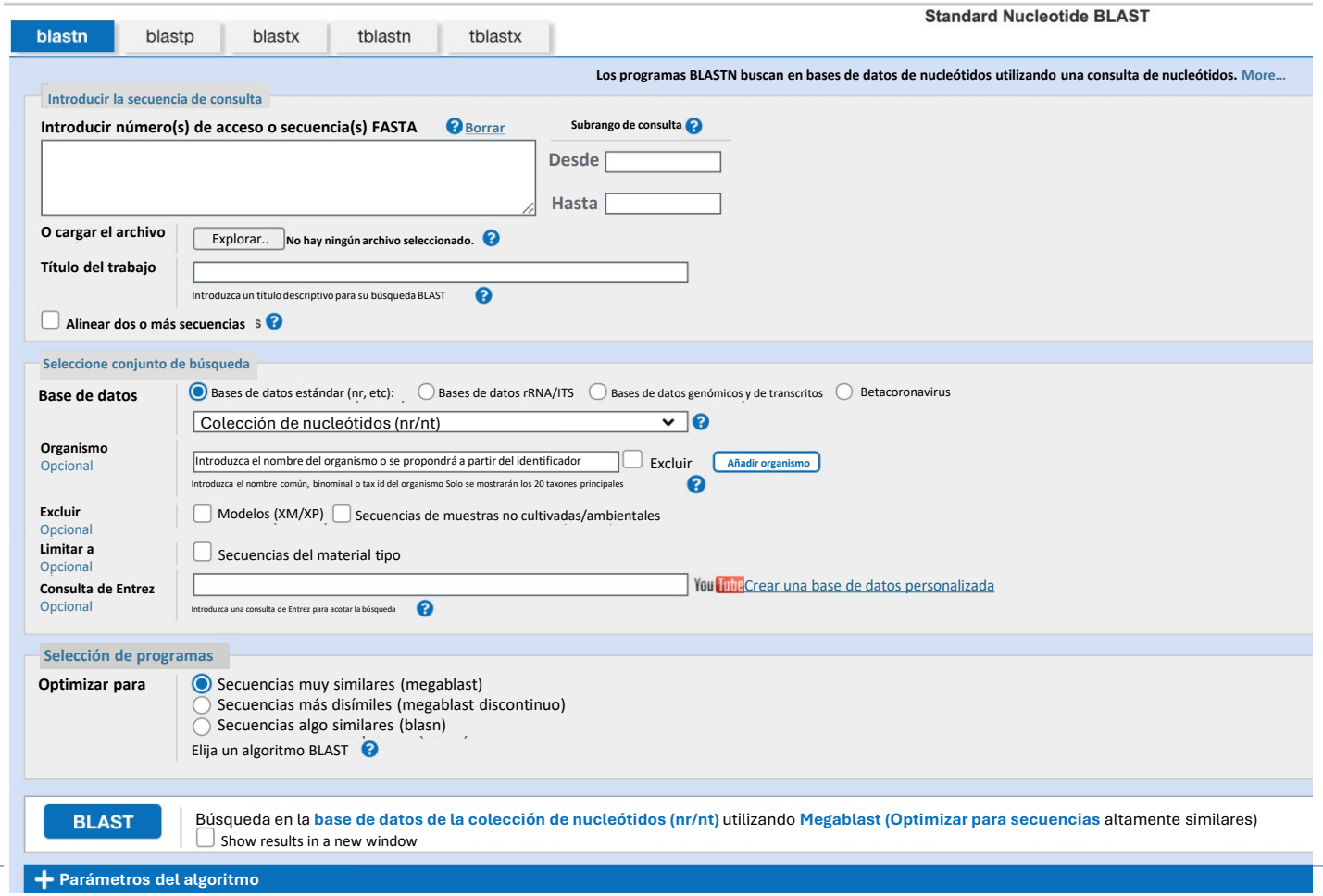
# Cómo buscar y extraer datos de GenBank

- A través del sitio web Entrez Nucleotide del NCBI (https://www.ncbi.nlm.nih.gov/nucleotide/). Las consultas de entradas (en la barra de búsqueda) pueden ser, p. ej., los números de acceso o el nombre del organismo



# Cómo buscar y extraer datos de GenBank

- Buscar secuencias de GenBank idénticas a las nuestras utilizando los distintos instrumentos de BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)



# Envíos a GenBank: inscribirse

- Es aconsejable inscribirse para tener una cuenta del NCBI antes de iniciar el proceso de envío (https://www.ncbi.nlm.nih.gov/account/?back_url=https%3A%2F%2Fwww.ncbi.nlm.nih.gov%2F)

# Envíos a GenBank: tipos de envíos

- BankIt https://www.ncbi.nlm.nih.gov/WebSub/ nos dirige a la base de datos de envíos adecuada

- Portal de envíos https://submit.ncbi.nlm.nih.gov/

## Enviar nuevas secuencias a GenBank

### ¿Qué tipo de datos de secuencia tiene?

- SARS-CoV-2
- ARN ribosómico (ARNr) o ARNr-ITS
- COX1 metazoico (animal multicelular)
- Virus de la gripe
- Norovirus
- Virus del dengue
- Genomas eucarióticos y procarióticos (secuenciación del genoma completo o completos)
- Ensamblado transcriptómico aleatorio (TSA)
- Lecturas de secuencias no ensambladas (SRA)

**Datos de secuencias no enumerados anteriormente (a través de BankIt): ARNm, ADN genómico, orgánulos, ARNnc, plásmidos, otros virus, bacteriófagos, construcciones sintéticas**

Requisitos para los envíos:

- Datos de secuencias en formato FASTA o de alineamiento
- Nombre del organismo u organismos en los que se aislaron los datos de la secuencia y cualquier otro dato descriptivo
- Características de la secuencia (ejemplos: CDS, gen, ARNt, con intervalos de nucleótidos y nombres de productos)

Inicio

## Portal de envío

Enviar al mayor repositorio público de información biológica y científica del mundo

Escriba unas palabras sobre los datos de la secuencia que va a enviar y seleccione una opción para obtener más información. También puede consultar la información de envío que aparece a continuación.

**¿Qué desea enviar?**

Escriba unas palabras sobre los datos de la secuencia.

Indique el tipo de secuencia | Sugerir herramienta

SARS-CoV-2 | 16S rRNA | genoma | ITS | SRA

# Envíos a GenBank: Requisitos para los envíos

- El envío debe incluir información sobre el organismo de origen y la anotación proporcionada por el remitente. (P. ej., identidad del organismo, fuente de recogida, localización geográfica, fecha de recogida, tipo de secuenciador, etc.)

- No deben enviarse:
  - Secuencias no contiguas
  - Secuencias de cebadores
  - Secuencias de proteínas sin envío de nucleótidos subyacentes
  - Secuencias que contienen una mezcla de secuencias genómicas y de ARNm
  - Secuencias sin equivalente físico (secuencias de consenso)
  - Secuencias con una longitud inferior a 200 nucleótidos

# Ejemplo: Envío al SRA

## Gestionar datos > Bioproyecto: PRJNA604723

| | |
|---|---|
| Acceso al bioproyecto | **PRJNA60723** **Análisis metagenómico de las comunidades bacterianas del agua y el suelo de los fulani y otros grupos de Nigeria**<br>7 Biomuestras<br>7 SRA |
| Estado | **Publicado en** |
| Fecha de publicación | 2020-05-20 |
| Creado | 2020-02-04 15:43 |
| Actualizado | 2021-03-16 16:13 |
| Título | Análisis metagenómico de las comunidades bacterianas del agua y el suelo de los fulani y otros grupos de Nigeria<br>Editar |
| Descripción | Este estudio pretendía analizar la comunidad bacteriana en las muestras ambientales obtenidas de una población nómada fulani y una población urbanizada estudiadas recientemente.<br>Editar |
| Pertinencia | Ambiental |

Subvenciones

| ID | Título | Entidad |
|---|---|---|
| ICM-2016-03232 | La microbiota intestinal de los fulani | Agencia Austriaca de Cooperación Internacional para la Educación, la Ciencia y la Investigación |

Editar

Publicaciones

Análisis metagenómico de las comunidades bacterianas del agua y el suelo de los fulani y otros grupos de Nigeria

Afolayan, A. O., Ayeni, F. A., Journal of Infection in Developing Countries. 30 de diciembre de 2020; 12(14)

Editar

# Ejemplo: Envío al SRA

**SRA (7)** Biomuestra (7)

Editar Seleccione los datos que desee modificar utilizando las casillas de verificación siguientes

Buscar

| Acceso | Título | Identificación de la biblioteca | Archivos | Nombre de la muestra | Estado | Fechas de publicación |
|---|---|---|---|---|---|---|
| SRR11015236 | Metagenomics_of_soil_sample_1_from_Jarawa_community | No_disponible | SJo1_S117_L001_R1_001.fastq.gz<br>SJo1_S117_L001_R2_001.fastq.gz | SJo1 | Publicado en | 2020-05-20 |
| SRR11015235 | Metagenomics_of_soil_sample_2_from_Jarawa_community | No_disponible1 | SJo2_S118_L001_R1_001.fastq.gz<br>SJo2_S118_L001_R2_001.fastq.gz | SJo2 | Publicado en | 2020-05-20 |
| SRR11015234 | Metagenomics_of_soil_sample_1_from_Fulani_community | No_disponible2 | SJe1_S115_L001_R1_001.fastq.gz<br>SJe1_S115_L001_R2_001.fastq.gz | SJe1 | Publicado en | 2020-05-20 |
| SRR11015233 | Metagenomics_of_soil_sample_2_from_Fulani_community | No_disponible3 | SJe2_S116_L001_R1_001.fastq.gz<br>SJe2_S116_L001_R2_001.fastq.gz | SJe2 | Publicado en | 2020-05-20 |
| SRR11015232 | Metagenomics_of_water_sample_1_from_Fulani_community | No_disponible4 | FJW1_S120_L001_R1_001.fastq.gz<br>FJW1_S120_L001_R2_001.fastq.gz | FJW1 | Publicado en | 2020-05-20 |
| SRR11015231 | Metagenomics_of_water_sample_2_from_Fulani_community | No_disponible5 | FJW2_S121_L001_R1_001.fastq.gz<br>FJW2_S121_L001_R2_001.fastq.gz | FJW2 | Publicado en | 2020-05-20 |
| SRR11015230 | Metagenomics_of_water_sample_1_from_Jarawa_community | No_disponible6 | JW_S122_L001_R1_001.fastq.gz<br>JW_S122_L001_R2_001.fastq.gz | JW | Publicado en | 2020-05-20 |

# Envíos a GenBank: Tras la aprobación, ¿qué sigue?

6 Envíos

| Envío | Título | Estado | | Actualizado |
|---|---|---|---|---|
| SUB8485181 | Bacterias lácticas de Kunu, 8 de noviembre de 2020 | ✔ Bioproyecto: **Tramitado**<br>PRJNA677960 : Bacterias lácticas de Kunu, 8 de noviembre de 2020<br>✔ Biomuestra: **Tramitada** (objetos)<br>Descargar archivo de atributos con accesos de biomuestra<br>✖ SRA: **Suprimida** | Gestionar datos | 16 de noviembre |
| SUB6921263 | Análisis metagenómico de las comunidades bacterianas del agua y el suelo de los fulani y otros grupos de Nigeria, 4 de febrero de 2020 | ✔ Biomuestra: **Tramitada**<br>PRJNA604723 : Análisis metagenómico de las comunidades bacterianas del agua y el suelo de los fulani y otros grupos de Nigeria, 4 de febrero de 2020<br>Prefijos de etiquetas de locus:<br>• Ninguno (SAMN13979619)<br>• Ninguno (SAMN13979620)<br>locustagprefix.txt<br>✔ Biomuestra: **Tramitada**<br>Cargado correctamente<br>(7 objetos)<br>Descargar archivo de atributos con accesos de biomuestra<br>✔ SRA: **Tramitado**<br>(7 objetos)<br>Descargar el archivo de metadatos con los accesos al SRA<br>Ver y gestionar mis datos de envíos al SRA | Gestionar datos | 20 de mayo |
| SUB7324554 | Determinación de las características de aislados de bacterias lácticas de la leche materna y de las heces de neonatos como posibles probióticos, 24 de abril de 2020 | ✔ Bioproyecto: **Tramitado**<br>PRJNA628165 : Determinación de las características de aislados de bacterias lácticas de la leche materna y de las heces de neonatos como posibles probióticos, 24 de abril de 2020<br>✔ Biomuestra: **Tramitada**<br>Cargado correctamente<br>(85 objetos)<br>Descargar archivo de atributos con accesos de biomuestra<br>✔ SRA: **Tramitado**<br>(85 objetos)<br>Descargar el archivo de metadatos con los accesos al SRA<br>Ver y gestionar mis datos de envíos al SRA | Gestionar datos | 29 de abril |

# Bases de datos de cepas concretas

- **PubMLST** (https://pubmlst.org/ ):
  Base de datos de libre acceso que contiene información genotípica y fenotípica de más de 100 microbios

# Bases de datos de cepas concretas

- **Enterobase**

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

**Bases de datos disponibles**

***Salmonella***
**Cepas:289925**
Ensamblados
- Preexistentes:**4932**
- De secuenciación de nueva generación:**284993**
- En curso:**5047**

Esquemas
- Achtman 7 Gene MLST:**289737**
- cgMLST V2 + HierCC V1:**284658**
- rMLST:**284636**
- wgMLST:**284219**

Inicio de la base de datos

***Escherichia/Shigella***
**Cepas:166108**
Ensamblados
- Preexistentes:**4932**
- De secuenciación de nueva generación:**156583**
- En curso:**2429**

Esquemas
- Achtman 7 Gene MLST:**166025**
- cgMLST V1 + HierCC V1:**156664**
- rMLST:**156461**
- wgMLST:**156225**

Inicio de la base de datos

***Streptococcus***
**Cepas:72798**
Ensamblados
- De secuenciación de nueva generación:**72798**
- En curso:**122**

Esquemas
- cgMLSTv1:**72781**

Inicio de la base de datos

***Clostridioides***
**Cepas:21326**
Ensamblados
- De secuenciación de nueva generación:**21326**
- En curso:**412**

Esquemas
- cgMLST V1 + HierCC V1:**21314**
- Griffiths 7 Gene:**21325**
- rMLST:**21325**
- wgMLST:**21312**

Inicio de la base de datos

***Vibrio***
**Cepas:11648**
Ensamblados
- De secuenciación de nueva generación:**11648**
- En curso:**83**

Esquemas
- rMLST:**11648**
- Vibrio cgMLST + HierCC:**11646**

Inicio de la base de datos

***Helicobacter***
**Cepas:5130**
Ensamblados
- Preexistentes:**1869**
- De secuenciación de nueva generación:**3261**
- En curso:**103**

Esquemas
- rMLST:**3257**

Inicio de la base de datos

***Yersinia***
**Cepas:4627**
Ensamblados
- Preexistentes:**861**
- De secuenciación de nueva generación:**3766**
- En curso:**3**

Esquemas
- Achtman 7 Gene:**4405**
- cgMLST V1 + HierCC V1:**3784**
- McNally 7 Gene:**3985**
- rMLST:**3767**
- wgMLST:**3766**

Inicio de la base de datos

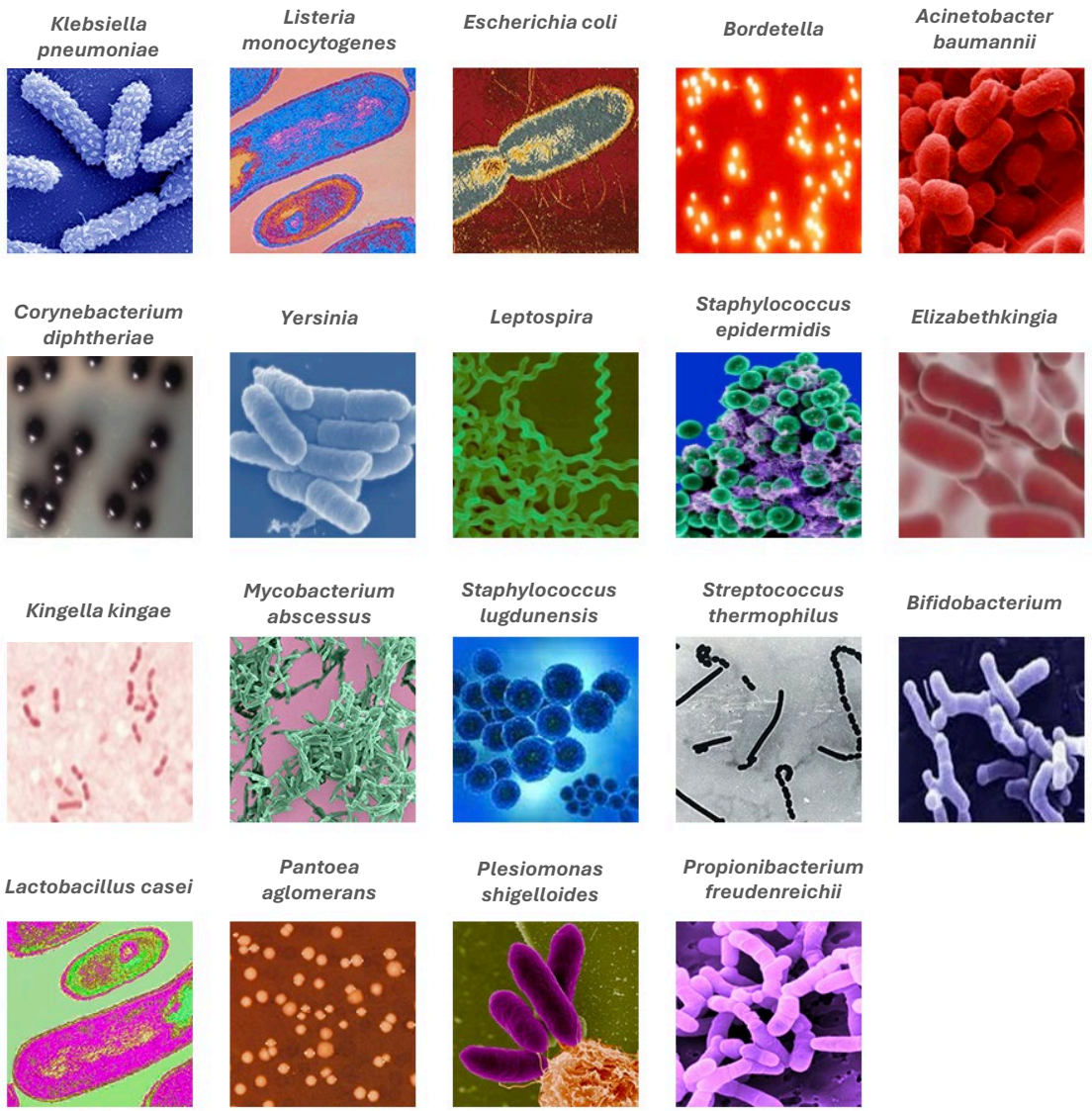
***Moraxela***
**Cepas:2567**
Ensamblados
- Preexistentes:**416**
- De secuenciación de nueva generación:**2151**
- En curso:**24**

Esquemas
- Achtman 7 Gene:**2568**
- rMLST:**2151**

Inicio de la base de datos

# Bases de datos de cepas concretas

- **BIGSdb-Pasteur** (https://bigsdb.pasteur.fr/): Alberga varias bases de datos basadas en la tipificación multilocus de secuencias (MLST), la tipificación basada en el genoma completo y esquemas suplementarios (en particular, genes de resistencia a los antimicrobianos o de virulencia).



# Base de datos del complejo de especies *Klebsiella*

- https://bigsdb.pasteur.fr/klebsiella/klebsiella.html para datos genómicos y definiciones genotípicas de aislados del complejo de especies *K. pneumoniae* (también denominado *K. pneumoniae* en sentido amplio, es decir, que incluye ***K. pneumoniae, K. quasipneumoniae, K. variicola*** y taxones relacionados) a partir de
  - ✓ tipificación multilocus de secuencias (MLST),
  - ✓ MLST de genoma central (cgMLST),
  - ✓ MLST ribosómico (rMLST),
  - ✓ tipificación capsular (secuenciación wzc y wzi) y
  - ✓ MLST de complejos génicos de virulencia

# Base de datos del complejo de especies *Klebsiella*

- La base de datos del complejo de especies *Klebsiella* contiene en realidad dos bases de datos:
  - ✓ Base de datos de secuencias y perfiles
  - ✓ Base de datos de aislados

- Para enviar aislados (con nuevos tipos de secuencia para la depuración y la asignación de tipificaciones de secuencias) o alelos o perfiles de MLST, primero hay que iniciar sesión en las bases de datos respectivas (los gestores de *Klebsiella* proporcionan un nombre de usuario y una contraseña temporales si se contacta con ellos [klebsiellaMLST@pasteur.fr]).

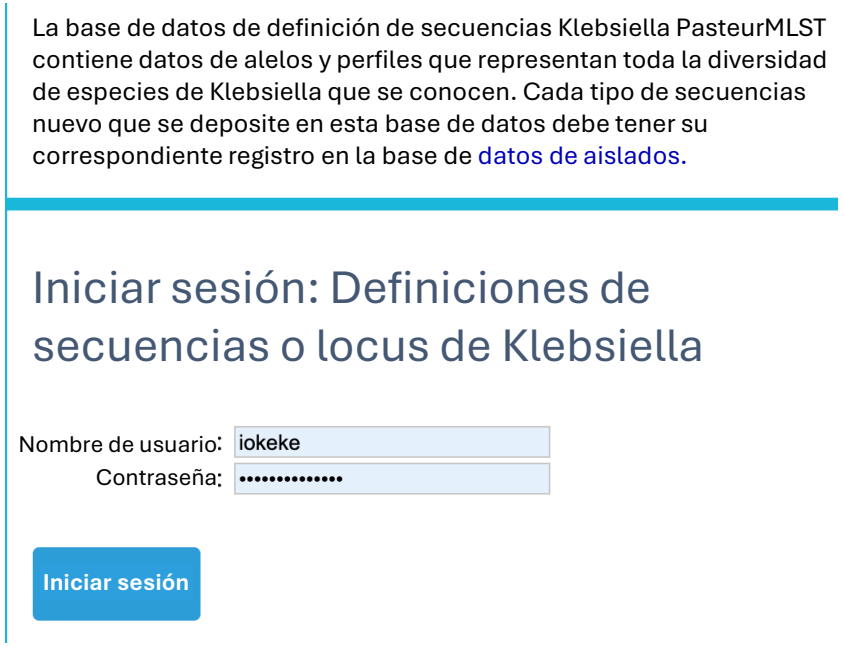
- También se pueden buscar alelos o perfiles de interés en la base de datos del complejo de especies Klebsiella y descargarlos para su posterior análisis

# Base de datos del complejo de especies *Klebsiella*

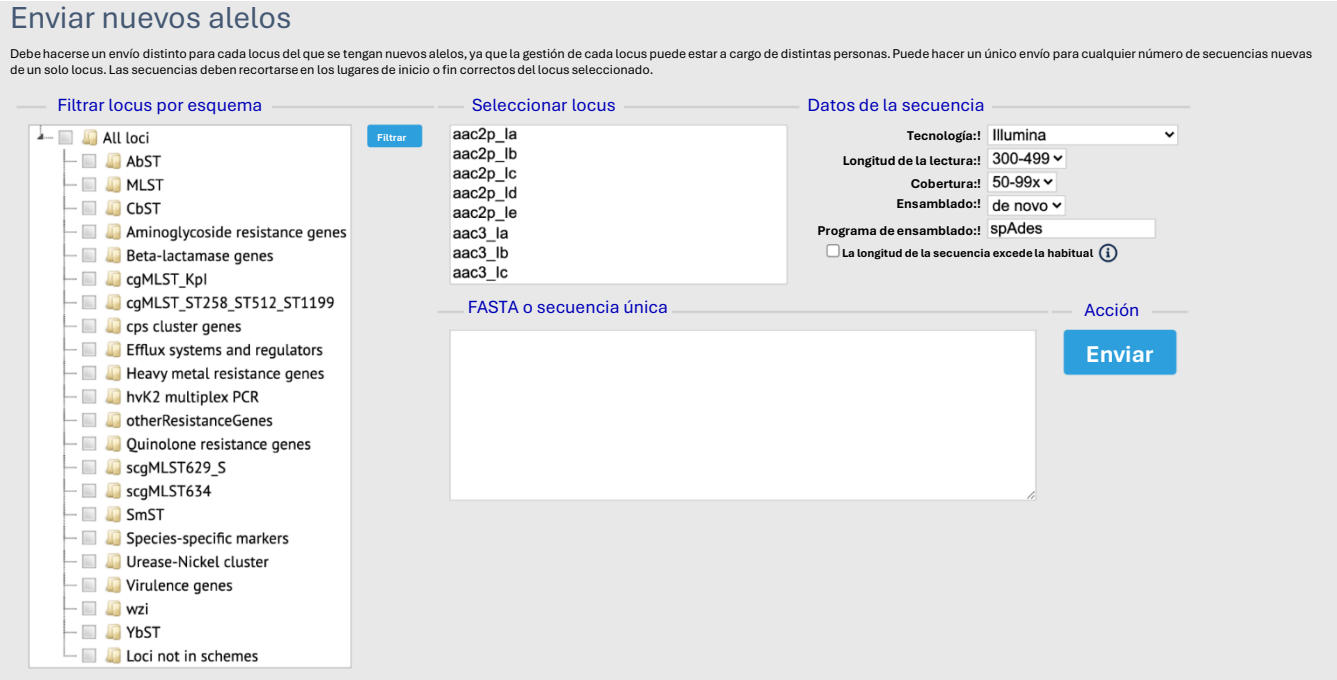
- Inicio de sesión en la base de datos de secuencias y perfiles (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef&page=login)

  → → →

- Iniciar sesión Envío Gestionar envíos Elegir tipo de envío (alelos o perfiles de MLST).



# Base de datos del complejo de especies *Klebsiella*: Tipo de envío (alelo/perfil de MLST) en la base de datos de secuencias y perfiles



# Base de datos del complejo de especies *Klebsiella*

- Inicio de sesión en la base de datos de aislados (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_isolates )

→ → →

- Iniciar sesión → Envío → Gestionar envíos → Elegir tipo de envío
  - ✓ aislados (sin archivos de ensamblado)
  - ✓ genomas (registros de aislados con archivos de ensamblado)

# Base de datos del complejo de especies *Klebsiella*: Tipo de envío (aislados/genomas) dentro de la base de datos de aislados

## Enviar nuevos aislados

Pegue sus aislados para añadirlos a la base de datos utilizando la plantilla que se facilita a continuación.

- Si lo desea, introduzca los alias (nombres alternativos) de sus aislados en forma de lista, separados por punto y coma (;).
- Si lo desea, introduzca las referencias de sus aislados en forma de lista de identificadores de PubMed, separados por punto y coma (;).
- También puede cargar otros campos de alelos junto con los demás datos de aislados. Para ello, solo tiene que crear una nueva columna con el nombre del locus.

Pegue un texto delimitado por tabuladores **(incluya una línea de encabezado de campo)**

Acción

Enviar

## Enviar nuevos aislados con ensamblados de genomas

Pegue sus aislados para añadirlos a la base de datos utilizando la plantilla que se facilita a continuación.

- Si lo desea, introduzca los alias (nombres alternativos) de sus aislados en forma de lista, separados por punto y coma (;).
- Si lo desea, introduzca las referencias de sus aislados en forma de lista de identificadores de PubMed, separados por punto y coma (;).
- También puede cargar otros campos de alelos junto con los demás datos de aislados. Para ello, solo tiene que crear una nueva columna con el nombre del locus.
- Introduzca el nombre del archivo FASTA de cóntigos del ensamblado en el campo assembly_filename y cargue este archivo como datos de apoyo. Los archivos FASTA pueden estar sin comprimir (.fas, .fasta) o comprimidos con gzip/zip (.fas.gz, .fas.zip). La carga tiene un límite de 500 archivos.
- Introduzca el nombre del método de secuencia utilizado en el campo sequence_method (valores permitidos: 454, Illumina, híbrido Illumina + ONT, híbrido Illumina + PacBio, Ion Torrent, Oxford Nanopore, PacBio, Sanger, Solexa, SOLiD, otros)

Pegue un texto delimitado por tabuladores **(incluya una línea de encabezado de campo)**

Acción

Enviar

# *¡Gracias!*