Exploración y análisis del genoma de *Escherichia coli* K12

Nombre del autor: Hely Salgado **Email:** heladia@ccg.unam.mx
Fecha: 13/sep/2024

Resumen (abstract)

Introducción

Escherichia coli K12 es una cepa modelo de bacterias que ha sido ampliamente estudiada debido a su importancia en biología molecular y microbiología. El genoma de *E. coli* K12, publicado en 1997, representa un avance significativo en la comprensión de los mecanismos genéticos de esta bacteria. La anotación detallada de sus genes y otros elementos genéticos a lo largo de los años ha permitido una comprensión más profunda de su biología y fisiología. Este estudio tiene como objetivo explorar y responder varias preguntas clave sobre el genoma de *E. coli* K12, incluyendo su tamaño, el número de cromosomas, los tipos de features, las fuentes de los datos de anotación, y la cantidad y distribución de genes y CDS. Estos análisis proporcionarán una visión integral del genoma de *E. coli* y facilitarán futuras investigaciones en el campo.

Metodología

Para llevar a cabo la exploración del genoma de *Escherichia coli* K12, se utilizaron datos de anotación disponibles públicamente. Los principales pasos metodológicos fueron los siguientes:

1. Software

Todo el análisis se hizo usando comandos unix. La versión del sistema operativo es CentOS Stream version 9.0. Servidor: tepeu.lcg.unam.mx

Para generar el actual reporte se usó stackedit - markdown [ref].

2. Obtención de datos

Los datos del genoma de *E. coli* K12 fueron descargados de la base de datos NCBI, con ID NC*000913.3 (URL: https://www.ncbi.nlm.nih.gov/nuccore/NC*000913.3/). Se utilizaron archivos en formato GFF (General Feature Format) versión xxxx y fastA.

```
|-- data
|   |-- coli_genomic.fna
|   |-- coli.gff
|   `-- coli_protein.fna
```

A continuación se describen los archivos:

Archivo	Descripción	Tipo
coli_genomic.fna	Secuencia de nucleotidos de *E. coli*	Formato FastA
coli.gff.	Anotación del genoma de *E. coli*	Formato gff
coli_protein.faa	Secuencia de aminoacidos de las proteinas de *E. coli*	formato FastA

Formato de los archivos

- `coli_genomic.fna` : formato fastA

```
> NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
```

Formato:

> a. La primera línea es información de la secuencia, iniciando con el identificador del genoma.

> b. Las siguientes líneas es la secuencia de DNA del genoma.

- `coli.gff` : anotación de features en el genoma

El contenido del archivo es

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM584v2
#!genome-build-accession NCBI_Assembly:GCF_000005845.2
##sequence-region NC_000913.3 1 4641652
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145

NC_000913.3     RefSeq  region  1       4641652 .       +       .       ID=NC_000913.3:1.>
NC_000913.3     RefSeq  gene    190     255     .       +       .       ID=gene-b0001;Dbx>
NC_000913.3     RefSeq  CDS     190     255     .       +       0       ID=cds-NP_414542.>
NC_000913.3     RefSeq  gene    337     2799    .       +       .       ID=gene-b0002;Dbx>
NC_000913.3     RefSeq  CDS     337     2799    .       +       0       ID=cds-NP_414543.>
```

Formato:

> a. Es un formato gff tabular, es decir cada dato es separado por tabulador.
>
> b. Cada renglón en el formato gff es una elemento genético anotado en el genoma, que se le denomina `feature` , éstos features pueden ser genes, secuencias de inserción, promotores, sitios de regulación, todo aquello que este codificado en el DNA y ocupe una región en el genoma de E. coli.
>
> c. Los atributos de cada columna par cada elemento genético son
> `1. seqname. Nombre del cromosoma 2. source. Nombre del programa que generó ese elemento 3. feature. Tipo de elemento 4. start. Posici`

Resultados

1. ¿De qué tamaño es el genoma de *Escherichia coli*?

Solución

- En la base de datos `Genome` en NCBI, en el genoma de E. coli, la versión 3 del genoma, se indica el tamaño del genoma que es de 4,641,652 bp.

```
LOCUS       NC_000913            4641652 bp    DNA     circular CON 09-MAR-2022
```

- También, en el archivo `coli.gff` en formato GFF del genoma de *E. coli*, se indica el tamaño del genoma.

```
##sequence-region NC_000913.3 1 4641652
```

2. ¿Cuántos cromosomas tiene *Escherichia coli*?

En la página del genoma de *E. coli*, se indica que es un sólo cromosoma. Lo que vamos a validar, es si el archivo `coli.gff` contiene solo features del cromosoma.

Archivo: `coli.gff`

- El archivo tiene comentarios que no hay que tomar en cuenta.
- El resto de las líneas vienen anotadas las features, la columna 1 viene anotado el cromosoma.

Algoritmo

1. Eliminar las líneas de comentarios. O bien no tomar en cuenta.
2. Tomar solo la columna 1, donde vienen los cromosomas.
3. Compactar/Quitar repeticiones

Solución

```
cut -f1 coli.gff | uniq
```

En resumen, el genoma de *E. coli* K12 tiene un tamaño de aproximadamente 4.6 millones de pares de bases distribuidos en un solo cromosoma circular.

3. ¿Qué tipos de features tiene el genoma de *Escherichia coli*?

Archivo :

-

Algoritmo

1.
2.

Solución

4. ¿Cuántos tipos de features se encuentran anotadas como parte del genoma de *Escherichia coli*?

Archivo:

-

Algoritmo

1.
2.

Solución

5. ¿Cuáles son las fuentes de los datos de anotación?

Archivo:

-

Algoritmo

1.
2.

Solución

6. ¿Cuántos genes y cuántos CDS tiene el genoma de *Escherichia coli*?

Archivo:

-

Algoritmo

1.
2.

Solución

7. ¿Cuántos orígenes de replicación tiene el genoma de *Escherichia coli*?

Archivo:

-

Algoritmo

1.
2.

Solución

8. ¿Cuántos genes hay en cada una de la cadenas del genoma de *Escherichia coli*?

Archivo:

-

Algoritmo

1.
2.

Solución

9. Escribe un archivo ordenado por cadena y región genómica.

Archivo:

-

Algoritmo

1.
2.

Solución

Discusión

Conclusiones

Referencias

1. Moreno, D. y Carrillo J. (2019). *Normas APA 7.^a^ edición. Guía de citación y referenciación* (Universidad Central, ed.). Universidad Central. Consultado el 07 de agosto de 2024. https://bitly.cx/hE17