Flujo de trabajo para el ensamblaje, anotación y comparación de genomas de bacteriófagos

(GUÍA TUTORIAL)

Estudiantes:

Luis Guillermo Gómez Orozco

Johnnatan Ruiz Escobar

Asesor:

Juan Esteban Pérez Jaramillo, Ph, D.

Co-asesor:

Cristian David Grisales, Biólogo

Instituto de Biología

Facultad de Ciencias Exactas y Naturales

Universidad de Antioquia

2021

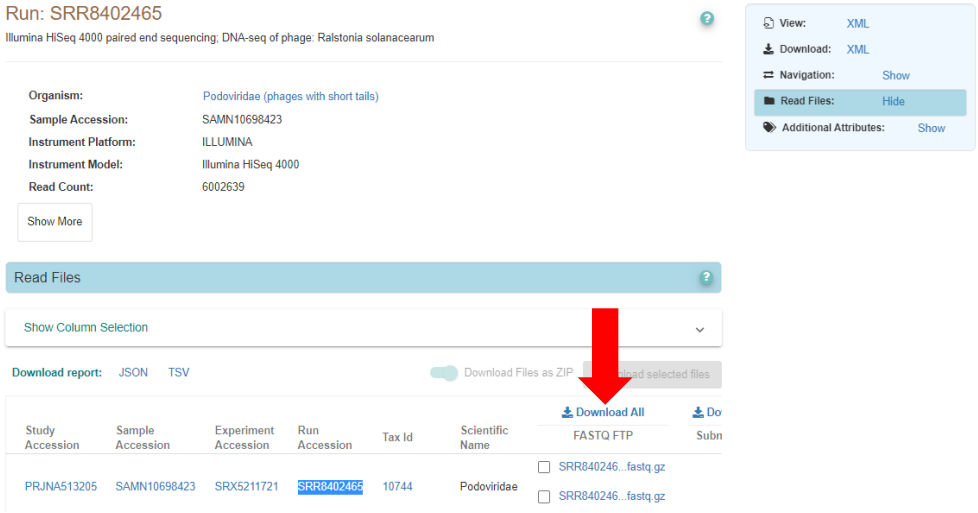
Contenido

Este tutorial muestra de manera sencilla el paso a paso para el tratamiento de genomas en crudo hasta la anotación y la comparación genómica mediante el uso de herramientas de libre acceso. Ninguno de los procesos realizados es absoluto, puesto que los parámetros dentro de cada herramienta pueden variar de acuerdo con la naturaleza de los datos y algunos procesos pueden ser omitidos. Es importante comprender que existen puntos clave como el pre-procesamiento, ensamblaje y anotación de los genomas, siendo estos 3 el eje principal a desarrollar durante el tutorial.

1. Descarga de datos

Lo primero es obtener un conjunto de datos con los cuales podamos trabajar, para ello procederemos a descargar secuencias de libre acceso del Archivo Europeo de Nucleótidos (ENA por sus siglas en inglés). Si usted ya tiene datos de secuenciación por favor omita estos pasos e inicie en el apartado 2 “Plataforma Galaxy y carga de datos”. En este punto del tutorial usaremos lecturas pareadas (paired-end) de la plataforma Illumina HiSeq del bacteriófago de ***Ralstonia solanacearum*** (código de acceso ENA SRR8402465). Esto es debido a una mayor facilidad y rapidez a la hora de ejecutar los programas, puesto que son secuencias cortas y de calidad promedio. Esto último nos permitirá realizar un preprocesamiento que admita ver la diferencia entre las lecturas iniciales y las procesadas.

Para descargar las secuencias, es necesario ir al link https://www.ebi.ac.uk/ena/browser/view/SRR8402465 . Nos desplazamos hacia abajo y en la columna de “FASTQ FTP” damos clic en “Download All”. Para descargar los archivos SRR8402465_1.fastq.gz y SRR8402465_2.fastq.gz.



Es buena práctica crear una nueva carpeta de trabajo, así mismo renombrar los archivos descargados (Ralstonia phage R1 y Ralstonia phage R2 para el forward y el reverse respectivamente). Estos estarán comprimidos como archivos .gz, los cuales podremos montar a la plataforma de Galaxy para dar inicio al preprocesamiento de los datos.

2. Plataforma Galaxy y carga de datos

El preprocesamiento de los datos se realizará en la plataforma de acceso libre Galaxy. Esta es una plataforma web de código abierto para la investigación intensiva en datos.

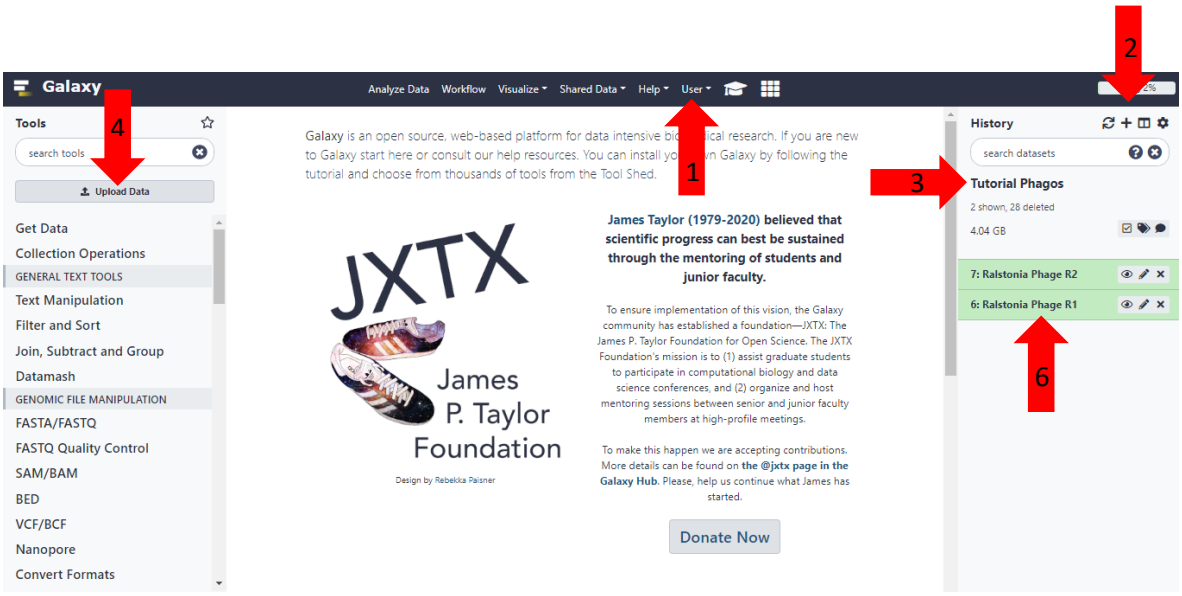
Sitio web: https://usegalaxy.org/

Instrucciones

1. Para poder manipular datos, lo primero será realizar el registro y la respectiva confirmación desde el correo a vincular con la plataforma. Para mayor información sobre el registro puede visitar el siguiente link (https://galaxyproject.org/support/account/). Si es de su interés aprender más sobre usegalaxy, puede visitar el tutorial completo accediendo a la pantalla de aprendizaje en la siguiente dirección (https://galaxyproject.org/learn/).

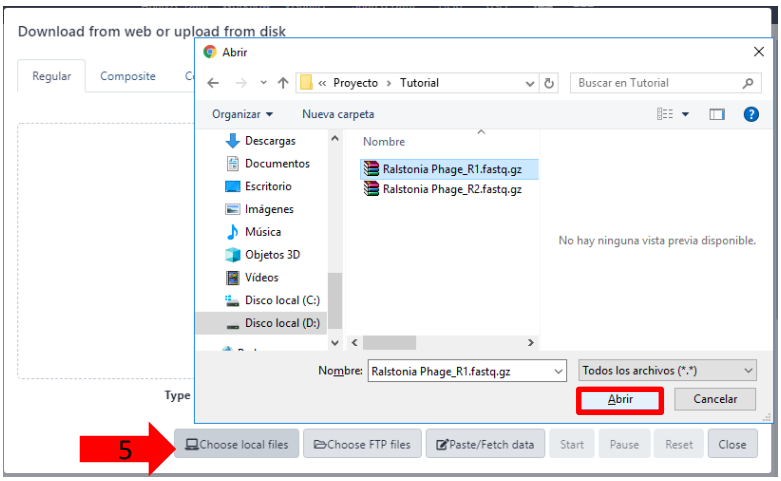
2. Una vez registrados e ingresados en la plataforma se procederá a crear una nueva "*History*" en la cual se ejecutarán todos los procesos de nuestro proyecto. Bastará con dar clic sobre el signo + ubicado en la parte superior derecha de la ventana.

3. Procederemos a renombrar la historia haciendo clic izquierdo sobre "*Unnamed history*". Le daremos un nombre a convenir con nuestro proyecto.



4. La carga de archivos se realizará desde la opción "*Upload Data*" en la parte superior izquierda. Cargaremos los datos que previamente se habían descargado desde la web del ENA.

5. En esta ventana podremos arrastrar los ficheros o cargarlos desde la opción "*Choose local file*". Seleccionamos los archivos desde la carpeta de origen, y procedemos a dar clic en abrir. Dependiendo del tamaño de nuestro archivo. fq o .fastq será el tiempo que se tarde la plataforma en cargar los datos también influye el flujo de trabajo en tiempo real sobre la plataforma y el ancho de banda de su internet. Para mayor información sobre la carga de archivos en Galaxy (https://galaxyproject.org/tutorials/upload/).



6. Sabremos que los archivos han sido cargados correctamente a la plataforma cuando, en la parte derecha podamos visualizarlos en color verde. Ahora estamos

preparados para iniciar el preprocesamiento de los datos crudos de nuestro genoma.

3. Preprocesamiento

El preprocesamiento es el primer acercamiento a la anotación correcta de nuestro genoma. En él vamos a mejorar la calidad de nuestros datos crudos. El control de calidad de los datos crudos sirve como un chequeo rápido para identificar y excluir datos de baja calidad, lo cual permite ahorrar gran cantidad de tiempo en los análisis posteriores, y evitar resultados erróneos. Las herramientas empleadas chequean la calidad de las bases (probabilidad de que la base asignada sea la correcta), la distribución de los nucleótidos, la distribución del contenido de GC, secuencias repetidas, entre otros parámetros (Guo *et al.,* 2014). Antes de intentar reunir un conjunto de datos, es una buena práctica examinar las lecturas para ver si son de buena calidad. Para ello utilizaremos una plataforma amigable y fácil de manejar llamada Galaxy (Afgan *et al.,* 2018). En esta web se aloja gran cantidad de herramientas bioinformáticas que facilitará el análisis preliminar de nuestros datos de secuenciación.

3.1 FastQC

FastQC (Andrews, 2010) tiene como objetivo proporcionar una forma sencilla de realizar algunas comprobaciones de control de calidad en datos provenientes de secuenciación.

La herramienta produce un texto básico y un archivo de salida HTML que contiene todos los resultados, incluidos los siguientes:

- Estadísticas básicas
- Calidad por secuencia de base
- Puntuaciones de calidad por secuencia
- Contenido por secuencia base
- Por contenido de GC base
- Por contenido de N base
- Secuencias sobrerrepresentadas

Sitio web: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ El anterior link es el sitio web del proyecto FastQC. En este puede descargar e instalar el programa, y tener acceso a tutoriales sobre el software.

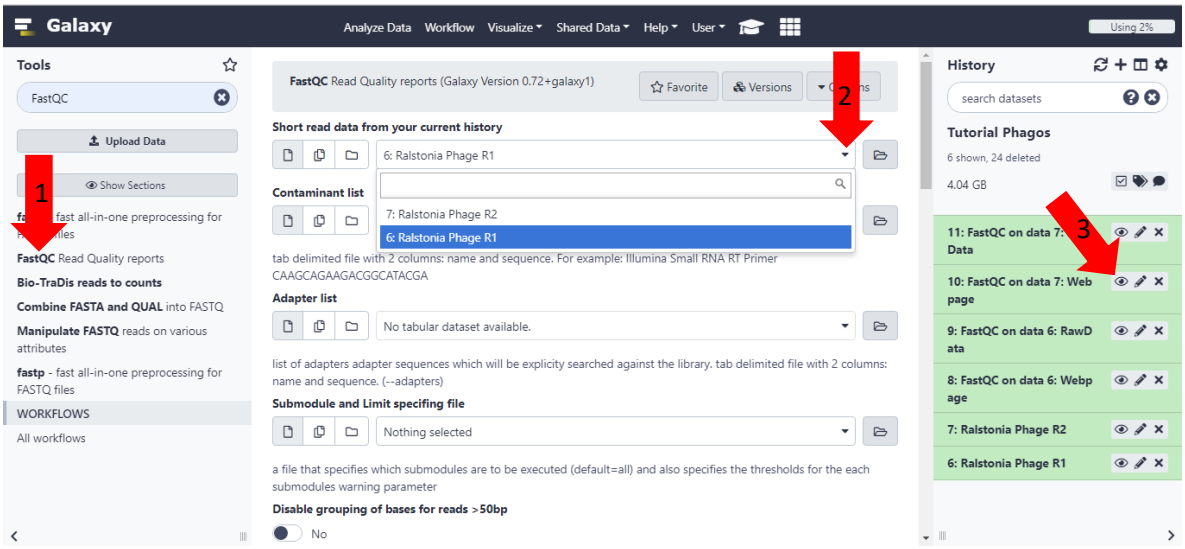
Compatibilidad: Está basado en el lenguaje Java, disponible para Windows, Linux y Mac OS. En este tutorial usaremos la versión 0.11.8 de FastQC vinculada a la plataforma Galaxy.

Entrada: Archivos de lecturas de secuencias forward y reverse (formato. fastq)

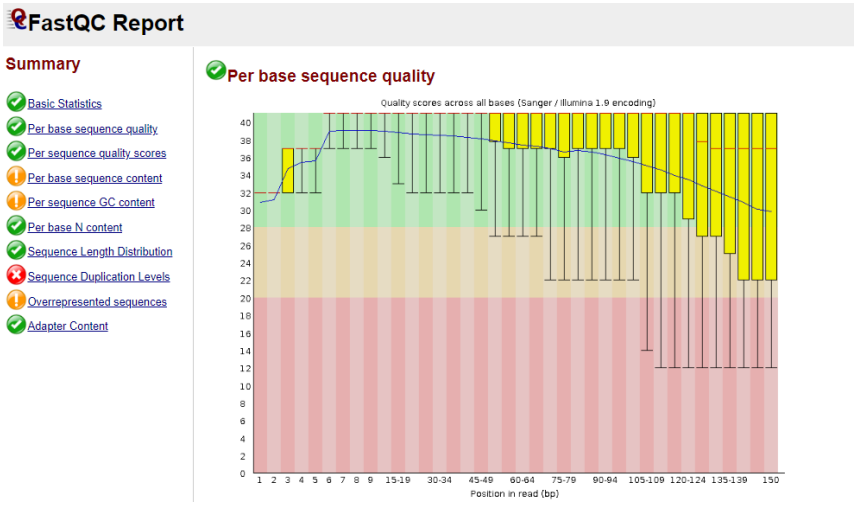
Instrucciones

1. En la parte superior izquierda en el apartado de “Tools” filtraremos por la palabra “FastQC” y daremos clic en la opción “**FastQC** Read Quality reports”.

2. En la opción “**Short read data from your current history**” que nos aparecen en la parte central de la pantalla seleccionamos el primer archivo a procesar. Haremos uso de los valores predeterminados, nos desplazamos hacia abajo y damos clic en el botón “**Execute**”



3. Una vez la plataforma ejecute el FastQC, veremos dos nuevos estados en el apartado de historial en la sección derecha de la plataforma. Para ver el reporte debemos dar clic en el icono con forma de ojo en el FastQC “Webpage”.



Para mayor información sobre el control de calidad y comprender cada uno de los gráficos del reporte de FastQC puede visitar el tutorial disponible en el link: https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

Nota: Deberá ejecutar nuevamente otro FastQC para el archivo faltante. En caso de obtener más de dos datos, con el fin de acelerar el proceso, se puede ejecutar un MultiQC (Elwels *et al.,* 2016)

Es aconsejable ejecutar el FastQC cada vez que los datos sean procesados por alguna otra herramienta, esto con el fin de tener un control sobre los cambios generados en las secuencias.

3.2 Cutadapt

Cutadapt (Martin, 2011) busca y elimina secuencias de adaptadores, cebadores, colas poli-A y otros tipos de secuencias no deseadas de sus lecturas de secuenciación. Todas las lecturas que estaban presentes en el archivo de entrada también estarán presentes en el archivo de salida.

Para archivos producto de secuenciación por Illumina se recomienda usar los valores pre-establecidos.

Sitio web: https://cutadapt.readthedocs.io/en/latest/index.html El anterior link es el sitio web de la herramienta, en este puede descargar he instalar el programa, y tener acceso a tutoriales sobre el software.

Compatibilidad: Se basa en la herramienta Cutadapt de código abierto disponible para Windows, Linux y Max OS X. En este tutorial usaremos la versión disponible para la plataforma Galaxy (Galaxy Versión 1.16.6 con Python 3.6.6).

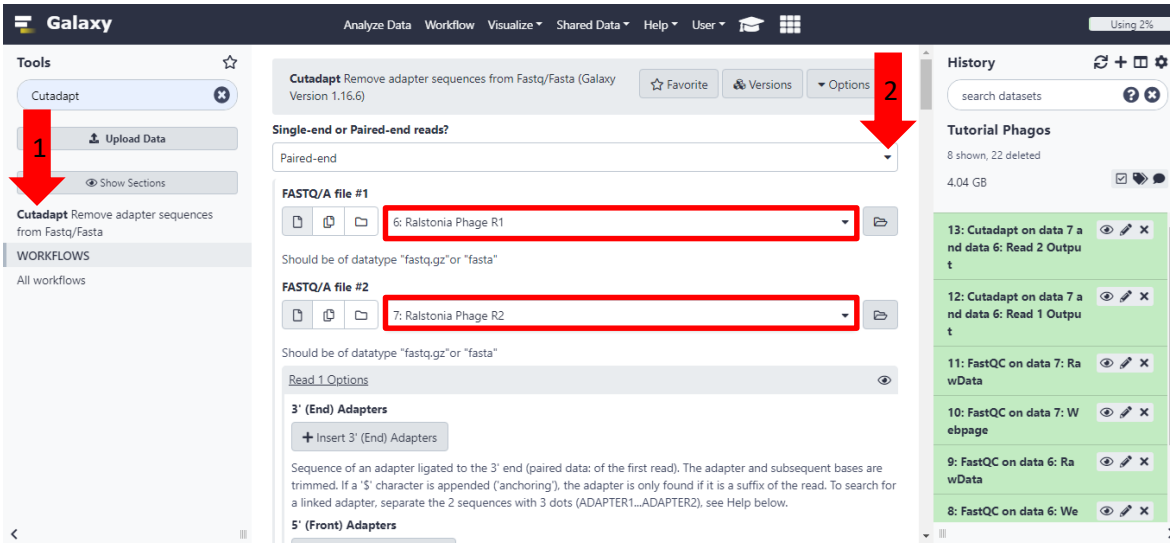
Entrada: Archivos de lecturas de secuencias cortas y comprimidas (formato.fastq.gz, fastq.bz, fastq o. fasta)

Instrucciones

1. Primeramente iremos a la sección de búsqueda y filtraremos por "**Cutadapt**" y seleccionaremos la opción de "**Cutadapt Remove adapter sequences from Fastq/Fasta**".

2. En el panel central daremos clic en la opción de "**Single-end or Paired-end reads**?". Aquí debemos elegir la opción que corresponda con la naturaleza de nuestros datos. En este tutorial estamos utilizando datos pareados (paired-end), por lo que daremos clic en este apartado. Punto seguido seleccionaremos los archivos en File #1 y File #2 correspondientes al fago de Ralstonia R1 y R2 respectivamente.

3. Procederemos a ejecutar la herramienta, bajando y dando clic en "**Execute**", recuerde que si la naturaleza de sus datos son provenientes de Illumina bastará con ejecutar los parámetros pre-establecidos. Si sus datos provienen de otra tecnología de secuenciación diferente a Illumina, puede visitar el tutorial de Cutadapt para validar la configuración que mejor se adapte a sus datos y necesidades en el link: https://cutadapt.readthedocs.io/en/stable/guide.html



Si queremos ver el reporte generado en el procesamiento, podemos ir al apartado "**Output Options**" en el panel central y marcar la opción de "**Report**" como "**Yes**"

3.3 Filter by quality

Filter by quality (Gordon, 2010) es una herramienta que filtra las lecturas según los puntajes de calidad. Se basa en el kit de herramientas FASTX de Assaf Gordon. La distribución del puntaje de calidad (de todos los ciclos) se calcula para cada lectura.

Compatibilidad: Está basado en el kit de herramientas FASTX desarrollada por Assaf Gordon. En este tutorial usaremos la versión 0.0.13 Filter by quality vinculada a la plataforma Galaxy.

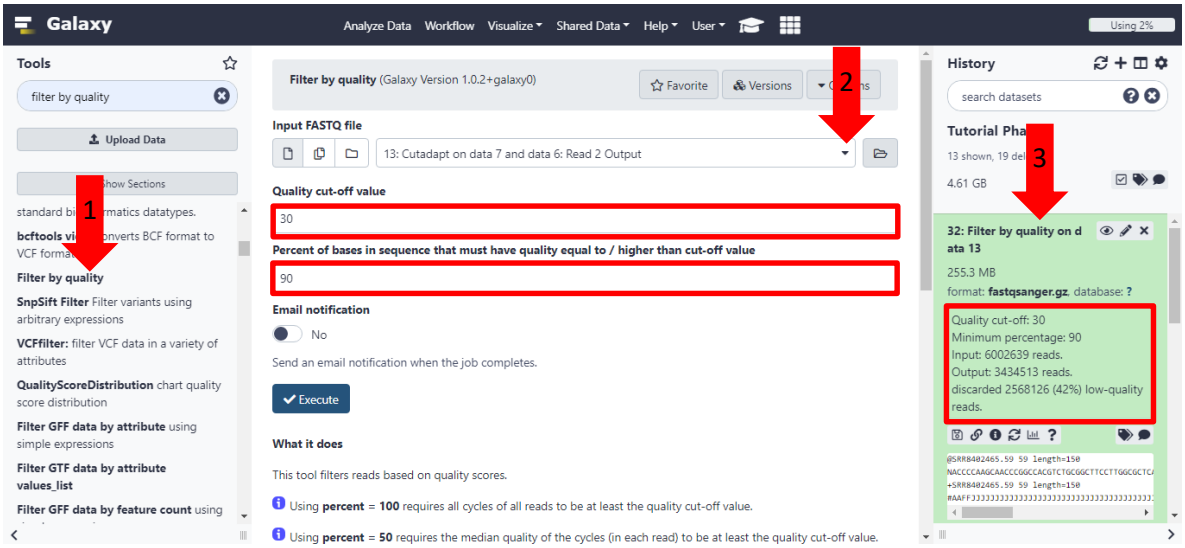
Entrada: Archivos de lecturas de secuencias cortas (formato. fasta y. fastq)

Instrucciones

1. En el buscador filtraremos por "**Filter by quality**" y daremos clic sobre la opción con este nombre.

2. En la opción "**Input FASTQ file**" seleccionamos el primer archivo a procesar, debe ser el archivo correspondiente a la primera lectura procesada previamente por el **Cutadapt**. Cambiaremos los valores del corte de calidad de acuerdo a la calidad de los datos de nuestra secuencia (estos valores se cambiarán dependiendo de la calidad de las lecturas). Recomendamos filtrar por un **Quality cut-off value** de 30 y porcentaje de bases en secuencia de 90%, si el valor Phred de sus lecturas es alto. Ver el apartado de **Per base sequence quality** en **FastQC** para verificar el valor Phred). Nos desplazaremos hacia abajo y daremos clic en el botón "Execute"

3. Para ver los resultados de la ejecución bastará con dar clic sobre el proceso en el historial. Se aprecia el porcentaje de lecturas que fueron descartadas durante el filtro de calidad.



Recuerde ejecutar nuevamente el **Filter by quality** para el archivo faltante Ralstonia Phage R2 previamente procesado con **Cutadapt.**

3.4 Trimmomatic

Trimmomatic (Bolger *et al.,* 2014) es un software rápido y multiproceso que se puede utilizar para remover secuencias de baja calidad, recortar ambos extremos de las secuencias y además elimina adaptadores cuando están presentes.

Los pasos de recorte actuales son:

- ILLUMINACLIP: Corte el adaptador y otras secuencias específicas de illumina de la lectura.
- SLIDING WINDOWN: Realice un recorte de ventana deslizante, cortando una vez que la calidad promedio dentro de la ventana cae por debajo de un umbral.
- LÍDER: corte las bases al comienzo de una lectura, si está por debajo de un umbral de calidad
- TRAILING: corte las bases al final de una lectura, si está por debajo de un umbral de calidad
- CROP: Corta la lectura a una longitud especificada
- HEADCROP: Corta el número especificado de bases desde el inicio de la lectura
- MINLEN: Elimine la lectura si está por debajo de una longitud especificada
- TOPHRED33: Convierta puntajes de calidad a Phred-33
- TOPHRED64: convierte puntajes de calidad a Phred-64

Sitio web: http://www.usadellab.org/cms/?page=trimmomatic El anterior link es el sitio web de la herramienta, en este puede descargar he instalar el programa, y tener acceso a tutoriales sobre el software.

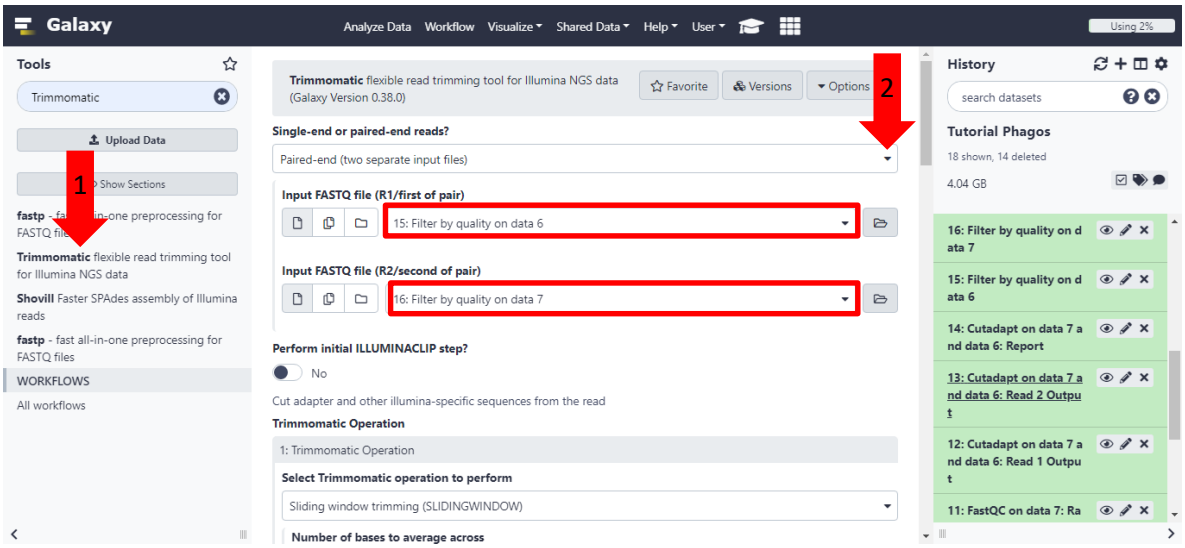
Compatibilidad: Está basado en el lenguaje Java, disponible para Windows, Linux y Mac OS. En este tutorial usaremos la versión 0.38.0 vinculada a la plataforma Galaxy.

Entrada: Funciona con FASTQ (utilizando puntuaciones de calidad phred + 33 o phred + 64, según la canalización de Illumina utilizada), ya sea FASTQ sin comprimir o con gzip. El uso del formato gzip se determina en función de la extensión .gz.

Instrucciones

1. En el buscador filtraremos por “**Trimmomatic**” y daremos clic sobre la opción “**Trimmomatic flexible read trimming tool for Illumina NGS data**”.

2. En el panel central daremos clic en la opción de “**Single-end or Paired-end reads**?”. Debemos elegir la opción que corresponda con la naturaleza de nuestros datos (En nuestro caso son datos separados entre archivos. Si sus datos son R1 y R2 pero están en un solo archivo elija la opción **Paired-end as collection**). Luego seleccionaremos los archivos en **Input FASTQ file (R1/first of pair)** y **(R2/second of pair)** correspondientes a los datos previamente procesados por **Filter by quality.**



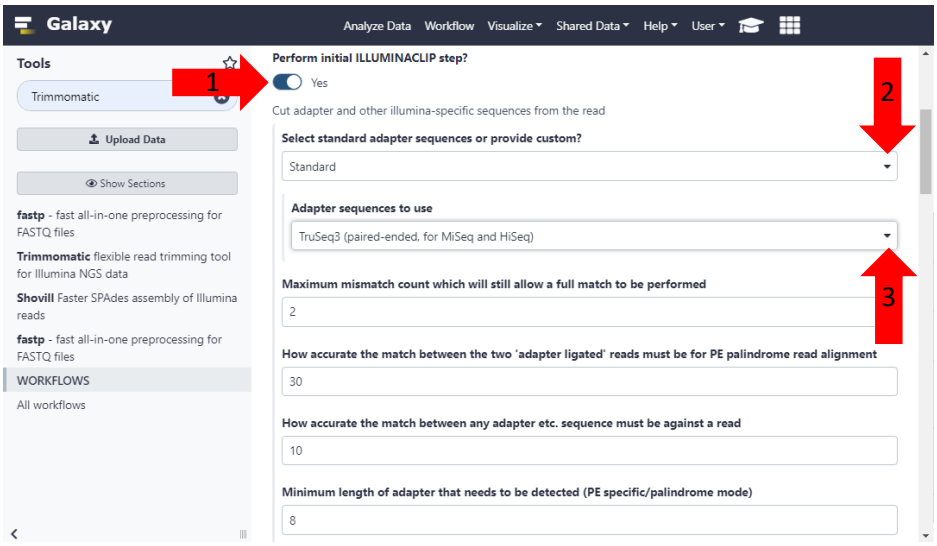
En este tutorial usaremos 4 procesos permitidos a ejecutar por el Trimmomatic.

3.4.1 Illuminaclip Adaptador de corte y otras secuencias específicas de lecturas Illumina.

1. Activaremos el “**Perform initial ILLUMINACLIP step?**” Esto desplegará un listado de opciones nuevas.

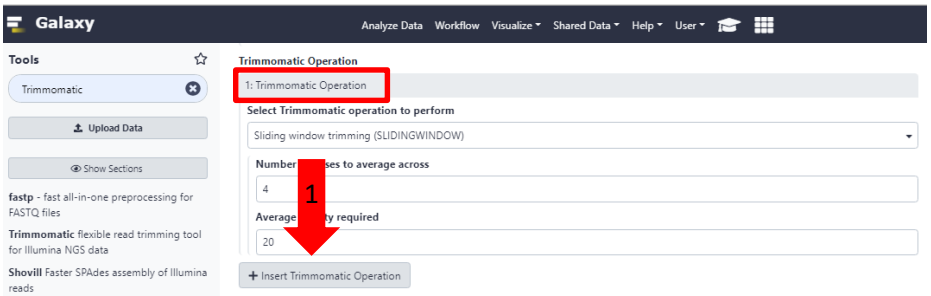
2. Elegiremos la opción “**Standard**”**.** Si conoce los adaptadores con los cuales se corrieron sus secuencias, elija la opción “**Custom**” y en el cuadro desplegado proporcione la secuencia del adaptador.

3. En la opción de “**Adapter sequences to use**” elegimos la naturaleza (Paired-end o singled-end) y la tecnología de secuenciación de nuestros datos. Las otras opciones se dejan en valores estándar.



3.4.2 Headcrop, Triling & Slidingwindow La primera herramienta corta el número especificado de bases desde el inicio de la lectura. La segunda corta las bases al final de una lectura, si está por debajo de un umbral de calidad. La última realiza un recorte de manera continua, cortando una vez que la calidad promedio dentro de la ventana cae por debajo de un umbral.

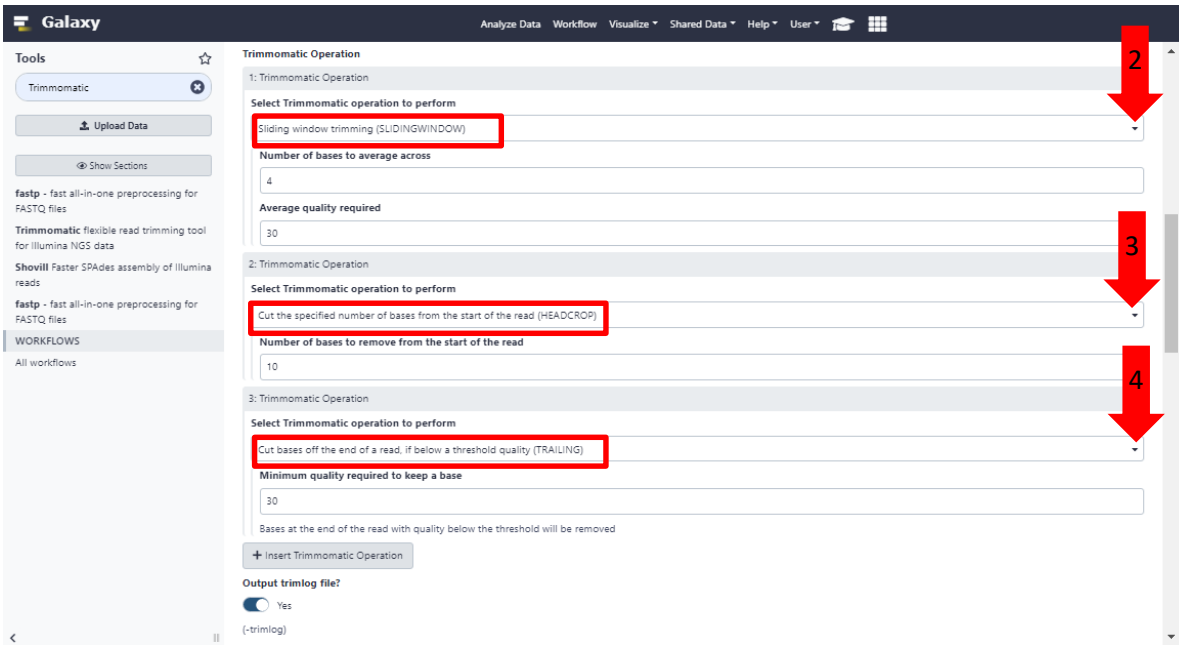
1. El primer paso a realizar será insertar dos operaciones adicionales a realizar por trimmomatic haciendo clic en el botón de "**+ Insert Trimmomatic Operation**".



2. En el primer apartado elegiremos la operación (SLIDINGWINDOW) y en las opciones sólo cambiaremos el "**Average quality required**" por un valor acorde con el valor phred de nuestra lectura, valor recomendado estándar entre 20 y 30.

3. En el segundo apartado elegiremos la operación (HEADCROP) y en su respectiva opción elegiremos el valor a convenir según la calidad de nuestros datos (apartado "**Per base sequence content**" del FastQC), valor recomendado estándar es 10.

4. En la última operación elegimos el proceso (TRAILING) y en su opción pondremos el valor Phred que creamos conveniente según la calidad de los datos de secuenciación, valor recomendado estándar entre 20 y 30.



Una vez realizadas las acciones y configuraciones pertinentes nos desplazaremos hacia abajo y daremos clic en el botón de "**Execute**"

3.5 Bowtie2

Bowtie2 (Langmead & Salzberg, 2012) es una herramienta rápida y de memoria eficiente para alinear las lecturas de secuenciación con secuencias de referencia largas. La plataforma de Galaxy para Bowtie2 permite seleccionar entre índices precalculados y definidos por el usuario para genomas de referencia. Permite usar un índice de genoma incorporados o cargar un genoma de interés como un archivo FASTA en el historial de Galaxy y seleccionarlo para mapear a genomas de referencia. Usaremos esta herramienta para descontaminar los reads de los fagos de *R. solanacearum.*

Sitio web: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Compatibilidad: A partir de la versión 2.3.5, Bowtie2 ahora admite la alineación de lecturas SRA. Las compilaciones pre-empaquetadas incluirán un paquete compatible con SRA. Si está compilando Bowtie2 desde la fuente, asegúrese de que Java esté disponible en su sistema.

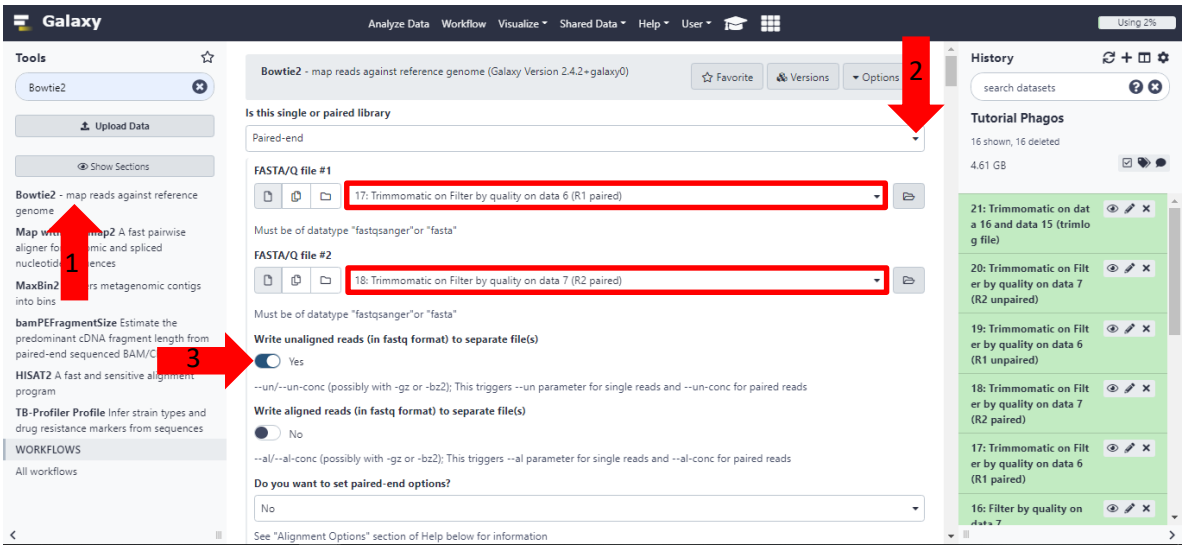
Entrada: Bowtie2 acepta archivos en formato Sanger FASTQ (single o paired-end).

Instrucciones

1. En el buscador filtraremos por “**Bowtie2**” y daremos clic sobre la opción “**Bowtie2** – map reads against reference genome”.

2. En el panel central elegimos la opción acorde a la naturaleza de nuestros datos (Paired-end según el tutorial) y en los apartados de **FASTA/Q file #1** y **file #2** seleccionamos los datos “Trimmados” del R1 y R2 respectivamente.

3. Seleccionaremos la opción “**Write unaligned reads**” para tener acceso a las lecturas procesadas que no se alinearán con el genoma de referencia.

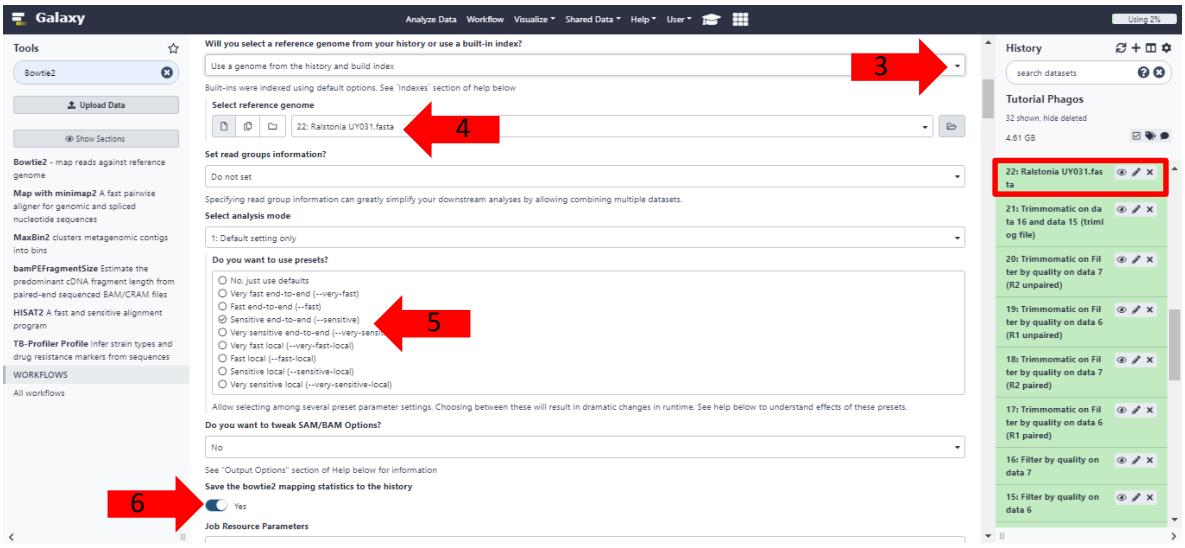


3. Nos desplazamos hacia abajo, y en el apartado de “**Will you select a reference genome from your history or use a built-in index?**” seleccionamos la opción de “**Use a genome from the history**” (si el hospedero de nuestro bacteriófago no se encuentra en el listado predeterminado, deberemos subir la secuencia del genoma de nuestro hospedero en formato .fasta como se enseña en el apartado **2. Plataforma galaxy y carga de datos**). En este tutorial estamos trabajando con un fago de *Ralstonia solanacearum*, por lo que deberemos cargar el genoma del hospedero como referencia. Lo podremos descargar desde la web del NCBI https://www.ncbi.nlm.nih.gov/.

4. En la opción “**Select reference genome**” seleccionamos el genoma de referencia previamente cargado a la plataforma

5. Según como deseemos que sea de riguroso el análisis debemos elegir la opción a convenir.

6. Daremos clic en la opción de “**Save the bowtie2 mapping statistics to the history**”

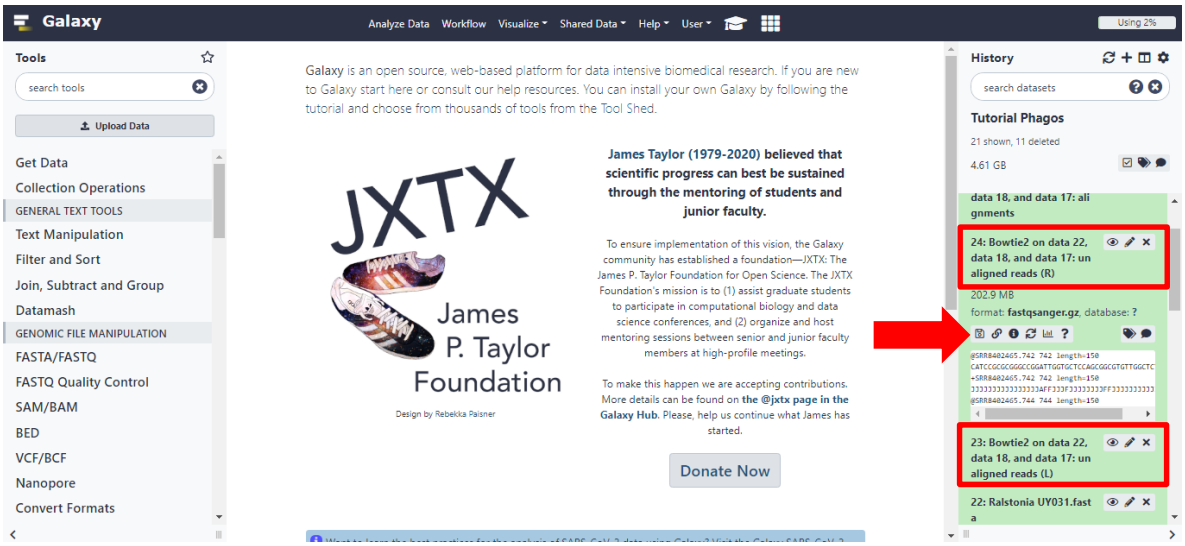


Los demás parámetros se dejarán tal como están, nos desplazaremos hacia abajo y daremos a "**Execute**". Para mayor información puede leer la guía de Bowtie2 Galaxy en el link https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html.

Sería interesante correr otro Bowtie2 tomando el genoma del Homo sapiens b38: hg38 canonical como genoma de referencia, debido a que es posible hallar contaminantes humanos en las muestras por manipulación de aislados.

3.6 Descarga de datos preprocesados

Una vez realizado el preprocesamiento de nuestros datos, procederemos a descargar las secuencias. Para ellos deberemos dar clic sobre el proyecto que deseemos descargar. A continuación, en la información que se despliega daremos clic en el icono del disco "**Download**". Esto iniciará de manera automática la descarga de los datos procesados por esta herramienta.



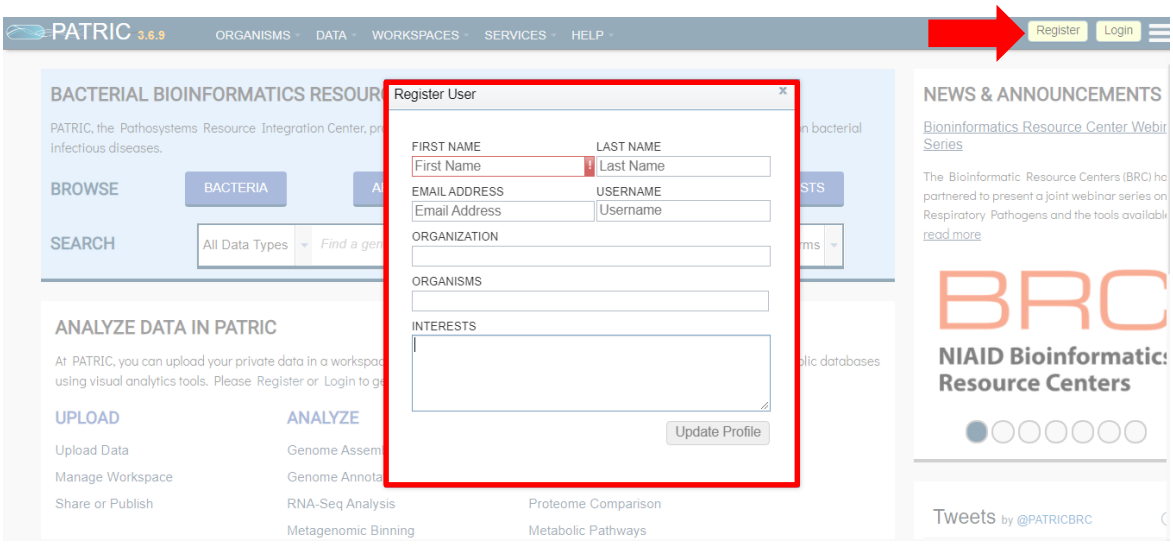
Recuerde renombrar y ubicar el archivo en la carpeta de trabajo para facilitar el proceso, así mismo el descargar los dos archivos (R1 **unaligned read R**) y (R2 **unaligned read L**).

4. Plataforma PATRIC, registro y carga de datos

Centro de Integración de Recursos de Patosystemas (PATRIC por sus siglas en ingles), proporciona datos integrados y herramientas de análisis para respaldar la investigación biomédica sobre enfermedades infecciosas bacterianas. Aquí puede cargar sus datos privados en un espacio de trabajo, analizarlos utilizando servicios de alto rendimiento y compararlos con otras bases de datos públicas utilizando herramientas de análisis visual.

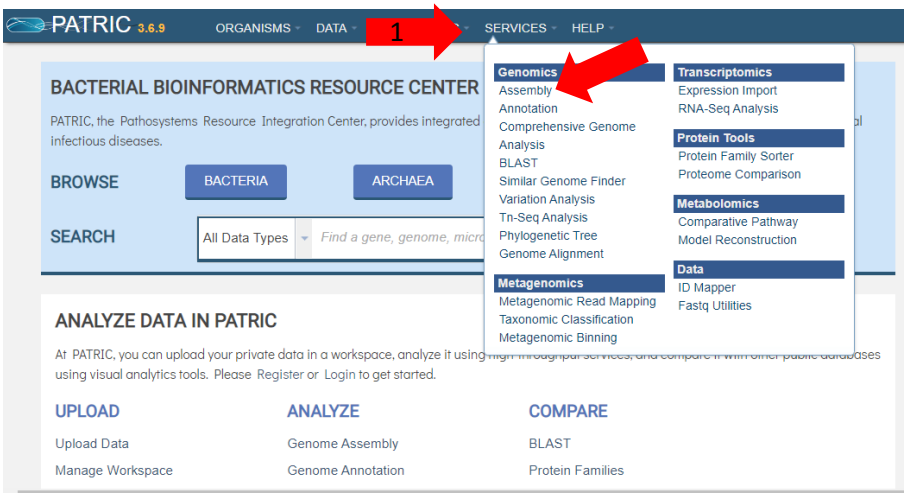
4.1 Registro

En este tutorial, realizaremos el ensamblaje de nuestro genoma mediante la plataforma PATRIC, a la que podremos acceder desde el link https://patricbrc.org/. Lo primero a realizar será el registro desde el botón "**Register**" en la parte superior derecha de la ventana. Posteriormente, en la nueva pestaña se ingresarán los datos personales, y como último paso se efectuará la confirmación desde el correo que elegimos vincular a la plataforma.



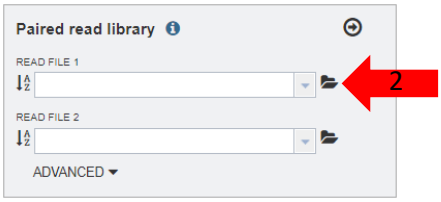
4.2 Carga de datos

1. En la parte superior de la página de PATRIC, busque la pestaña "**Services**" y haga clic en ella, luego en el cuadro desplegable haga clic en "**Assembly**"



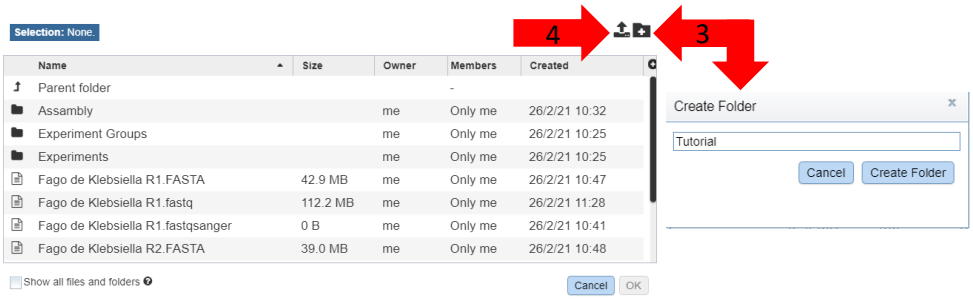
Esto abrirá la página de inicio del ensamblaje, donde podremos subir nuestros datos pareados o no pareados (paired-end o singled-end) dependiendo de la naturaleza de nuestros datos.

2. Como nuestros datos son pareados nos dirigiremos al apartado de "**Paired read library**" y daremos clic en el icono sombreado con forma de carpeta para subir el primer archivo.



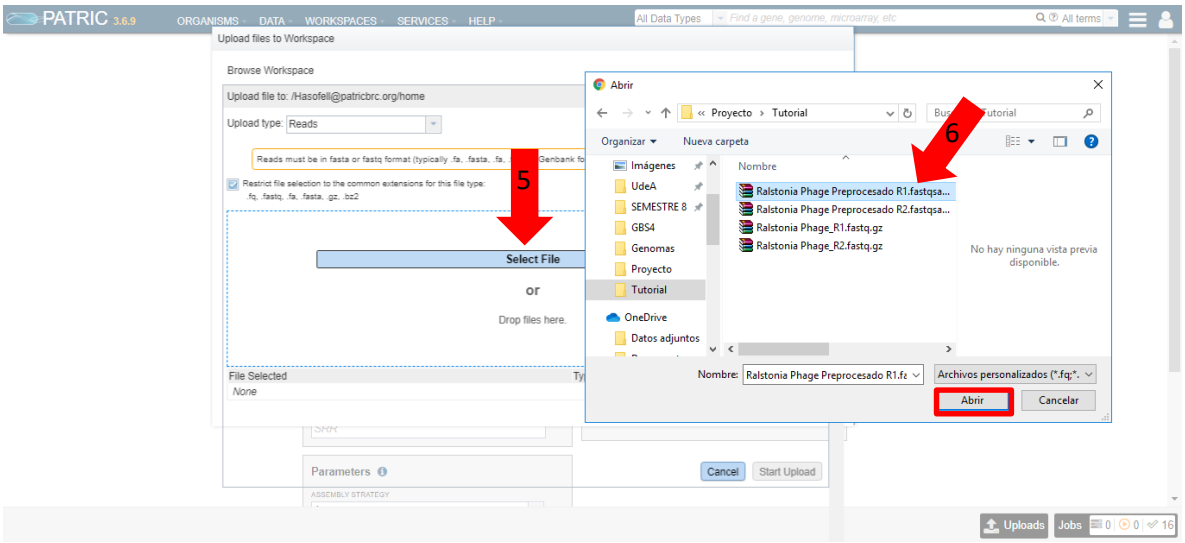
3. En la nueva ventana, lo primero será crear una carpeta en el sistema, en la que podremos guardar nuestros procesos. Para ello daremos clic en el icono con forma de carpeta y, en la ventana emergente podremos nombrar la carpeta.

4. Ahora cargaremos la secuencia, previamente preprocesadas y descargada desde Galaxy, correspondientes a nuestro archivo (R1). Daremos clic en el icono con flecha apuntando hacia arriba.

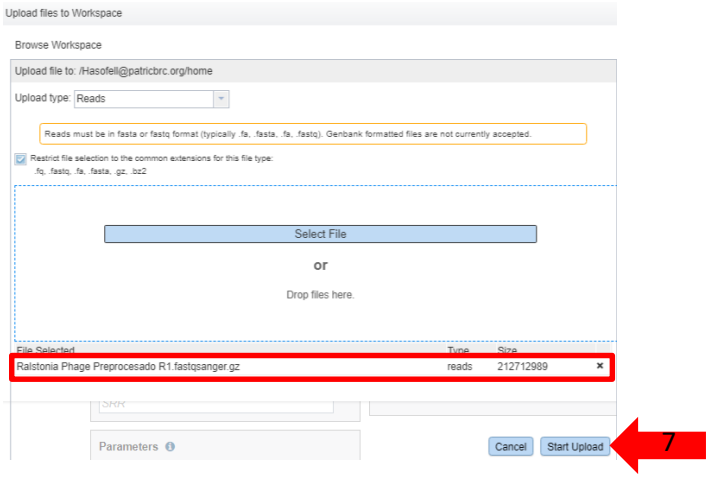


5. Esto abrirá una nueva ventana que nos permitirá seleccionar los archivos guardados en la computadora, para ello daremos clic en "**Select File**".

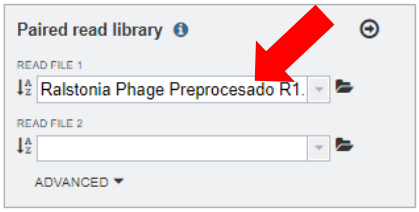
6. En la nueva ventana buscaremos el lugar donde guardamos nuestros datos Preprocesados, seleccionamos y damos clic en "**Abrir**"



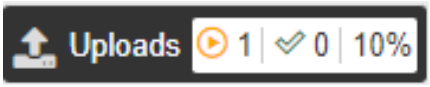
7. Una vez seleccionado, completará automáticamente el nombre del archivo. Punto seguido dar clic en el botón “**Start Upload**” para dar inicio a la carga de datos.



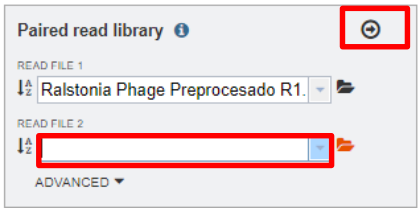
Esto completará automáticamente el nombre del documento en el cuadro de texto.



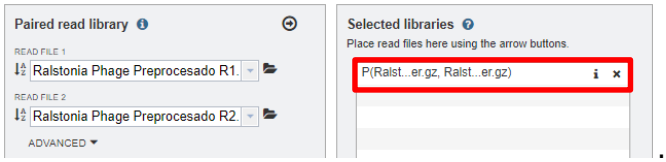
Puede ver el proceso de carga del archivo en la parte inferior derecha de la página de PATRIC. Esto puede tardar varios minutos, dependiendo del tamaño de los archivos.



8. Repita el proceso para cargar el segundo par de la lectura (R2).



9. Una vez los datos se hayan subido a la plataforma al 100% haga clic en el icono de una flecha dentro de un círculo en la parte superior derecha de la ventana, esto finalizará la carga de datos y moverá los archivos al cuadro “**Selected libraries**”.



Ahora podremos trabajar con estos datos.

5. Ensamblaje

El ensamblaje *de novo* es el proceso de fusionar lecturas de secuencias superpuestas en secuencias contiguas (contigs) sin el uso de ningún genoma de referencia como guía. Los ensambladores más eficientes para secuencias de lectura corta suelen ser aquellos que emplean gráficos de *Bruijn* para producir un ensamblaje (Compeau *et al.,* 2011).

5.1 Ensamblaje del genoma (SPAdes)

SPAdes (Bankevich *et al.,* 2012) es un algoritmo de ensamblaje de genoma que fue diseñado para conjuntos de datos bacterianos unicelulares y multicelulares. SPAdes utiliza *k-mers* para construir el gráfico de *Bruijn* inicial y en las siguientes etapas realiza operaciones teóricas de gráficos que se basan en la estructura del gráfico, la cobertura y las longitudes de secuencia. Además, ajusta los errores de forma iterativa. SPAdes es una herramienta de código libre, alojado en el Centro de Biotecnología Algorítmica (CAB por sus siglas en ingles), web en la que podrás hallar versiones de descarga y guías completas.

En este tutorial haremos uso de la herramienta SPAdes para ensamblar *de novo* nuestro genoma de interés. Rihtman y colaboradores (2016). Demostraron que Spades es significativamente mejor que los ensambladores Velvet y Ray a la hora de ensamblar genomas de bacteriófagos, con un desempeño del 98,6% contra 84,6% y 89,7% respectivamente. Para ello nos valdremos de la plataforma de libre acceso PATRIC, en la que previamente realizamos el registro y la carga de datos.

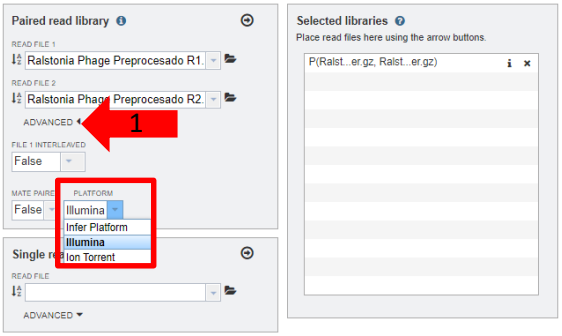
Sitio Web: https://cab.spbu.ru/software/spades/

Compatibilidad: SPAdes es compatible con sistemas operativos como Linux, Windows y Mac. funciona con tecnologías de secuenciación como lo son Ion Torrent, PacBio, Oxford Nanopore e Illumina.

Entrada: Admite secuencias en formato Fastq

Instrucciones

1. Lo primero será elegir la plataforma bajo la que se secuenciaron nuestros datos. Para ello daremos clic en la opción "**ADVANCED**" en las opciones que se despliegan daremos clic "**PLATAFORM**" y seleccionamos la opción pertinente.

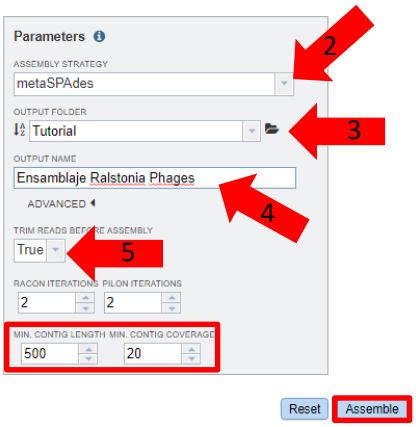


2. En el apartado de "**Parameters**" seleccionaremos el software con el cual queremos realizar el ensamblaje del genoma, haciendo clic en la opción "**ASSEMBLY STRATEGY**". Esta plataforma permite el uso de diversas herramientas, nosotros usaremos SPAdes.

3. Seleccionaremos la carpeta en la cual se van a alojar los diferentes archivos de salida generados por la anotación del genoma, dando clic en "**OUTPUT FOLDER**". Esta carpeta fue creada previamente en el punto 3 del *apartado 4.2 Carga de datos.*

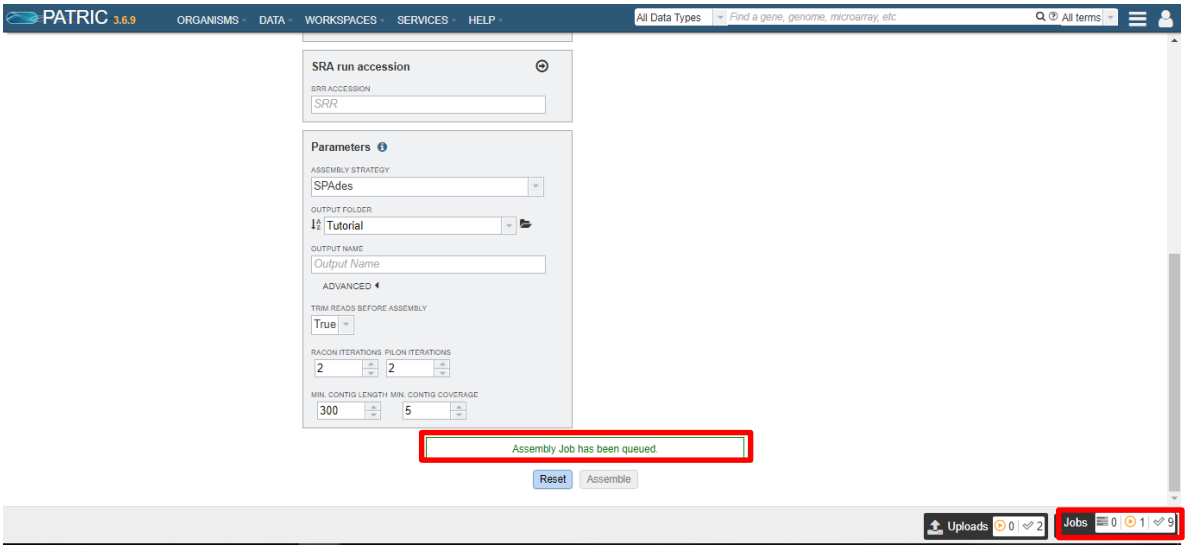
4. En la opción "**OUTPUT NAME**" daremos nombre a nuestro proceso de ensamblaje.

5. En la opción "**TRIM READS BEFORE ASSEMBLY**" cambiaremos a "**True**". Los valores de "**MIN. CONTIG LENGTH"** y "**MIN. CONTIG COVERAGE**" los podremos cambiar de acuerdo a lo estrictos que queramos ser a la hora de ensamblar nuestro genoma. Con nuestros datos, recomendamos usar un valor de 500 y 20 respectivamente como se muestra en la imagen de abajo.



Finalmente daremos clic en el botón de "**Assemble**"

Si desea tener mayor información sobre los parámetros avanzados y las demás herramientas de ensamblaje en la plataforma PATRIC, puede visitar los tutoriales en el link https://docs.patricbrc.org/tutorial/genome_assembly/assembly2.html



Este mensaje indica que nuestro proyecto será subido a la cola de procesos por ejecutar. Será agregado a la lista de "**Jobs**" (trabajos) en ejecución, pestaña en la que si damos clic podremos ver los trabajos ejecutados y en ejecución. Sólo resta esperar la total ejecución.

Status	ID	Service
running	3419286	GenomeAssembly2
completed	3408084	GenomeAssembly2
failed	3408056	GenomeAssembly2
completed	3408049	GenomeAssembly2
failed	3408046	GenomeAssembly2
failed	3408044	GenomeAssembly2
completed	3407768	GenomeAssembly2
completed	3407041	GenomeAssembly2
failed	3407038	GenomeAssembly2
completed	3406156	GenomeAssembly2
completed	3406108	GenomeAssembly2
completed	3406105	GenomeAssembly2
completed	2670937	Annotation
completed	2670886	Annotation
completed	2670044	Annotation
completed	2323619	Annotation

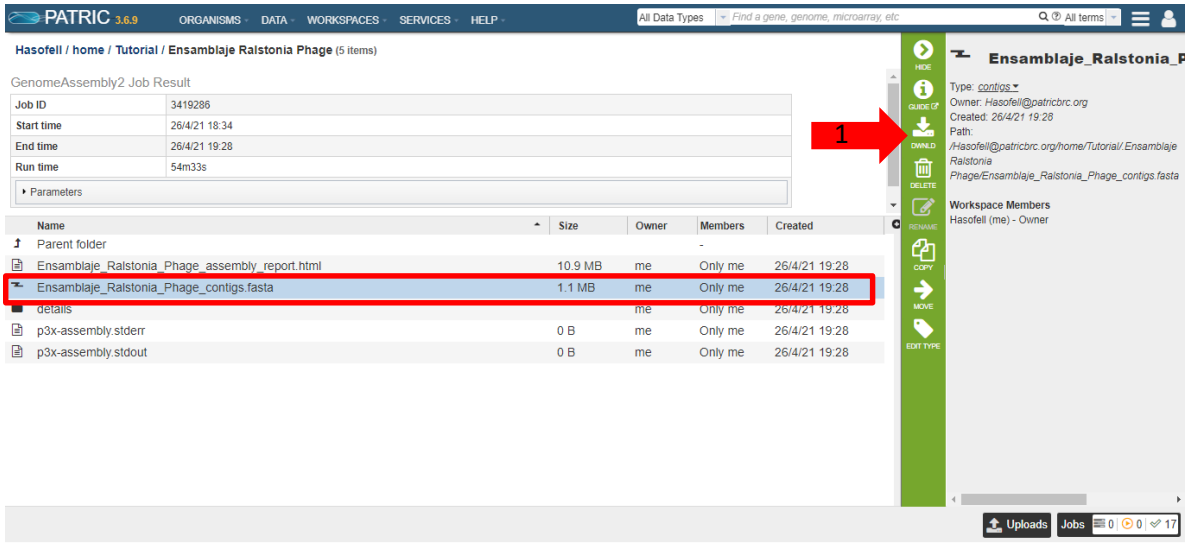
1 - 21 de 21 resultados

5.1.1 Descarga datos ensamblados

Una vez el ensamblaje del genoma haya terminado, el "**Status**" cambiará a "**completed**". Para ver los resultados daremos doble clic sobre el proceso, en esta ventana, podremos ver dos archivo, uno contiene todos los reportes del ensamblaje en formato .html, y otro archivo contig.fasta con las secuencias nucleotídicas del ensamblaje, este último es el archivo de interés.

1. Seleccionamos el archivo contig.fasta y daremos clic en el botón de "**DWNLD**" para iniciar la descarga.

Recuerde renombrar el archivo y guardarlo en la carpeta de trabajo.



5.2 Análisis calidad del ensamblaje

Existen algunas métricas que permiten evaluar la calidad del ensamblaje cuantitativamente. Se calcula generalmente la talla mínima, máxima y media de los contigs, así como la talla total del ensamblaje, la cual debe coincidir con la talla esperada del genoma. Pero el principal valor estadístico es el valor N50 y es definido como la longitud de los contigs, tal que usando contigs de igual o mayor tamaño al valor de N50, estos corresponden a la mitad de las bases del genoma ensamblado. Para una mejor valoración de la calidad, desde el punto de vista cualitativo, se puede realizar el alineamiento de las lecturas con los contigs obtenidos, proceso denominado remapeo (Nagarajan & Pop, 2013).

El análisis de calidad del ensamblaje será realizado en la plataforma de libre acceso Usegalaxy.

5.2.1 Quast

QUAST (Mikheenko *et al.,* 2018) es una herramienta de evaluación de la calidad. La herramienta evalúa los ensamblajes del genoma calculando varias métricas. Funciona con o sin genoma de referencia.

Sitio web: http://quast.sourceforge.net/metaquast

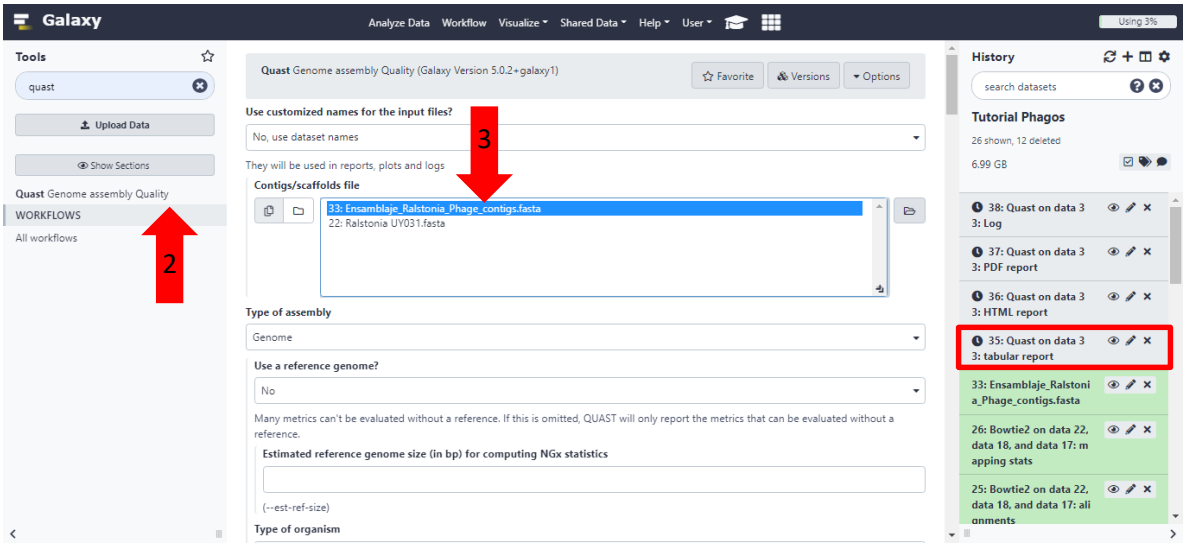
Entrada: La herramienta acepta ensamblajes y genomas de referencia en formato FASTA. Los archivos pueden estar comprimidos con zip, gzip o bzip2.

Instrucciones

1. Cargar los datos del contig.fasta en la plataforma Galaxy como se enseña en el apartado *2. Plataforma Usegalaxy y carga de datos.* Podremos seguir trabajando en el mismo Historial o bien crear uno nuevo.

2. Una vez cargados los datos a la plataforma Galaxy, en el apartado de búsqueda, filtraremos por "**Quast**" y daremos clic en la opción "**Quast Genome assembly Quality**".

3. En el panel central elegimos el contig.fasta, dejamos todos los parámetros en predeterminados y tras desplazar hacia abajo daremos clic en el botón de "**Execute**".



Para visualizar los resultados del proceso, una vez haya terminado, deberemos dar clic en el icono con forma de ojo en el "**tabular report"**.

Assembly	Ensamblaje_Ralstonia_Phage_contigs_fasta
# contigs (>= 0 bp)	22
# contigs (>= 1000 bp)	9
Total length (>= 0 bp)	65922
Total length (>= 1000 bp)	63900
# contigs	10
Largest contig	38385
Total length	64573
GC (%)	61.78
N50	38385
N75	3228
L50	1
L75	4
# N's per 100 kbp	0.00

Si desea realizar un proceso con mayor robustez, puede visitar el Link http://quast.sourceforge.net/docs/manual.html para comprender todas las métricas.

5.2.2 Bandage

Bandage (Wick *et al.,* 2015) es un programa GUI que permite a los usuarios interactuar con los gráficos de ensamblaje hechos por ensambladores *de novo* como Velvet, SPAdes, MEGAHIT y otros. Los gráficos de ensamblaje *de novo* contienen no solo contigs ensamblados, sino también las conexiones entre esos

contigs, que anteriormente no eran fácilmente accesibles. Bandaje visualiza gráficos de ensamblaje, con conexiones, utilizando algoritmos de diseño de gráficos.

Sitio Web: La herramienta no cuenta con una página web, su ejecutable está alojado en GitHub https://rrwick.github.io/Bandage/

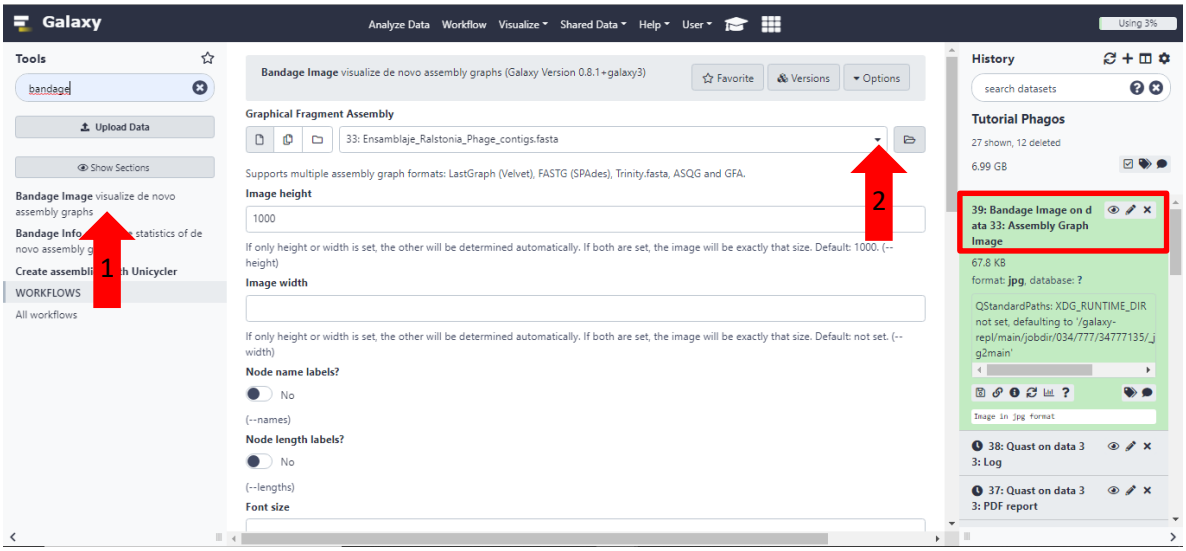
Compatibilidad: Es compatible con Windows, Mac y Linux

Entrada: La herramienta acepta ensamblajes en formato FASTA. Los archivos pueden estar comprimidos con zip, gzip o bzip2.

Instrucciones

1. En el buscador filtramos por "**Bandage**" y elegimos la opción "**Bandage Image visualize de novo assembly graphs**"

2. En el panel central, la opción "**Graphical Fragment Assembly**" elegimos el archivo contig.fasta, podemos dejar las métricas como preestablecido y ejecutamos.



Para ver el gráfico asociado a nuestro ensamblaje, damos clic en el icono con forma de ojo en el proceso correspondiente al Bandage en el panel del historial.

5.3 Refinamiento del genoma (Polishing)

La finalidad del Polishing es mejorar significativamente los ensamblajes del genoma en borrador corrigiendo bases, arreglando ensamblajes incorrectos y llenando huecos (gaps). Se hace uso de herramientas "multifuncionales"

totalmente automatizada para corregir conjuntos de borradores y variantes de secuencia de llamada de varios tamaños, incluidas inserciones y eliminaciones muy grandes.

5.3.1 Generando archivos SAM

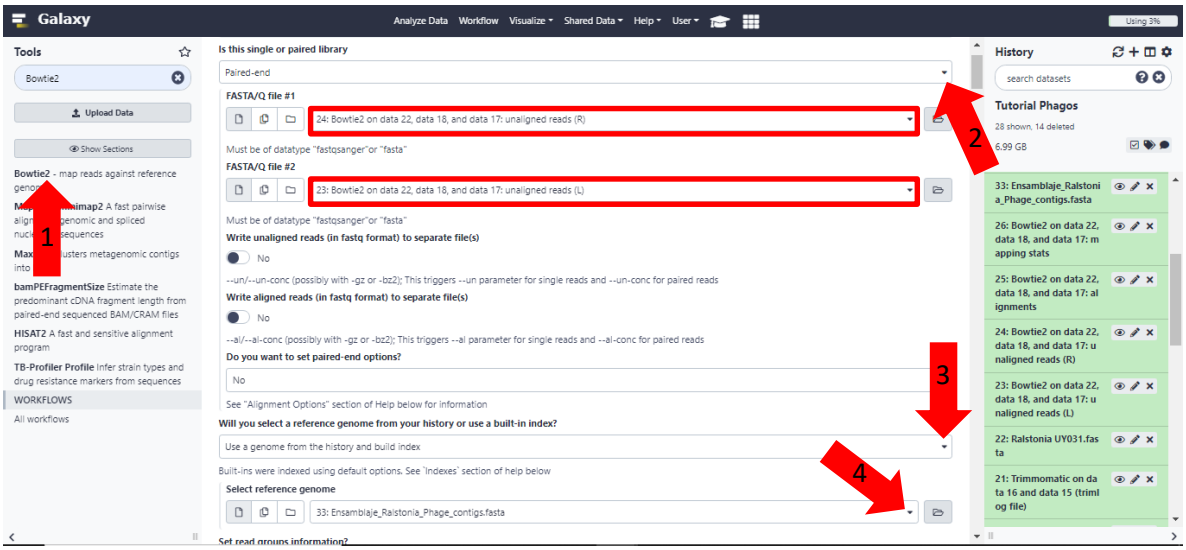
El primer paso a seguir, para poder realizar el pulido de nuestro ensamblaje mediante el programa Pilon, es generar un archivo en formato SAM mediante **Bowtie2**.

1. Filtramos en el buscador de herramientas por la palabra "**Bowtie2**" y la abrimos.

2. En el panel central en el apartado "**Is this single or paired library**" elegimos la opción acorde con la naturaleza de nuestros datos (paired-end o single-end). En las opciones **FASTA/Q file #1** y **FASTA/Q file #2,** seleccionar los archivos que descargamos en el paso **3.6 Descarga de datos preprocesados,** R1 **unaligned read R** y R2 **unaligned read L** respectivamente.

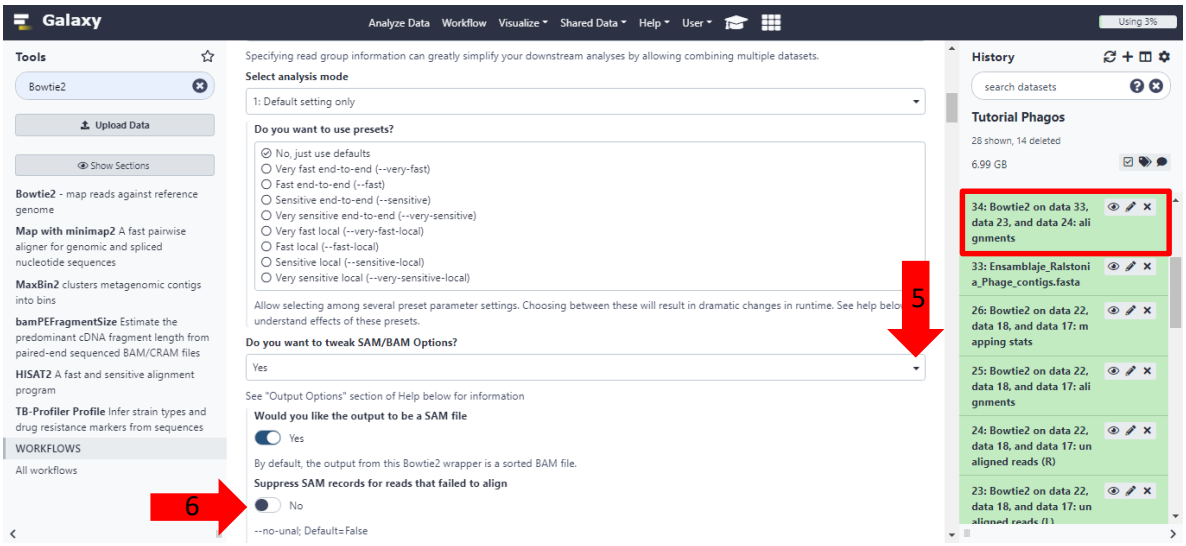
3. En el apartado "**Will you select a reference genome from your history or use a built-in index?**" seleccionamos la segunda opción "**Use a genome from this history**". Esto nos permitirá seleccionar un archivo previamente subido a nuestro historial.

4. En el apartado "**Select reference genome**" debemos elegir el archivo contig.fasta que descargamos previamente en el paso **5.1.1 Descarga datos ensamblados** y cargamos en el galaxy en el paso **5.2.1 Quast** instrucción 1.



5. Bajamos en el panel central y cambiamos la opción "**Do you want to tweak SAM/BAM Options?**" a "**YES**"

6. En las opciones que se despliegan, en la que dice “**Would you like the output to be a SAM file**”, debemos cambiarla a “**YES**” y ejecutamos.

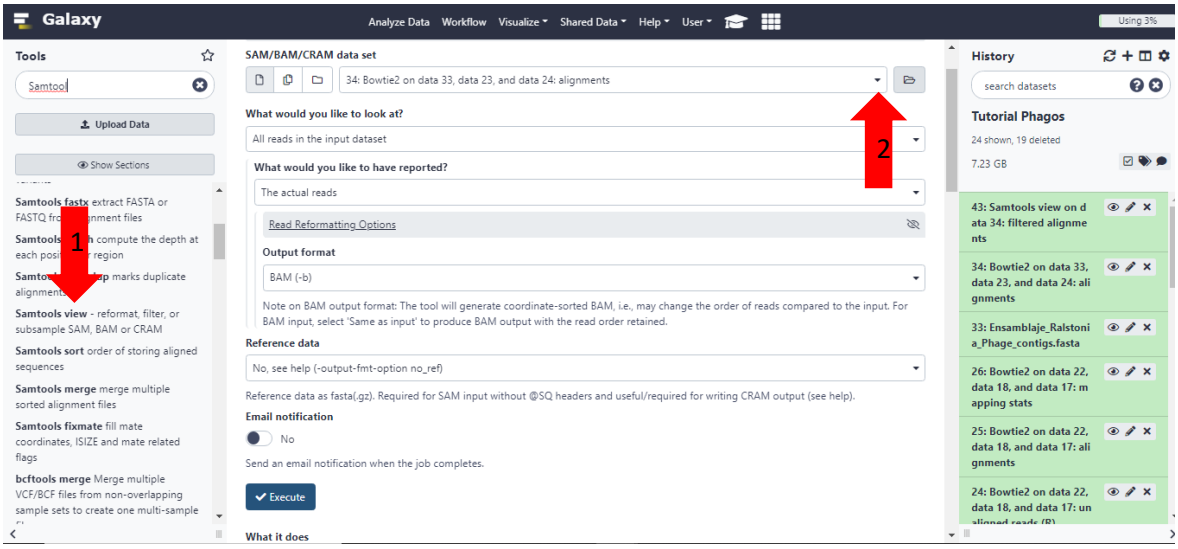


El archivo SAM generado lo usaremos para convertirlo a formato binario BAM, para ello utilizaremos la herramienta Samtool (Li H. *et al., 2009*)

5.3.2 Convirtiendo SAM a BAM

1. En el buscador, filtramos por Samtool y elegimos la opción “**Samtools view** - reformat, filter, or subsample SAM, BAM or CRAM”

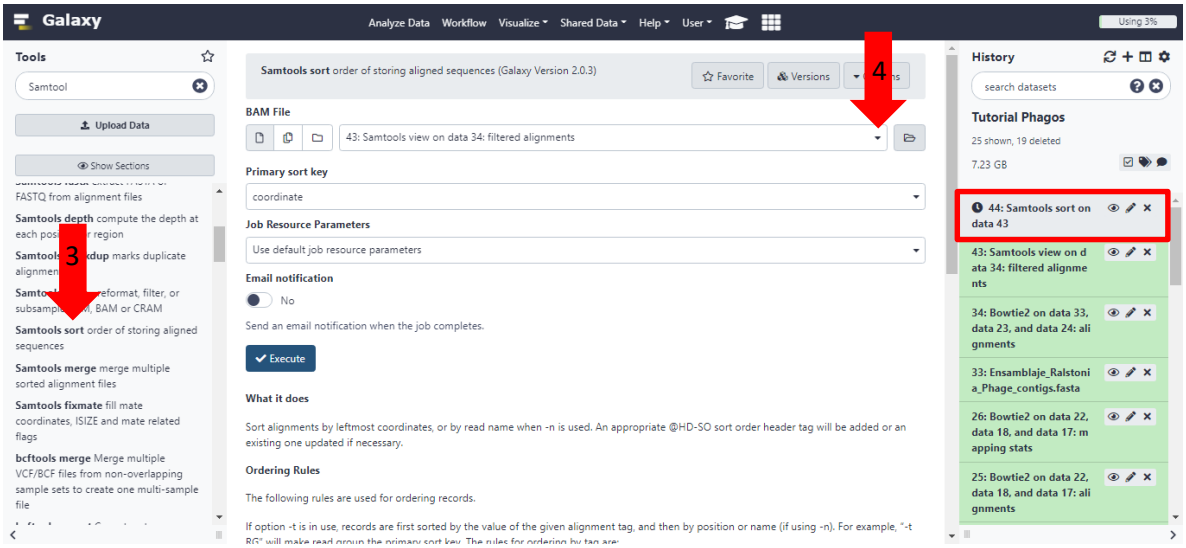
2. En el panel central, en la opción “**SAM/BAM/CRAM data set**” cargaremos el archivo SAM creado en el paso anterior y ejecutamos.



3. Para reordenar las alineaciones por coordenadas de los datos contenidos en el archivo BAM, con el fin de optimizar el proceso de “Polishing”, en el panel

izquierdo buscaremos la opción de “**Samtools sort order of storing aligned sequences**” y daremos clic.

4. En la opción “**BAM File**” seleccionamos el archivo BAM creado anteriormente y ejecutamos.



El archivo de salida de este proceso será usado para realizar el pulido de nuestro genoma mediante el uso de herramientas todo en uno.

5.3.3 Pilon

Pilon (Walker *et al.,* 2014) es una herramienta de software que se puede utilizar para mejorar automáticamente los conjuntos de borradores o para encontrar variaciones entre cepas. Pilon utiliza el análisis de alineación de lectura para identificar inconsistencias entre el genoma de entrada y la evidencia en las lecturas. Luego intenta realizar mejoras en el genoma de entrada, que incluyen:

- Diferencias de base única.
- Pequeños indels.
- Eventos de sustitución de bloque o indel más grandes.
- Relleno de huecos.
- Identificación de ensamblajes incorrectos locales, incluida la apertura opcional de nuevos huecos.

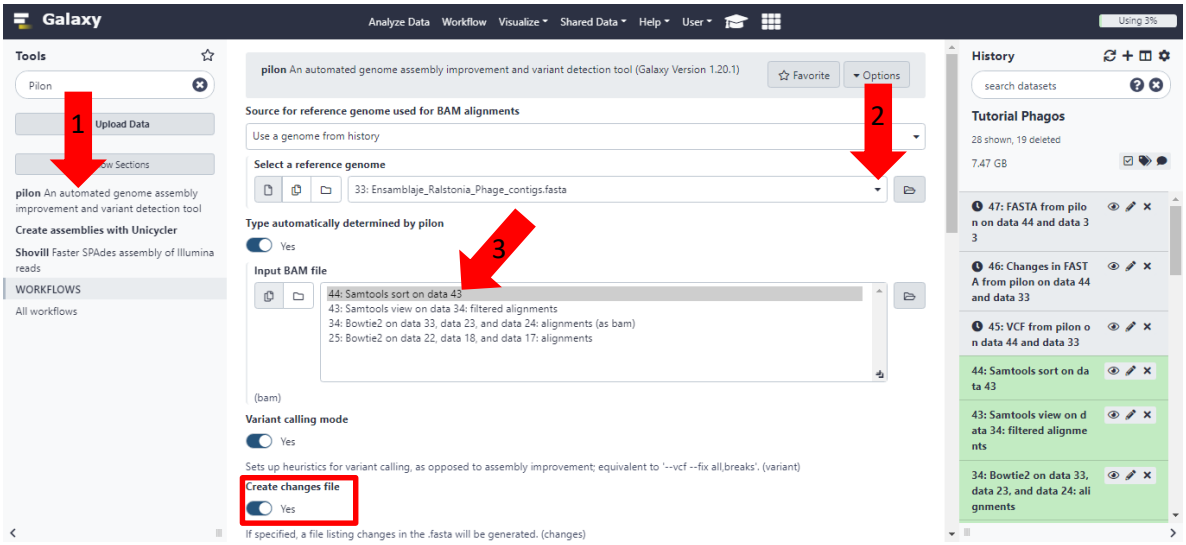
Entrada: Pilon requiere como entrada un archivo FASTA del genoma junto con uno o más archivos BAM de lecturas alineadas con el archivo FASTA de entrada.

Instrucciones:

1. En el buscador, filtramos por “**Pilon**” y seleccionamos la opción “**Pilon An automated genome assembly improvement and variant detection tool**”.

2. En la opción “**Select a reference genome**” seleccionamos nuestro archivo ensamblado, es decir el contigs.fasta.

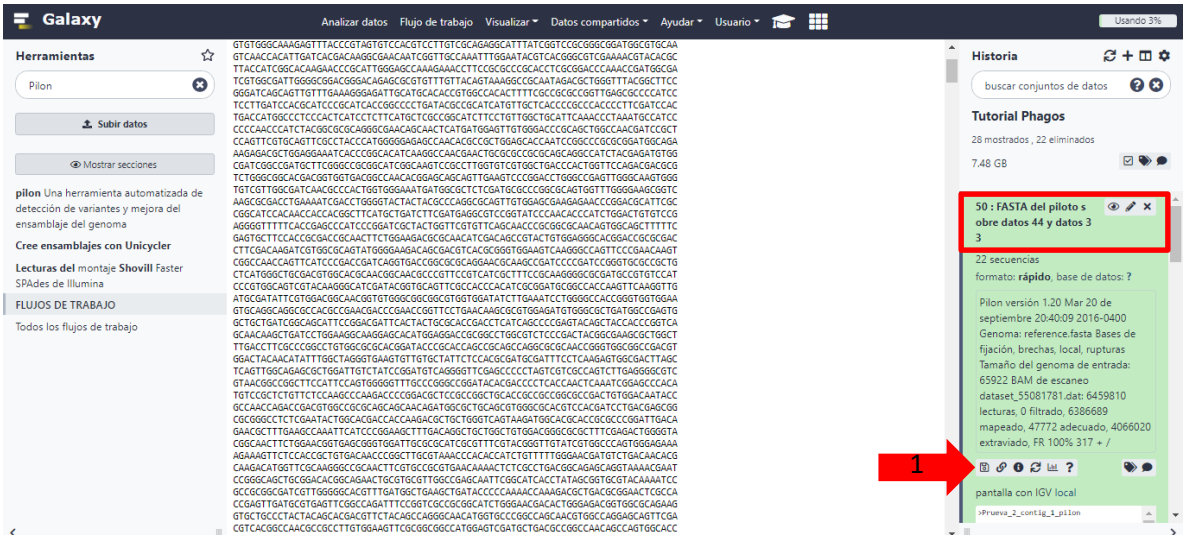
3. En el “**Input BAM file**” seleccionamos el archivo BAM ordenado.



En el panel central, si queremos ver el archivo con los cambios generados, debemos cambiar la opción “**Create changes file**” a “**Yes**” y como último ejecutamos el proceso.

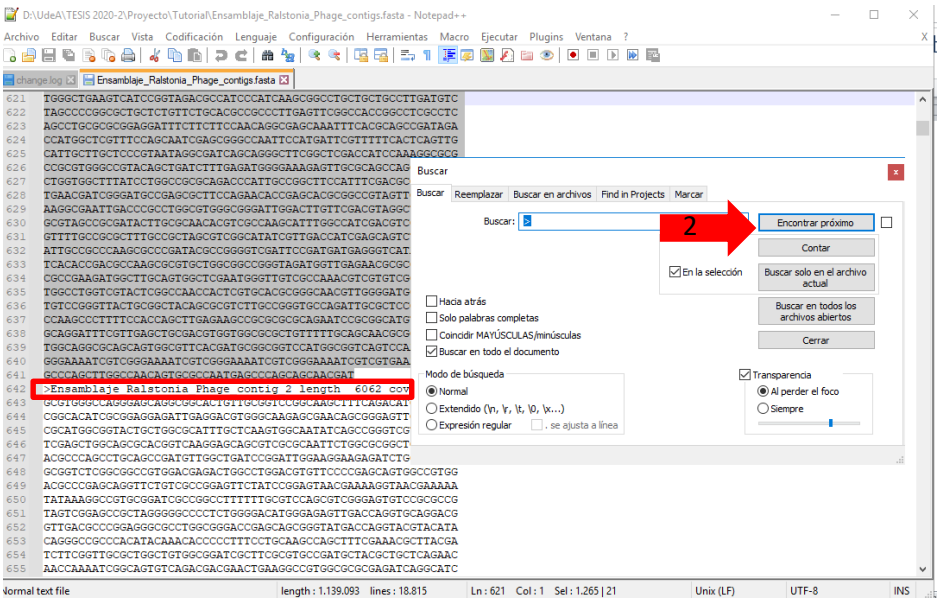
5.3.3.1 Descarga de datos post polishing

1. Vamos al proceso Fasta compilado por el Pilon y descargamos el archivo contigs.fasta procesado, dando clic en el icono con forma de disquete.

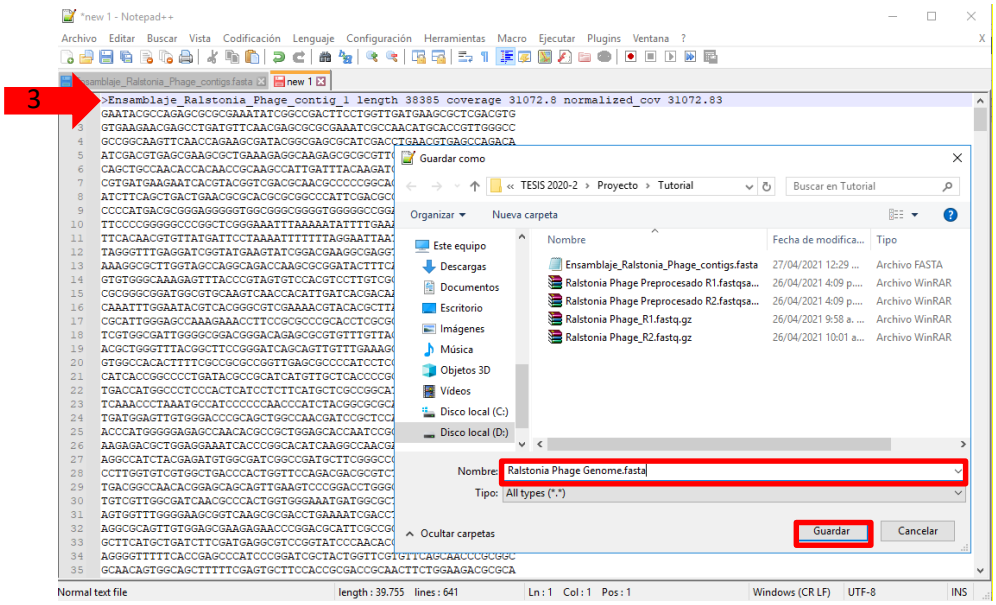


2. Abrimos el archivo descargado con el editor de texto preferido, ubicamos el puntero en la primera fila. Vamos al apartado de buscar y filtramos por el símbolo

“**>**”. Esto nos desplazara entre contigs, necesitaremos copiar el contig de interés (Ver los resultados del **Bandage** en el archivo .html generado tras el ensamblaje, al igual que las demás métricas para saber cuál contig es de interés), en nuestro caso es el primer contig. Seleccionamos todas las secuencias nucleotídica y la copiamos (Esta no es una práctica común en el análisis de genomas, nosotros la utilizamos debido a la fragmentación de nuestro ensamblaje).



3. Creamos un nuevo archivo de texto en el que pegaremos la secuencia nucleotídica del contig de interés. Buscamos en el apartado de archivo la opción de “**Guardar como**”, vamos la carpeta de trabajo y renombramos el archivo poniendo la extensión **.fasta** al final, como último cliquear en “**guardar**”.



Este archivo es el que usaremos para realizar la anotación de nuestro Genoma.

6. Anotación del genoma

La anotación del genoma consiste en interpretar o extraer la información que en él se encuentra, esto es importante porque así podemos darle un significado biológico a los elementos dentro del genoma y por ende podemos realizar posteriores análisis teniendo en cuenta este componente biológico. La anotación de un genoma comprende principalmente dos etapas: la anotación estructural (predicción de regiones codificantes) y la anotación funcional (asignación de información biológica de las regiones codificantes predichas).

6. 1 Anotación del genoma por medio de (PATRIC)

En este caso haremos uso del servicio de anotación de PATRIC, el cual utiliza el kit de herramientas RAST (RASTtk) para realizar una anotación del genoma. Una vez llevado a cabo el proceso obtendremos archivos de salida los cuales contiene nuestro genoma con los componentes estructurales ya definidos e inclusive con posibles funciones biológicas para las regiones predichas.

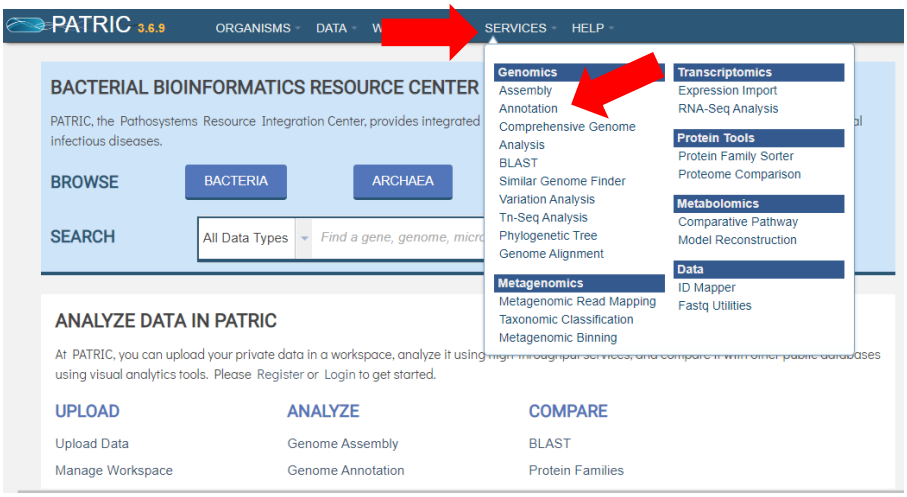
Sitio web: http://rast.nmpdr.org

Entrada: Archivos Fasta o FastQ.

Instrucciones:

1. En la parte superior de la plataforma PATRIC, busque el menú "**Services**" y haga clic sobre él, posteriormente seleccione el submenú de anotación, allí debemos cargar nuevamente nuestros datos que en este caso corresponden al archivo que guardamos con el nombre de **Ralstonia Phage Genome.fasta.**

Nota: Para realizar el proceso de cargar nuestros datos podemos dirigirnos nuevamente al apartado **4.2** "**Carga de datos**".



Esto abrirá el interfaz del servicio de anotación del genoma que utiliza el kit de herramientas RAST (RASTtk), debemos entonces proporcionar la información que se nos solicita.

2. Cliqueamos sobre la flecha que se encuentra en la casilla de "**Contigs**" para luego proporcionar el archivo **.fasta** el cual contiene nuestro genoma ensamblado.

3. En el apartado de "**Domain**" debemos seleccionar el dominio taxonómico del cual hace parte nuestro genoma en cuestión. En este caso elegiremos **Bacteriophage**.

4. En la ventana de "**Taxonomy Name**" debe especificarse la categoría taxonómica a nivel de género o menor para así obtener las últimas predicciones de las familias de proteínas que se han registrado para nuestro genoma. Seleccionamos entonces **Bacteriophage sp**.

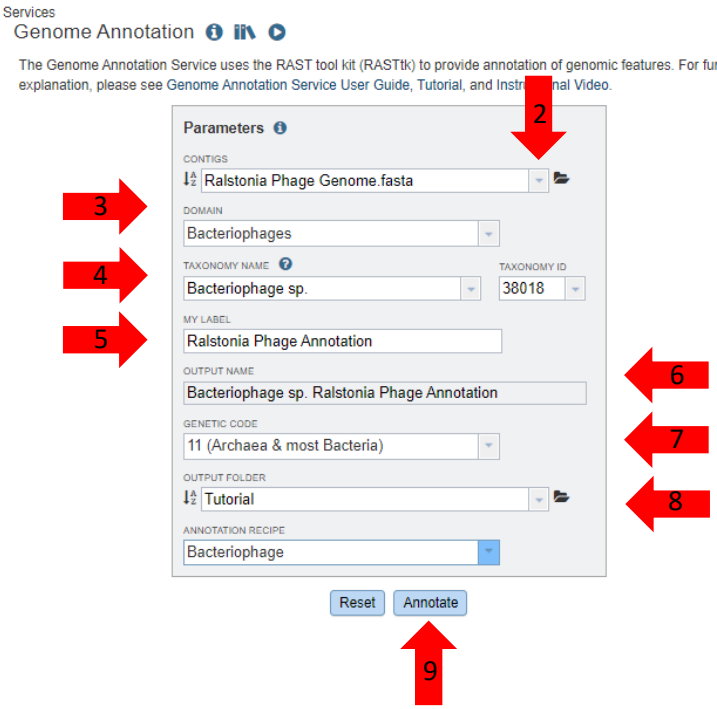
5. En la casilla "**My Label**" deberemos proporcionar el nombre con el cual queremos identificar nuestro documento una vez se realice el proceso de la auto anotación.

6. El "**Output Name**" es el nombre combinado entre "**Taxonomy Name**" y "**My Label**". Cuando finalice el proceso, este será el nombre con el cual encontraremos nuestro archivo en el área de trabajos de PATRIC.

7. En la ventana de "**Genetic Code**" deberemos especificar cuál es la traducción de codones que se utilizaran para la llamada de los genes. En este caso dejaremos la opción de se encuentra por defecto.

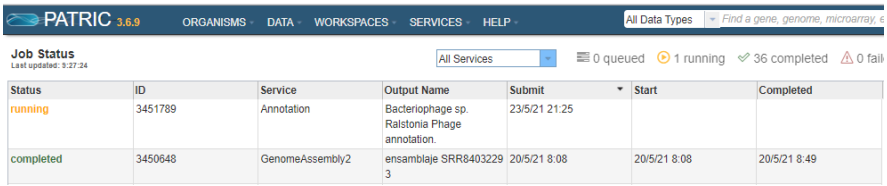
8. En “**Output folder**” seleccionamos la carpeta del espacio de trabajo en la cual queremos que se guarden los resultados. Esta carpeta fue creada con anterioridad con el nombre de **Tutorial**.

9. Una vez se hayan definido los parámetros específicos para nuestra anotación, procederemos entonces a dar clic en el botón de “**Annotate**” para dar inicio a este proceso.



Aparecerá entonces un mensaje debajo del panel principal el cual indica que nuestro trabajo está ahora en cola, a la espera de iniciar su ejecución (esto puede tardar algunos segundos).

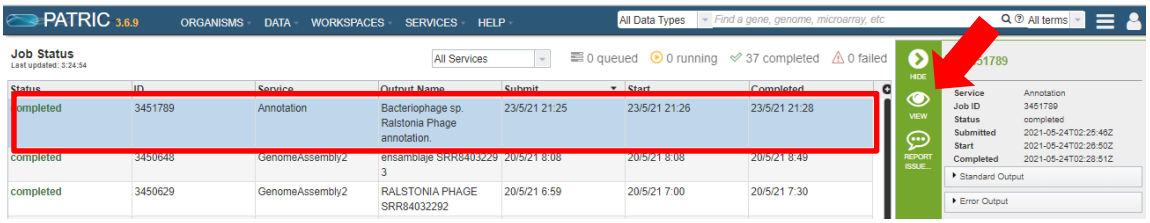
Para ver el progreso de nuestro trabajo basta con dar clic sobre el indicador de trabajos, el cual se encuentra en la esquina inferior derecha de la plataforma PATRIC. Esta acción nos llevará a una ventana, donde se encuentra el progreso de todas las actividades que estamos realizando en el momento, además de un registro de las actividades ejecutadas con anterioridad.



6. 1.1 Descarga de genoma anotado

Podemos darnos cuenta que se ha llevado a cabo el trabajo solicitado una vez el estado del trabajo cambia a "**Completed**". Para revisar entonces los archivos que ha generado este proceso deberemos seleccionar el trabajo y dar clic en el icono 

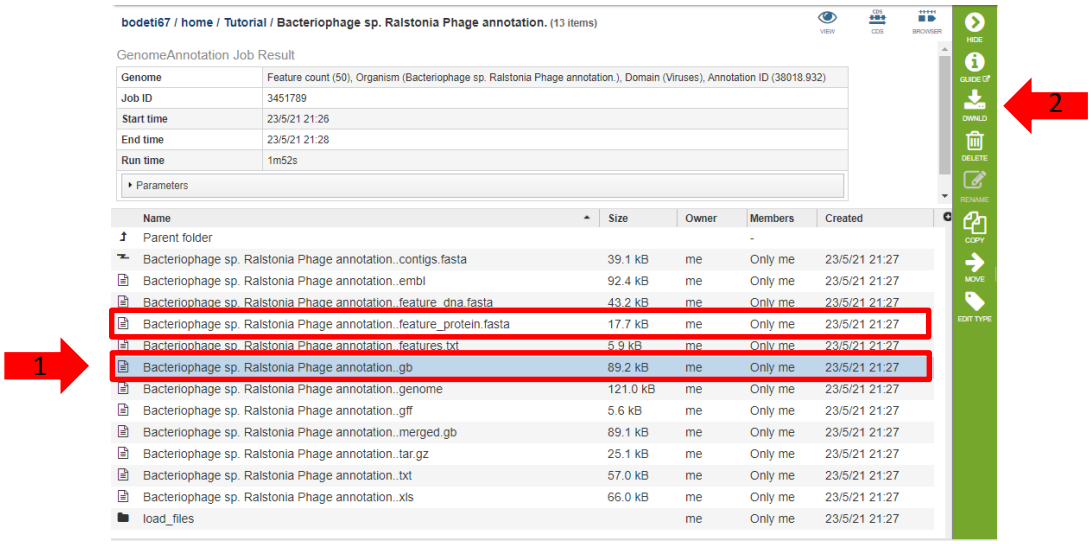
que se encuentra sobre la barra vertical de color verde.



Esto abrirá nuevamente una ventana en la cual se muestra un encabezado que describe nuestro trabajo. Además, allí se localizan todos los archivos de salida que están a nuestra disposición. Para descargar cualquier archivo realice los siguientes pasos:

1. Seleccionar el archivo de interés. En este caso haremos uso del archivo .**gb** el cual contiene nuestro genoma anotado en formato GenBank.

2. Dar clic sobre el botón de descargas "**DWNLD**", disponible en la barra de acciones.



Recuerde renombrar el archivo y guardarlo en la carpeta de trabajo.

Si desea tener mayor información sobre los demás archivos generados por la anotación en la plataforma de PATRIC, por favor, diríjase al link https://docs.patricbrc.org/user_guides/services/genome_annotation_service.html

6. 2 Herramientas bioinformáticas para la anotación funcional

Para realizar la anotación estructural actualmente existen muchas herramientas de uso fácil que realizan con gran eficacia la llamada de los genes y demás elementos genéticos dentro de un genoma, por ejemplo, RASTtk utiliza tanto Glimmer3 (Delcher *et al.,* 2007) como Prodigal (Hyatt *et al.,* 2010) para hacer más preciso este proceso. Por otro lado, la anotación funcional es un proceso un poco más complejo, ya que no está tan automatizado y requiere en muchos casos la comparación de los genes encontrados en la anotación frente a diferentes bases de datos públicas, con el fin de intentar generar anotaciones funcionales mucho más completas debido a que este es material de gran importancia para los posteriores análisis comparativos.

En los siguientes pasos presentaremos una breve introducción a algunas de las herramientas web que son ampliamente utilizadas para la anotación funcional en fagos. Tenga en cuenta que estas no son de uso obligatorio y por el contrario el usuario puede utilizar distintas bases de datos y herramientas dependiendo de sus criterios y fines de investigación.

6. 2 .1 BLAST

BLAST (Basic Local Alignment Search Tool), es un software ampliamente utilizado para realizar alineamientos locales de secuencias de ADN, ARN o aminoácidos. Este programa tiene la capacidad de comparar una secuencia problema (query cover) frente a todas las secuencias de una base de datos en específico, además genera significancia estadística de las coincidencias encontradas. Esto es de gran ayuda puesto que esta herramienta puede ser útil a la hora de inferir relaciones estructurales, funcionales o evolutivas entre diversas secuencias, y de este modo identificar nuevos miembros de alguna familia de genes o proteínas.

El NCBI-BLAST es un software el cual usa por defecto el NCBI para realizar búsquedas de secuencias en sus bases de datos. En esta ocasión haremos uso de la herramienta BLASTp (herramienta de búsqueda de alineación local básica para secuencias de proteínas) desde la página del NCBI la cual en este caso utiliza base de datos de proteínas. Si el parecido entre las secuencias es superior al 90%, puede tratarse de proteínas homólogas y es bastante probable que las anotaciones de las secuencias en las bases de datos también sean válidas para nuestra muestra problema. Este procedimiento es especialmente útil en aquellas secuencias a las que no se les asignó una posible función luego de la auto-anotación.

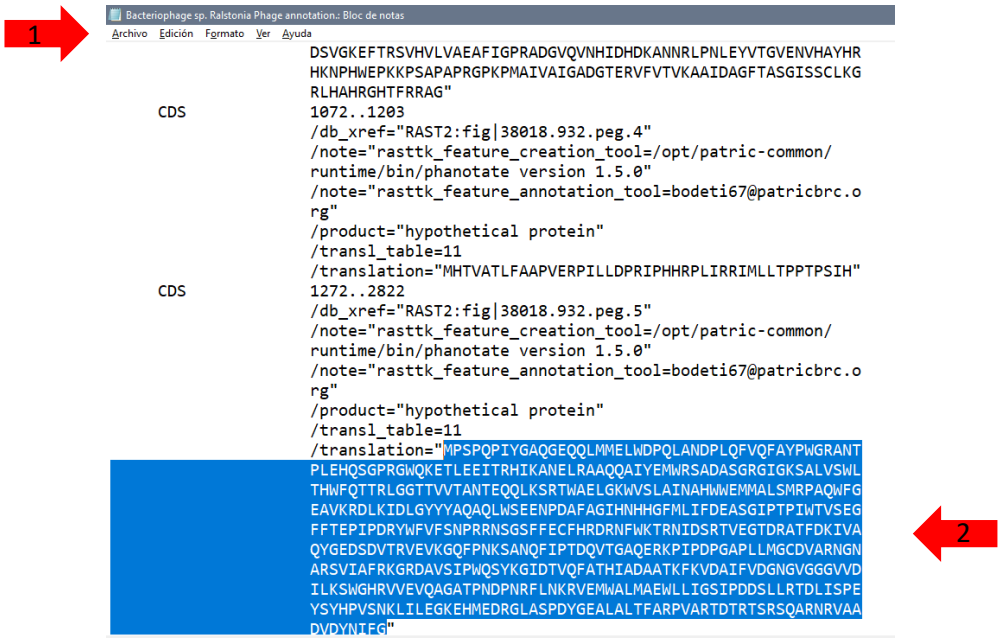
Sitio web: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Entrada: Archivos Fasta con secuencias ADN, ARN o aminoácidos.

Instrucciones:

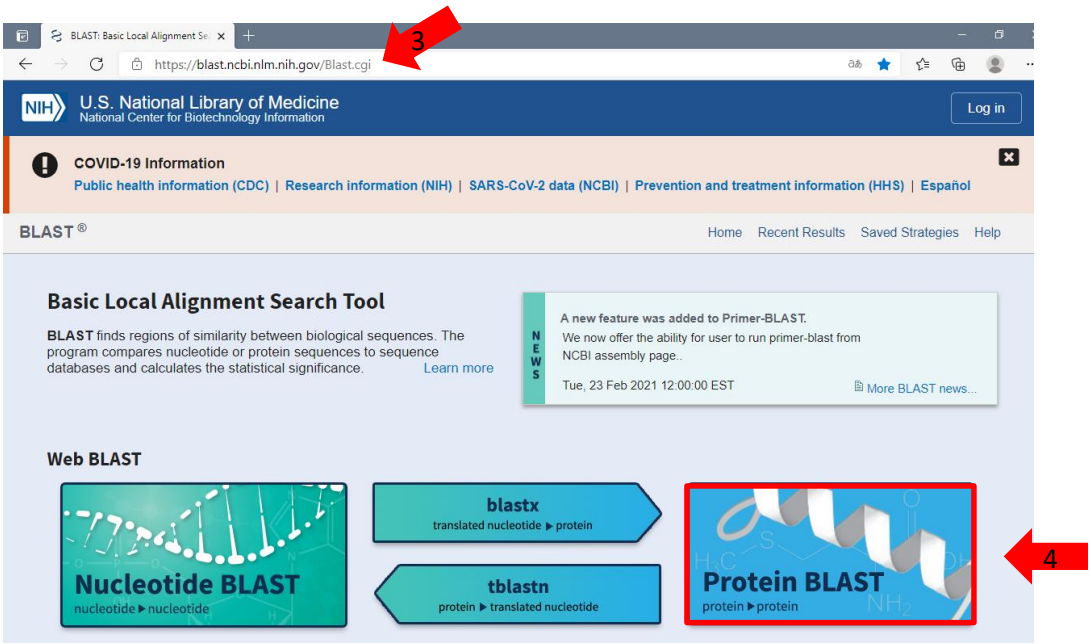
1. Dentro de nuestra carpeta de trabajo abrimos el archivo en formato **.gb** identificado con el nombre **Bacteriophage sp. Ralstonia Phage annotation,** el cual corresponde a la auto anotación realizada en PATRIC.

2. Copie toda la secuencia de aminoácidos **“translation”** correspondiente al CDS del cual desea realizar el alineamiento.



3. Abrir una nueva ventana desde su navegador de preferencia e ingresar al sitio web del NCBI-BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

4. Seleccione la ventana de protein BLAST.

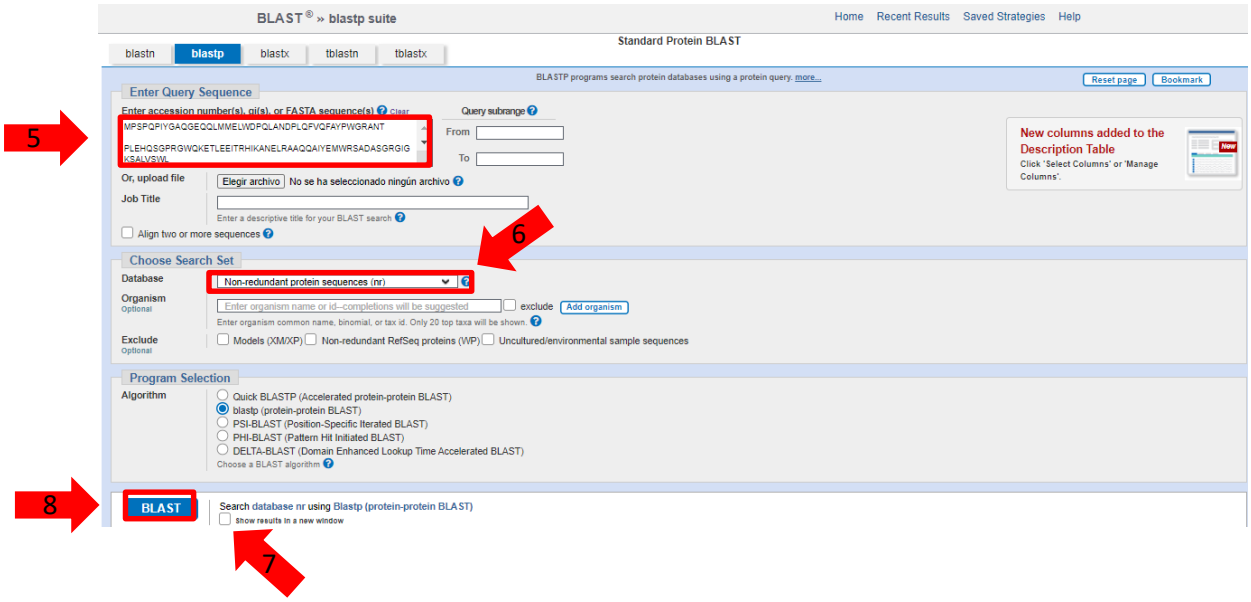


5. Dentro del rectángulo que aparece bajo **“Enter query sequence”** pegar la secuencia de aminoácidos que copió anteriormente en el paso 2. (Si desea introducir un nombre a esta búsqueda, diríjase al rectángulo **“Job Tittle”**)

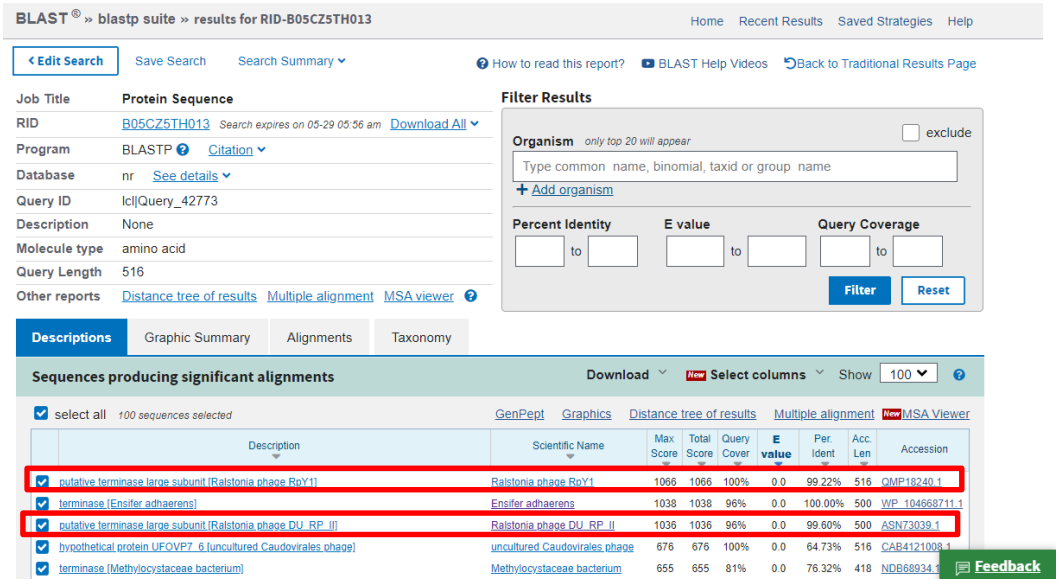
6. Asegúrese que dentro de la caja **“Choose Search Set”** se encuentre seleccionada la base de datos **“Non-redundant protein sequences (nr)”.**

7. Si desea que los resultados se muestren en otra ventaja, debe hacer clic sobre el cuadro que aparece en la parte inferior de la pantalla.

8. Para dar inicio a la búsqueda, presione el botón azul **“BLAST”.** Inmediatamente se abrirá una nueva ventana la cual contiene información sobre el estado de la búsqueda. Lo siguiente es esperar a que se generen los resultados.



El programa BLASTp realizará una búsqueda y luego mostrará en una nueva página los resultados, representados en diferentes gráficas y estadísticas, los cuales nos brindan mayor información sobre las secuencias de las cuales BLAST encontró mayor coincidencia con la secuencia consulta. La siguiente imagen es una representación de cómo debería lucir la página una vez finaliza el proceso.



Diríjase a la siguiente guía de apoyo para obtener mayor información sobre el uso de BLAST y de cómo interpretar mejor los resultados: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

6. 2 .2 HHpred

HHpred (Zimmermann *et al.,* 2018) es un servidor interactivo el cual permite realizar la búsqueda de homologías a partir una sola secuencia proteica en diferentes bases de datos al mismo tiempo. Este método de búsqueda de secuencias en diversas bases de datos ha facilitado la rápida detección de información funcional referente a secuencias no caracterizadas. Su gran sensibilidad ha aportado inclusive en algunos casos al entendimiento de la evolución en algunas proteínas, esto gracias a que a menudo detecta relaciones homólogas más remotas entre las mismas. Debido a su velocidad de ejecución, facilidad de uso, flexibilidad, sensibilidad, bases de datos y demás, HHpred se ha postulado en los últimos años como una herramienta ampliamente utilizada por los investigadores para realizar inferencias significativas basadas en homología.

Sitio web: https://toolkit.tuebingen.mpg.de/tools/hhpred

Entrada: Secuencia de proteínas (en formato FASTA o como texto sin formato).

Instrucciones

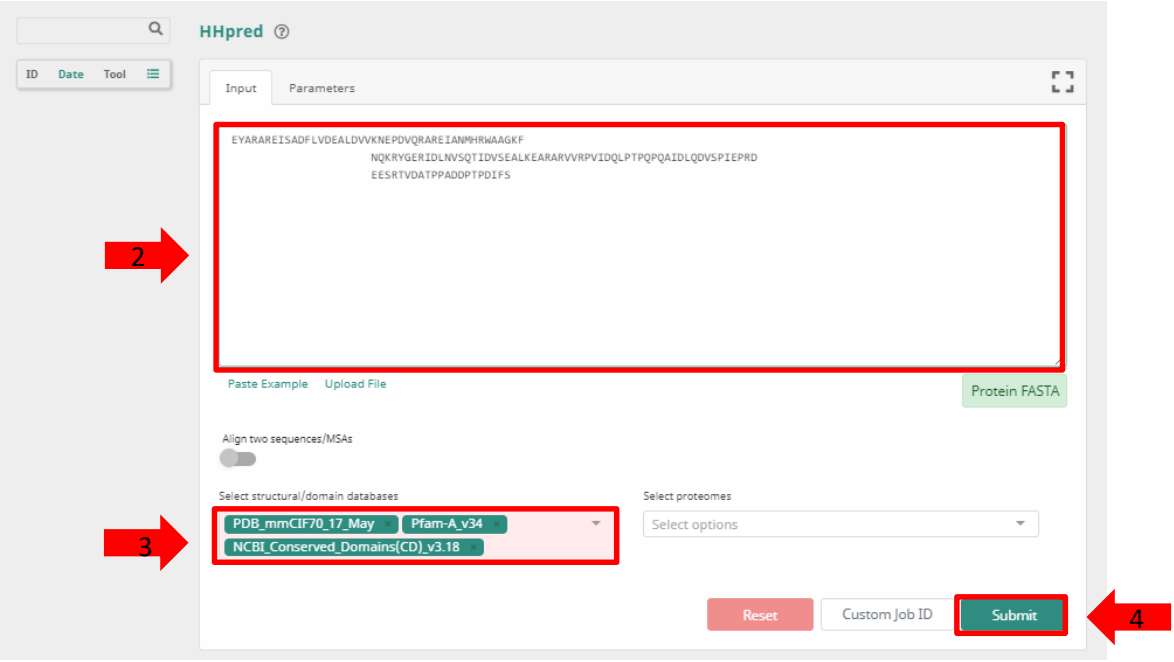
1. **I**ngresa al siguiente link para abrir la ventana principal de HHpred: https://toolkit.tuebingen.mpg.de/tools/hhpred

2. Pegar la secuencia consulta (secuencia proteica) dentro del recuadro en blanco de la pestaña **“**input”. Estas secuencias se pueden recuperar del archivo

protein.FASTA descargado anteriormente en el apartado **6.1.1 "Descarga de genoma anotado".**

Opcional: puede cargar directamente su archivo seleccionando "**Upload file".**

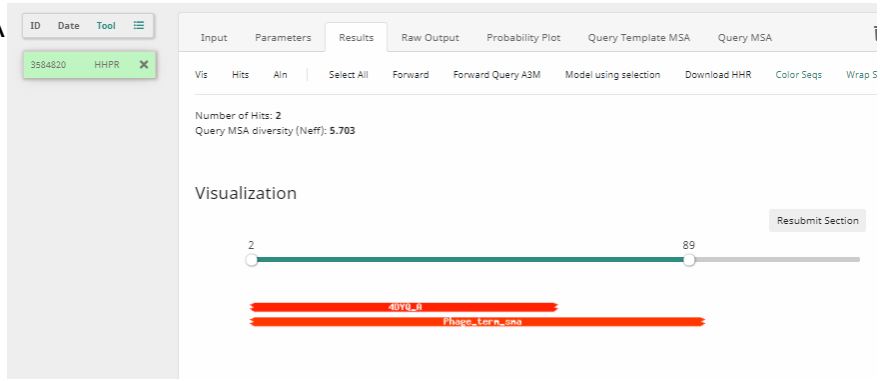
3. Elija las bases de datos contra las cuales desea comparar su secuencia. Para esto seleccione la lista desplegable en **"Select structural/domain databases".** Aconsejamos seleccionar las bases de datos que aparecen en la imagen siguiente**.**



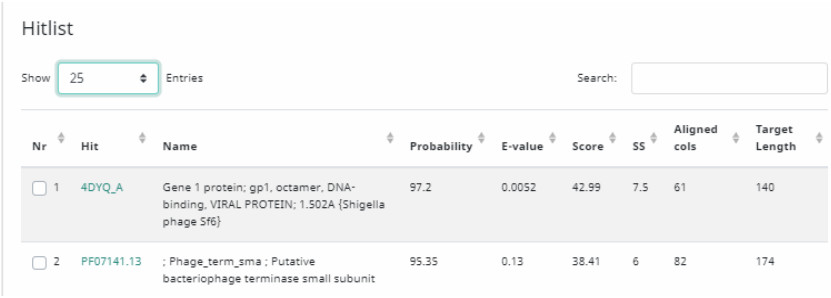
4. Inicie su búsqueda presionando el botón **"Submit"** ubicado en la parte inferior derecha del recuadro principal**.** Espere algunos minutos hasta que se realice el proceso. Sabremos que el proceso finalizó cuando el indicador del trabajo pase de color amarillo a verde.

En la nueva ventana emergente se presentan 3 principales secciones. A) visualización, muestra en barras la secuencia a consultar y todas aquellas con las que se encontró coincidencia, además, muestra la cobertura con respecto a la secuencia problema. B) lista de aciertos, proporciona un listado de estadísticas relacionadas con la probabilidad de acierto para las secuencias relacionadas con la consulta. C) Alineaciones, en este apartado se presenta las alineaciones que fueron acertadas, así como aquellas que no coincidieron entre la secuencia consulta y las secuencias en las bases de datos.

A



B



C

6.2.3. tRNAscan-SE

tRNAscan-SE (Chan & Lowe, 2019) es una herramienta ampliamente usada con el fin de encontrar secuencias no codificantes de ARN de transferencia (tRNA) en un genoma. Este software toma la secuencia genómica y realiza una búsqueda dentro de bases de datos de tRNA, y utiliza modelos de covarianza avanzados para mejorar la predicción y clasificación de los tRNA que se encuentran dentro de nuestra secuencia consulta. Luego de realizado el proceso, tRNAscan-SE proporciona al usuario las coordenadas genómicas, la estructura secundaria y la función predicha de los tRNA que se encuentran dentro del genoma.

Sitio web: http://trna.ucsc.edu/tRNAscan-SE/

Archivos de entrada: Archivo FASTA con el genoma completo en secuencia de nucleótido.

Compatibilidad: La versión completa de tRNAscan-SE se encuentra disponible principalmente para sistema operativo UNIX, sin embargo, Se ha desarrollado un sitio web con una versión muy completa de este software.

Otras opciones: ARAGORN http://www.ansikte.se/ARAGORN/

Instrucciones

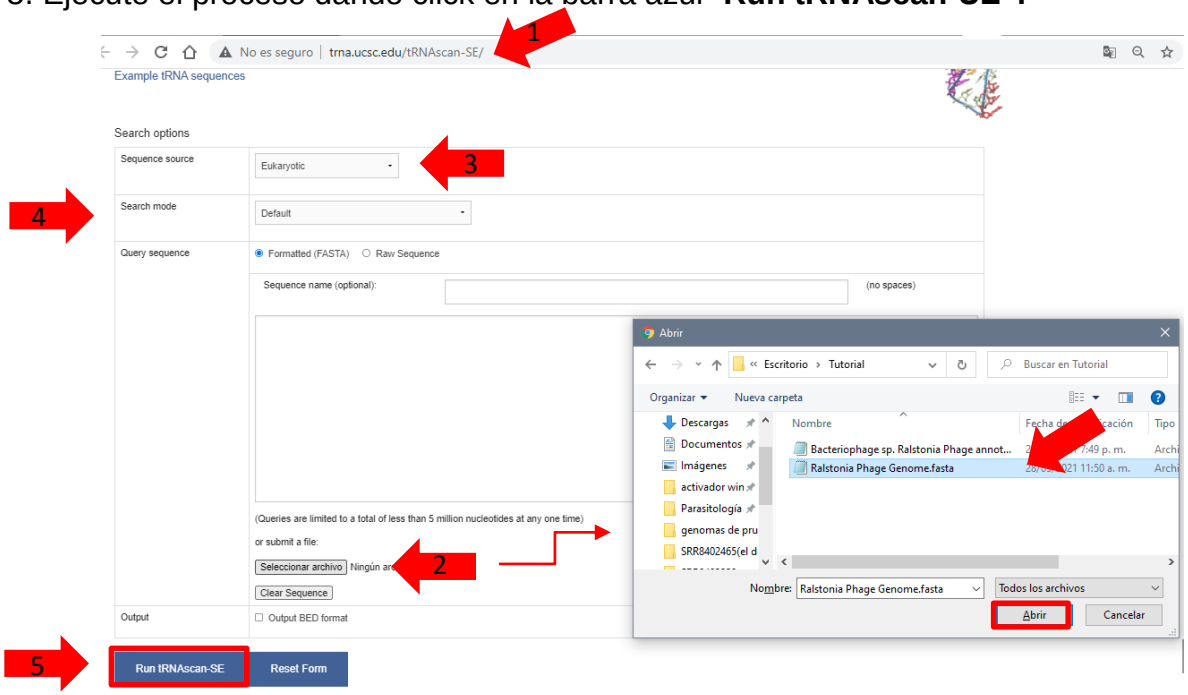
1. Ingrese al servidor web de tRNAscan-SE copiando el siguiente link: http://trna.ucsc.edu/tRNAscan-SE/

2. Dar click sobre "**Seleccionar Archivo**" para cargar la secuencia consulta. En este trabajo la secuencia consulta se encuentra dentro de la carpeta **"Tutorial"** con el nombre de "**Ralstonia Phage Genome.fasta",** este archivo contiene el ensamblaje del genoma en formato FASTA.

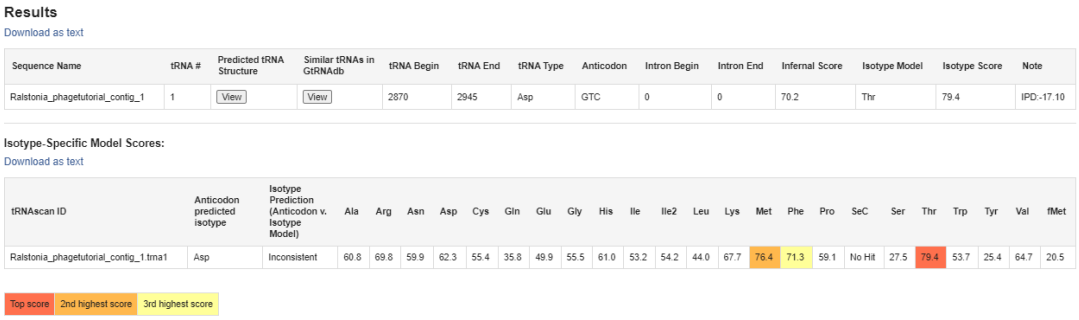
3. Si conoce la fuente de la secuencia deberá proporcionarla en el apartado "**Sequence source",** dentro de los parámetros de búsqueda. En este caso se utilizó la opción "**Bacterial"**.

4. Asegurarse que la "**Sequence source"** aparezca por defecto.

5. Ejecute el proceso dando click en la barra azul "**Run tRNAscan-SE".**



La búsqueda del genoma tardará algunos segundos, el tiempo de espera depende del tamaño de la secuencia y de la cantidad de búsquedas que la página web está realizando en el momento. Cuando se complete la búsqueda los resultados se muestran representados en dos tablas principales que incluye coordenadas, puntuaciones isotipo del tRNA, puntuaciones de los genes predichos y demás especificaciones.



Results

Download as text

Sequence Name	tRNA #	Predicted tRNA Structure	Similar tRNAs in GtRNAdb	tRNA Begin	tRNA End	tRNA Type	Anticodon	Intron Begin	Intron End	Infernal Score	Isotype Model	Isotype Score	Note
Ralstonia_phagetutorial_contig_1	1	View	View	2870	2945	Asp	GTC	0	0	70.2	Thr	79.4	IPD:-17.10

Isotype-Specific Model Scores:

Download as text

tRNAscan ID	Anticodon predicted isotype	Isotype Prediction (Anticodon v. Isotype Model)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Ile2	Leu	Lys	Met	Phe	Pro	SeC	Ser	Thr	Trp	Tyr	Val	fMet
Ralstonia_phagetutorial_contig_1.trna1	Asp	Inconsistent	60.8	69.8	59.9	62.3	55.4	35.8	49.9	55.5	61.0	53.2	54.2	44.0	67.7	76.4	71.3	59.1	No Hit	27.5	79.4	53.7	25.4	64.7	20.5

Top score | 2nd highest score | 3rd highest score

Para mayor información sobre cómo hacer mejor uso de la herramienta tRNAscan-SE y de cómo interpretar los resultados, por favor, diríjase al siguiente link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768409/

Opcional: Si desea corroborar o comparar los resultados encontrados en tRNAscan-SE, aconsejamos utilizar la herramienta de sitio web ARAGORN (Laslett & Canback, 2004), esta herramienta aunque es menos usada también localiza con buena precisión secuencias de tRNA en un genoma. Puede hacer uso de ARAGORN desde el sitio web http://www.ansikte.se/ARAGORN/

6.2.4 TMHMM

TMHHM (Krogh *et al.,* 2001) es una herramienta la cual funciona bajo modelos ocultos de Markov, cuya finalidad es predecir e identificar la topología de proteínas transmembrana dentro de una secuencia genómica. Se ha demostrado que TMHHM predice correctamente entre el 97 y 98% las hélices estructurales de proteínas transmembrana. Además, tiene gran capacidad para distinguir entre proteínas solubles y de membrana. Este alto grado de precisión ha permitido la identificación de un buen número de secuencias de grandes colecciones de genomas.

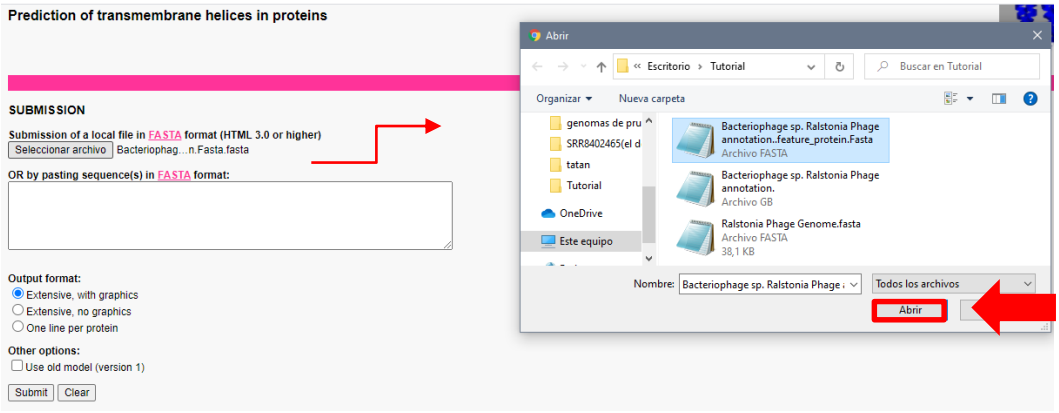
Sitio web: http://www.cbs.dtu.dk/services/TMHMM/

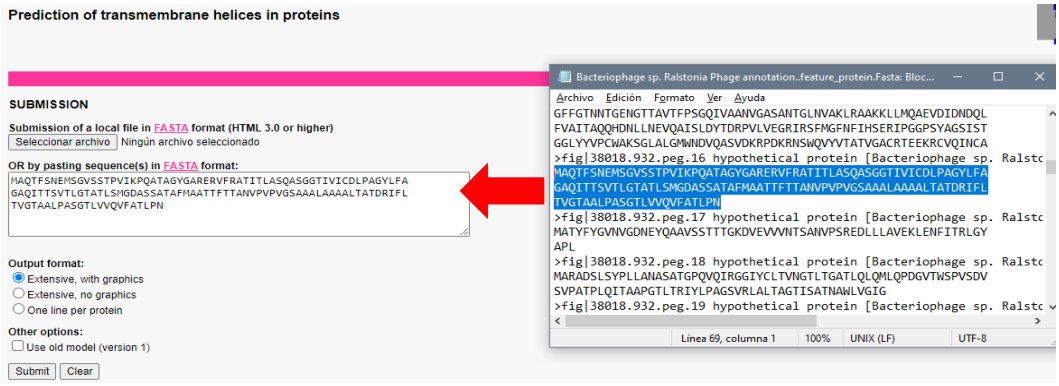
Entrada: Secuencias o archivos protein.Fasta

Instrucciones:

1. Copiar el siguiente link en su buscador de preferencia: http://www.cbs.dtu.dk/services/TMHMM/

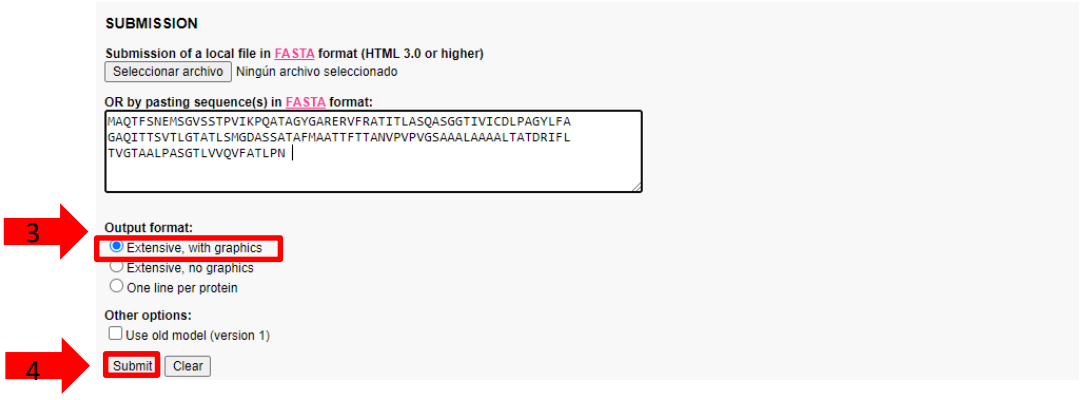
2. Ingresar la secuencia consulta en la plataforma. Para esto podemos cargar el archivo protein.Fasta que descargamos anteriormente en el apartado 6.1.1 “**Descarga de genoma anotado”**, o bien, seleccionar una única secuencia de este archivo y pegarla en el rectángulo blanco.





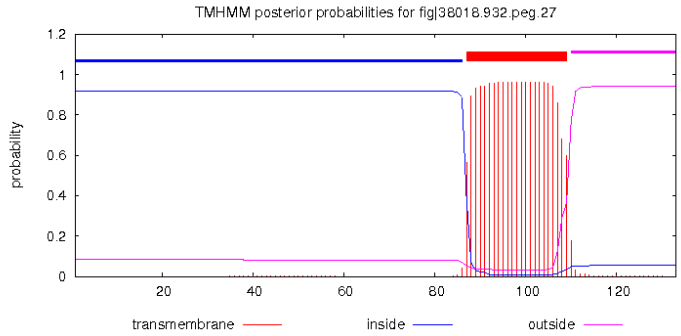
3. En la opción **“Output format”** seleccione la opción **“Extensive, with graphics”,** ya que se proporciona un poco más de información.

4. Presionar el botón **“Submit”** que se encuentra en la parte inferior izquierda, para dar inicio al procesamiento de los datos.



El resultado de este procedimiento muestra principalmente una tabla, donde se detallan algunas estadísticas de las secuencias predichas como proteínas de membrana o relacionados a ellas, se puede observar la longitud de la secuencia, el número de hélices transmembrana predichos, la ubicación de las proteínas transmembranas etc. Además, se adjunta una gráfica con las probabilidades de que las hélices predichas sean transmembrana, una hélice interior o exterior.

```
# fig|38018.932.peg.27 Length: 133
# fig|38018.932.peg.27 Number of predicted TMHs:  1
# fig|38018.932.peg.27 Exp number of AAs in TMHs: 21.09411
# fig|38018.932.peg.27 Exp number, first 60 AAs:  0.01545
# fig|38018.932.peg.27 Total prob of N-in:        0.91706
fig|38018.932.peg.27    TMHMM2.0        inside      1    86
fig|38018.932.peg.27    TMHMM2.0        TMhelix     87   109
fig|38018.932.peg.27    TMHMM2.0        outside     110  133
```



Opcional: Puede utilizar la herramienta SOSUI (Hirokawa *et al.,* 1998) para ampliar un poco más la información sobre la clasificación y las estructuras secundarias en las proteínas de membrana predichas por TMHHM**.** https://harrier.nagahama-i-bio.ac.jp/sosui/sosui_submit.html.

7. Comparación genómica

La genómica comparativa es la rama de la genética que estudia las diferencias y similitudes entre el genoma de diferentes especies, más o menos cercanas evolutivamente. Puede ayudar a determinar qué genes están relacionados con funciones concretas entre y dentro de las especies. Su objetivo es encontrar las huellas moleculares del proceso evolutivo que están escondidas en el genoma de las diferentes especies que conocemos. Comparar el genoma de dos especies diferentes puede ayudarnos a determinar su relación filogenética. Cuanto más próxima sea la divergencia de ambas especies a partir de una especie ancestral, más similares serán sus genomas, tanto en su composición como en la posición de sus genes. Este concepto se conoce como sintenia (Lien *et al.,* 2016).

7.1 Descarga de datos y generación de archivo de comparación

Previamente a realizar una comparación mediante cualquier herramienta es necesario tener los genomas de los organismos que queremos comparar, y generar sus respectivos archivos de comparación. A continuación, explicaremos paso a paso estos dos puntos.

7.1.1 Descarga de secuencias desde el NCBI

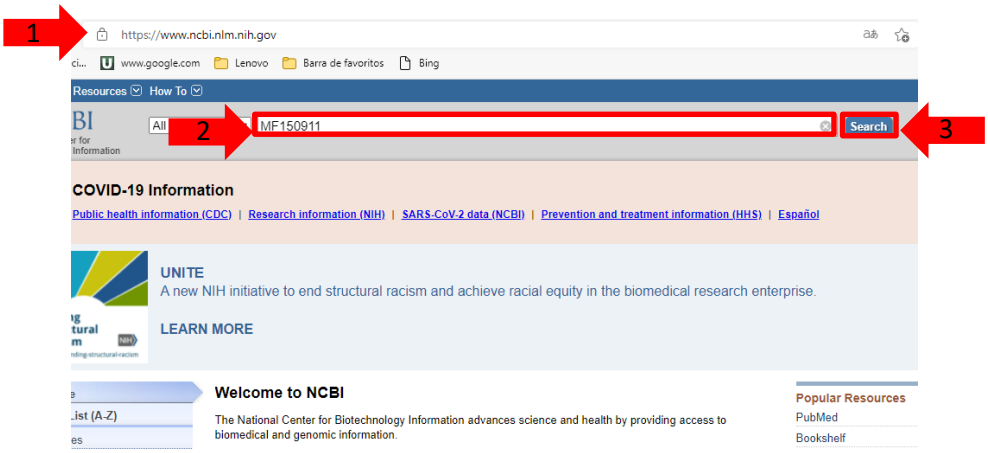
Sitio Web: https://www.ncbi.nlm.nih.gov/nuccore

Instrucciones:

1. Ingrese al sitio web: https://www.ncbi.nlm.nih.gov/nuccore

2. Pegar dentro de la barra "**Search"** de la base de datos el número de acceso del genoma que desea extraer, En este caso (MF150911) que corresponde al fago de ***Ralstonia solanacearum*** DU_RP_II.

3. Oprima sobre "**Search"** para buscar su secuencia.

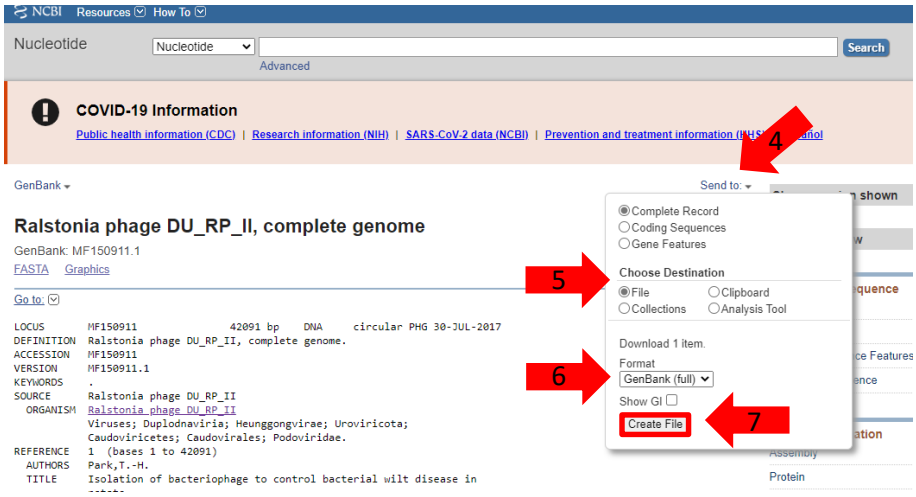


4. En la nueva pestaña damos clic sobre "**Send to".**

5. Dentro del apartado "**Choose Destination"** seleccionamos el círculo que corresponde a "**File".**

6. Damos clic sobre "**Format"** el cual muestra el formato GenBank (gb) como predeterminado. Al realizar esto se despliega un listado de los múltiples formatos bajos los cuales podemos descargar la secuencia. Seleccionamos el formato GenBank (full).

7. Damos clic sobre "**Create file"** e inmediatamente iniciara la descarga.



Aconsejamos guardar este archivo con el nombre de Ralstonia Phage DU_RP_II.gb dentro de la carpeta **"Tutorial".** Además**,** también deberemos descargar el genoma de este fago en formato .FASTA, para hacerlo deberemos seguir todo el procedimiento mencionado anteriormente, pero ahora en el paso 6 deberemos seleccionar formato FASTA en lugar de GenBank (full).

NOTA: Recuerde descargar el formato. FASTA y .gb del fago RpY1 con código de acceso (MN996301). Todos estos archivos serán útiles en el apartado de comparación genómica.

7.1.2 Generar archivos de comparación usando BLASTn

El archivo de comparación contiene información sobre qué tan similares o diferentes son las diversas regiones que conforman las secuencias que se pretende comparar. Esto se logra utilizando un sistema de puntuación generado a partir de la alineación de dichas secuencias. La generación del archivo de comparación es fundamental para utilizar la herramienta ACT, pues gracias a este se podrá visualizar las regiones compartidas o no entre los genomas, lo cual nos da una perspectiva más amplia sobre la similitud de las secuencias.

Nota: La forma en que se genera en este caso los archivos de comparación es solo una de las varias opciones que existen para originar estos datos. Si desea puede explorar algunas otras opciones como tBLASTx, MEGABLAST.

Sitio Web: https://blast.ncbi.nlm.nih.gov/Blast.cgi

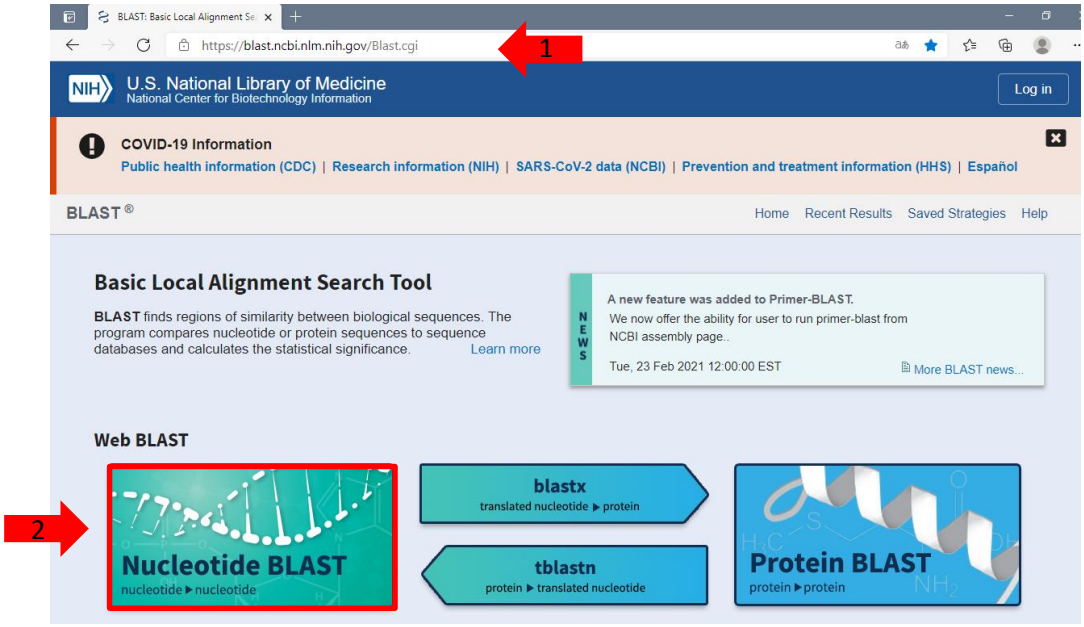
Entrada: Formato FASTA de las secuencias que se quieren comparar.

Instrucciones

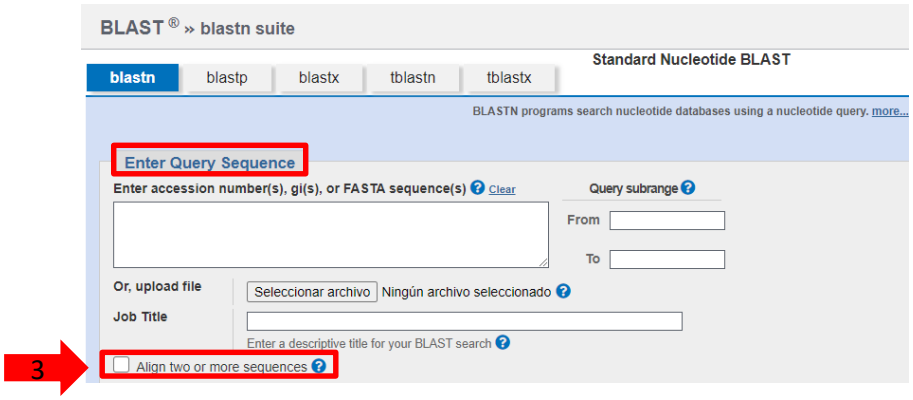
En este trabajo realizamos la comparación genómica con los fagos de Ralstonia DU_RP_II (MF150911) y el RpY1 (MN996301), los cuales mostraban buena similitud de secuencias con nuestro fago. Estos archivos fueron descargados anteriormente en el apartado 7.1.1 "**Descarga de datos y generación de archivo de comparación**"

1. Abrir una nueva ventana desde su navegador de preferencia e ingresar al sitio web del NCBI-BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

2. Seleccione la ventana **"Nucleotide BLAST".**



3. En la nueva ventana marque la opción "**Align two or more sequences",** que se encuentra dentro del apartado **"Enter query sequence".** Esto abrirá una nueva dentro de la cual podremos agregar otra secuencia.

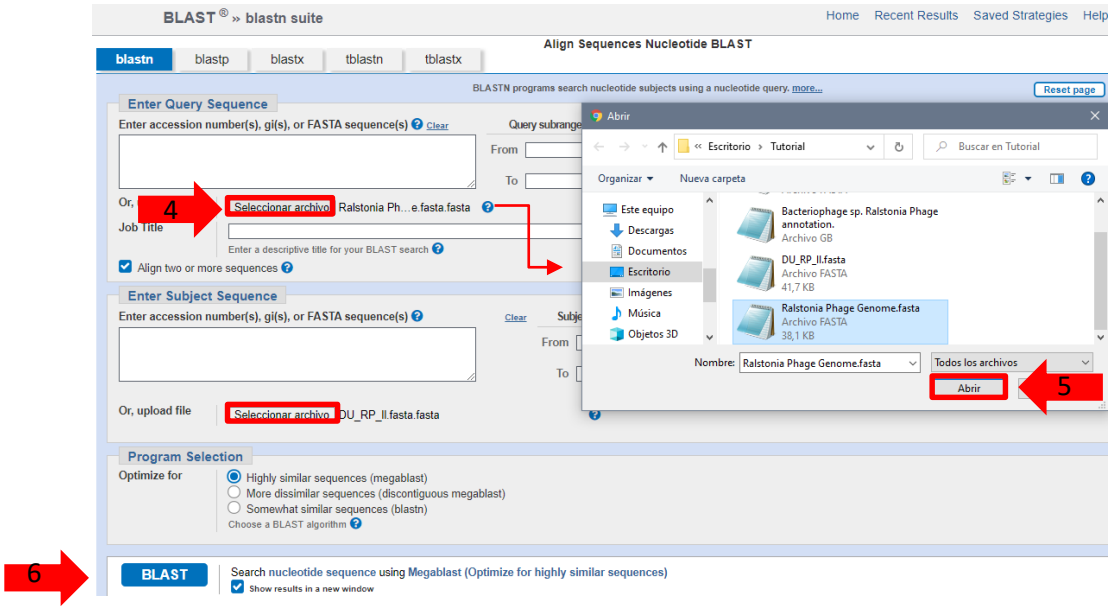


4. Presione el botón **“Seleccionar archivo”** y especifique el lugar en el que se encuentran las secuencias a partir de las cuales quiere generar el archivo de comparación. Recuerde que estas secuencias se encuentran dentro de la carpeta **“Tutorial”.**

5. Seleccione el archivo que desea cargar y presione **“Abrir”.** Sabremos que nuestra secuencia ha cargado cuando aparece el nombre de este al lado de la opción **“Seleccionar archivo”.**

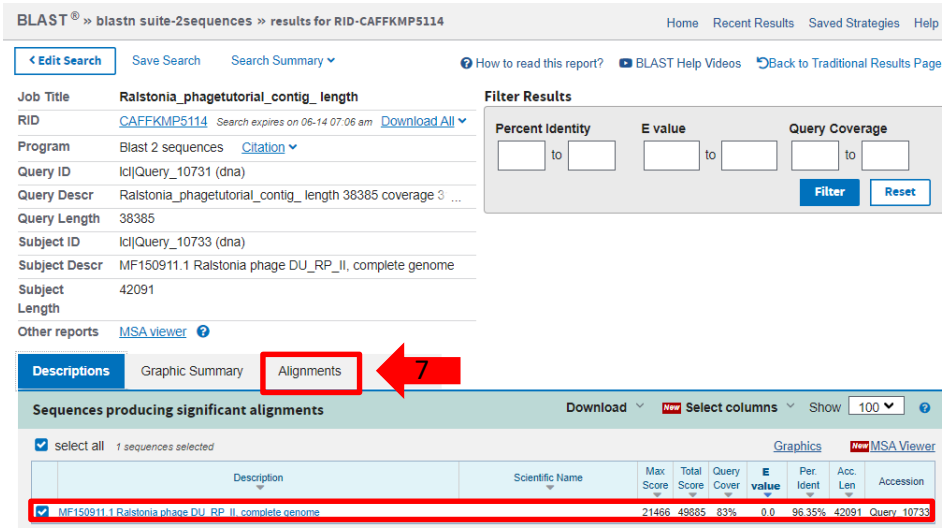
Debe realizar nuevamente los procedimientos de los pasos 3 y 4 en el otro rectángulo para cargar la secuencia contra la cual quiere realizar la comparación. En este caso generaremos un archivo de comparación a partir de las secuencias de los fagos NJ-P3, la cual guardamos previamente con el nombre **“Ralstonia Phage Genome.fasta”** y el fago DU_RP_II (MF150911) secuencia descargada desde el NCBI.

6. Presionar el botón BLAST para generar la comparación. Este proceso puede tardar algunos segundos.

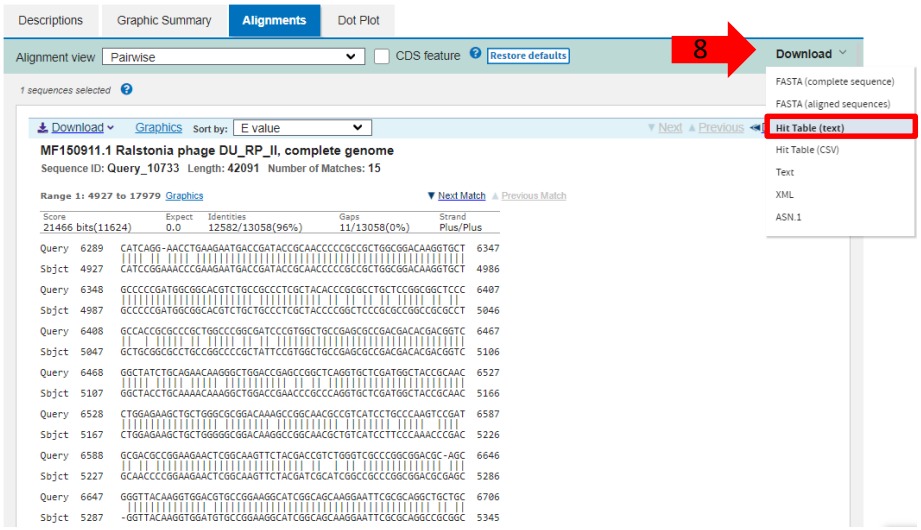


El anterior procedimiento nos llevará inicialmente a una nueva ventana la cual nos muestra información relacionada sobre el proceso que acabamos de llevar a cabo, además, nos presentan una breve descripción sobre el alineamiento realizado.

7. Seleccionar el recuadro “**Alignments”** para obtener mayor información sobre el alineamiento.



Ahora podemos visualizar los alineamientos que realizó BLASTn para las dos secuencias que ingresamos. Lo siguiente será descargar el archivo de comparación.



8. Dar clic sobre “**Download”,** que se encuentra en la parte superior derecha en la sección de “**Alignments”.** Allí seleccione la opción de descarga “**Hit Table (text)”.**

Por último, recomendamos guardar el archivo generado dentro de la carpeta de trabajo “**Tutorial”,** de igual modo debemos cambiar el nombre de este, en nuestro caso lo llamamos “**Ralstonia Phage Genome.fasta VS Ralstonia Phage DU_RP_II.fasta”**

Nota: Deberá generar nuevamente otro archivo de comparación, pero ahora con nuestro fago (NJ-P3) contra el fago RpY1 (MN996301). El archivo final deberá

llevar el nombre de "**Ralstonia Phage Genome.fasta VS Ralstonia Phage RpY1.fasta".**

7.2 Artemis Comparison Tool (ACT)

Artemis Comparison Tool (ACT) (Caevert *et al.,* 2005) es un visualizador gráfico el cual permite realizar comparaciones entre diferentes genomas y sus respectivas anotaciones. Debido a que este programa está basado en Artemis, se conservan algunas de sus funciones principales, entre estas, la fácil interacción y gran capacidad de navegación que el investigador tiene dentro de los genomas de interés, lo cual permite un mayor análisis de los datos. Además, los componentes gráficos no son tan complejos de entender y utilizar, lo que los hace muy informativos a la hora de distinguir entre regiones conservadas, inserciones y reordenamientos en cualquier nivel, desde el genoma hasta las diferencias de pares de bases.

Sitio Web: https://www.sanger.ac.uk/tool/artemis-comparison-tool-act/

Compatibilidad: ACT es compatible con sistemas operativos como Linux, Windows y Mac.

Entrada: Formatos EMBL, Genbank, GFF y FASTA

Para la correcta instalación de ACT aconsejamos instalar la versión más actualizada de Java, ya que este programa está escrito en este lenguaje de programación. En el siguiente link puede encontrar la información necesaria para descargar e instalar ACT en su equipo: http://sanger-pathogens.github.io/Artemis/ACT/.

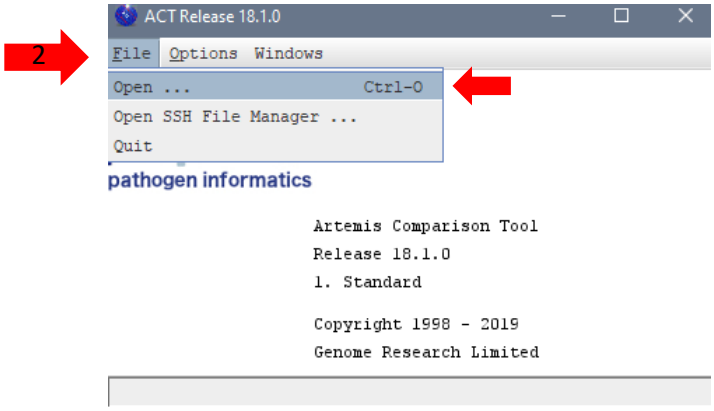
Si prefiere hacer uso del manual de ACT por favor ingrese a la siguiente página: https://sanger-pathogens.github.io/Artemis/ACT/act-manual.pdf

Instrucciones

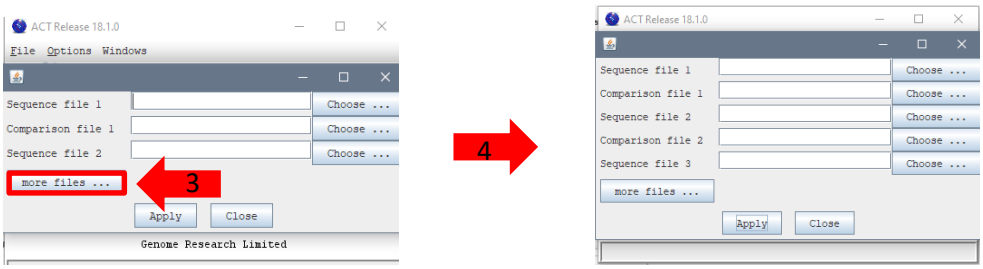
Una vez instalado el programa en su equipo, debe seguir los siguientes pasos para lograr una visualización de la comparación entre sus secuencias dentro de la ventana principal de ACT.

1. Ejecute ACT dando doble clic sobre el icono principal que aparece en el escritorio de su computadora.

2. Seleccione el menú **"File"** y luego la opción **"Open".** Se habilitará una nueva ventana, donde debemos cargar nuestras secuencias y sus respectivos archivos de comparación.



3. En la siguiente ventana dar clic en el botón "**More files"** para aumentar el número de archivos que se pueden cargar.



4. Ahora debemos cargar los datos seleccionando **"Choose local file"**. **"Sequence file"**, corresponde a las secuencias en formato .fasta de los genomas que descargamos desde el NCBI, por otro lado, **"Comparison file"**, hace referencia a los archivos de comparación que generamos y descargamos anteriormente desde BLASTn. Los archivos que necesitamos en esta ocasión se encuentran dentro la carpeta de trabajo **"Tutorial"**.

Nota: El orden en que se cargue los archivos es de suma importancia ya que en este mismo orden serán visualizados en ACT. Por tal motivo recomendamos cargar los archivos siguiendo el mismo patrón que aquí enseñamos:

Sequence 1: Ralstonia phage DU_RP_II.fasta

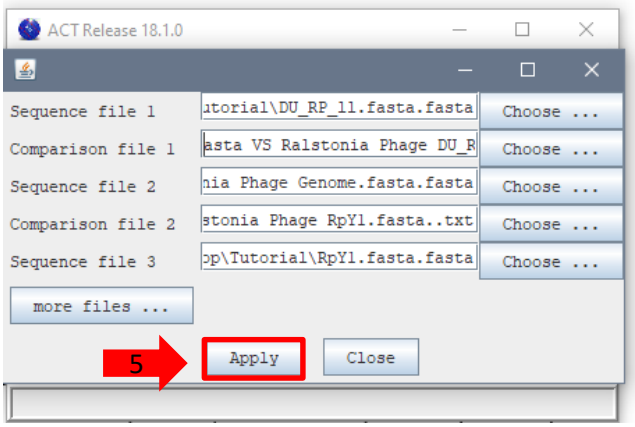
Comparison file 1: Ralstonia Phage Genome.fasta VS Ralstonia Phage DU_RP_II.fasta

Sequence 2: Ralstonia Phage Genome.fasta

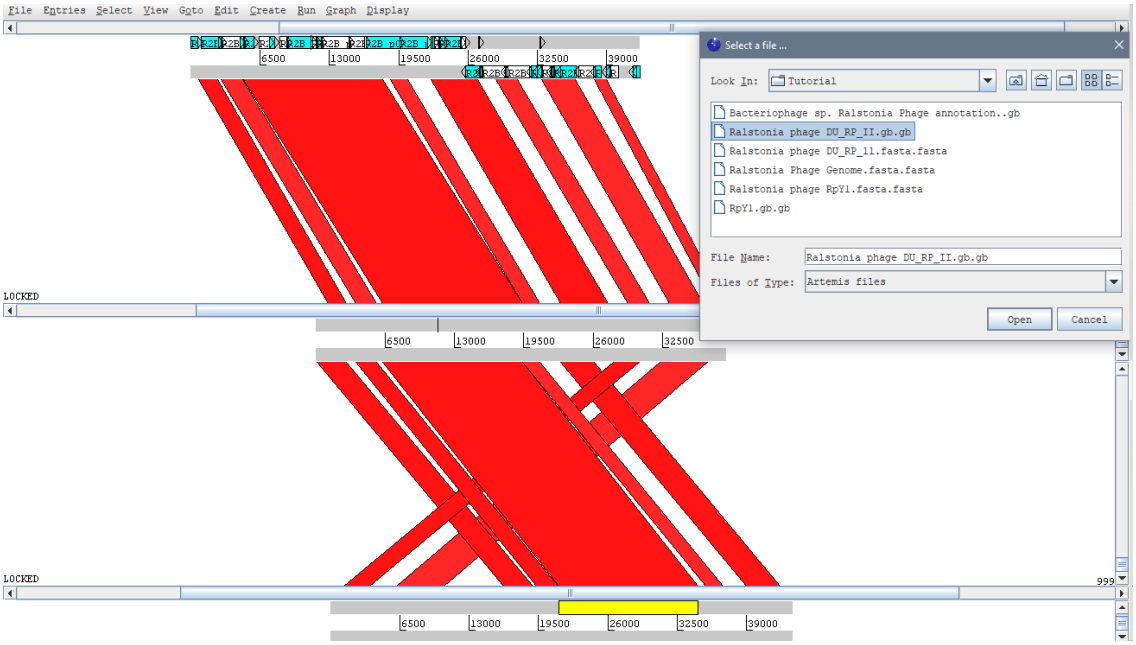
Comparison file 2: Ralstonia Phage Genome.fasta VS Ralstonia Phage RpY1.fasta.

Sequence 3: Ralstonia phage RpY1.fasta.

5. Una vez cargadas las secuencias oprimimos clic sobre el botón **“Apply”** para visualizar las comparaciones.



Si desea cargar las respectivas anotaciones para cada fago dirigase al menu **“File”**, elija el fago sobre el cual quiere cargar la anotación (Ralstonia phage DU_RP_ll. fasta en este caso), luego seleccione **“Read And Entry”** y cargue el archivo genbank (.gb) “**Ralstonia phage DU_RP_ll.gb”,** el cual contiene las características de la anotación para este fago. El archivo .gb se descargó anteriormente desde el NCBI. Esto se puede realizar para cada genoma cargado dentro de ACT.



En esta guía tutorial se concluye que existen diversos algoritmos y herramientas cuyo uso puede variar de acuerdo a las características genómicas del organismo objeto de estudio, la naturaleza de los datos y las preguntas que quiere solucionar el investigador. Es sumamente importante tener claro que existen algunos pasos que siempre deben de tenerse en cuenta cuando se realiza un análisis genómico. En esta guía se abarcan estos pasos, comprendidos como análisis de calidad de los datos, ensamblaje del genoma, anotación y por último la comparación genómica. Todo esto se llevó a cabo mediante el uso de herramientas web de libre acceso y de fácil manejo, con las cuales se ensambló y anotó el genoma del fago NJ-P3, presentó una longitud de 38.385 pb, con 49 CDS, de los cuales fue posible predecir funciones putativas para 25 genes codificantes. Este fago, además, mostró una alta similitud de secuencias con los fagos DU_RP_ll y RpY1.

El estudio de los genomas de bacteriófagos para fines investigativos es un tema que motiva a investigadores nuevos e experimentados en el tema. Muchos de estos investigadores no cuentan con los conocimientos básicos sobre análisis de genomas, puesto que en su gran mayoría este conocimiento se fundamenta en el uso programas de líneas de comando, que a su vez requieren de avanzadas habilidades en programación. Por tal motivo se plantea que trabajos como este (**Flujo de trabajo para el análisis de datos, ensamblaje, anotación y comparación de genomas de bacteriófagos**), son de gran importancia para promover e incentivar a personas sin habilidades en programación, facilitando el entendimiento de aspectos fundamentales para el estudio de genomas y aportando al conocimiento sobre las características genómicas de bacteriófagos en general.

8. Glosario

Adaptador: Secuencias de oligonucleótidos que se une a los extremos 5-3´de cada uno de los fragmentos que conforman la librería de secuenciación. Actúan en la amplificación y secuenciación de dichos fragmentos.

Anotación: Conjunto de Procesos mediante los cuales se identifican las diferentes estructuras genómicas dentro de un genoma y se lleva a cabo su posterior clasificación funcional, implementando principalmente diferentes bases de datos.

Archivo BAM: (*Binary Alignment/Map format*) es la versión comprimida del archivo SAM, donde se presentan los datos en código binario.

Archivo SAM: Archivos que contienen información relacionada a alineamientos de secuencias.

Contig: Longitud de secuencia continua *in silico* generada por alineamiento de lecturas de secuencias que se solapan.

Control de calidad: Conjunto de procedimientos realizados sobre datos generados a partir de NGS que sirve como un chequeo rápido para la identificación y posterior exclusión de datos con problemas de calidad.

Datos crudos: Datos sobres los cuales no se ha realizado ningún procedimiento previo.

Ensamblador: Herramienta bioinformática que funciona bajo algoritmos que permiten el ensamblaje de pequeñas secuencias en secuencias más largas y contiguas (contigs).

Formato.FASTA: Formato que contiene las secuencias en plano de ADN, ARN o aminoácidos.

Formato.FASTQ: Tipo de formato dentro del que se almacena la secuencia nucleotídica con sus respectivas puntaciones de calidad.

Formato.GB: Formato dentro del cual se muestran diversas características relacionadas con los datos y su respectiva anotación.

Forward o R1: Lecturas generadas en dirección 5-3´.

Gaps: Ausencia de uno o más nucleótidos en una de las hebras de ADN O ARN.

Genoma: Conjunto de genes que hacen parte de un organismo.

Genoma de referencia: Genoma altamente secuenciado y ensamblado que se utiliza como molde para alinear nuevas lecturas secuenciadas.

Indels: hace referencia a la “inserción o deleción" que se puede presentar dentro de una secuencia nucleotídica.

In silico: Término referido a procesos que se llevan a cabo mediante el uso de herramientas computacionales.

Lecturas (reads): Pequeñas secuencias obtenidas de tecnologías de secuenciación.

Next Generation Sequencing (NGS): Tecnologías diseñadas para secuenciar gran cantidad de datos de ADN y ARN de forma masiva. Se caracterizan principalmente por generar una gran cantidad de datos a un menor tiempo y costo por base.

Paired-end: Cada uno de los fragmentos de la librería es secuenciado por ambos extremos y al mismo tiempo.

Output: Archivos específicos obtenidos luego de llevar a cabo un proceso mediante el uso de una herramienta bioinformática.

Reverse o R2: Lecturas generadas en dirección 3´-5´.

Secuenciación: Conjunto de métodos y técnicas bioquímicas utilizadas con el fin de determinar el orden de los nucleótidos dentro de una muestra de ADN o ARN.

Single-end: Indica que cada fragmento de la librería se secuencia por un solo extremo.

Valor Pherd: Medida de la calidad en la identificación de los nucleótidos generados por la tecnología de secuenciación.

9. Bibliografía

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research, 46(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Andrews S.2010. FastQC: a quality control tool for high throughput sequence data. [Internet]. Fecha de acceso: 20 de marzo de 2021. Disponible en: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology: a journal of computational molecular cell biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. Bioinformatics (Oxford, England), 21(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nature biotechnology, 29(11), 987–991. https://doi.org/10.1038/nbt.2023

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in molecular biology (Clifton, N.J.)*, *1962*, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)*, *23*(6), 673–679. https://doi.org/10.1093/bioinformatics/btm009

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048.

Gordon, A. (2010). FASTQ/A short-reads pre-processing tools. [Internet]. Fecha de acceso: 20 de marzo de 2021. Disponible en: http://hannonlab.cshl.edu/fastx_toolkit/

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. Briefings in bioinformatics, 15(6), 879–889. https://doi.org/10.1093/bib/bbt069

Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics, 14(4), 378–379.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11, 119. https://doi.org/10.1186/1471-2105-11-119

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305(3), 567–580.

Langmead, Ben y Steven L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. En: Nature Methods 9.4, pág.357-359.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research*, *32*(1), 11–16. https://doi.org/10.1093/nar/gkh152

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15; 25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K., Olav Vik, J., … Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature, 533(7602), 200–205. https://doi.org/10.1038/nature17164

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), pp. 10-12. doi:https://doi.org/10.14806/ej.17.1.200

Mikheenko Alla, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich. (2018). Versatile genome assembly evaluation with QUAST-LG, Bioinformatics. 34 (13): i142-i150. doi: 10.1093 / bioinformatics / bty266

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. Nature reviews. Genetics, 14(3), 157–167. https://doi.org/10.1038/nrg3367

Rihtman, B., Meaden, S., Clokie, MR, Koskella, B. y Millard, AD (2016). Evaluación de la tecnología Illumina para la secuenciación de alto rendimiento de genomas de bacteriófagos. PeerJ, 4, e2055. https://doi.org/10.7717/peerj.2055

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualisation of de novo genome assemblies. Bioinformatics, 31(20), 3350-3352.

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of molecular biology, 430(15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007