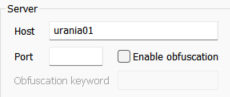
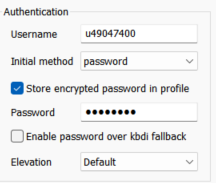
**Protocolo de obtención de alineamientos de péptidos a partir de reads**

**Paso 1: Instalación y conexión de Bitvise SSH Client**

En el apartado “Server” debemos rellenar el campo de host con “urania01”.



En el apartado “Authentication” debemos introducir nuestro usuario y contraseña.



Luego pulsamos el botón “Log in” para conectarnos.

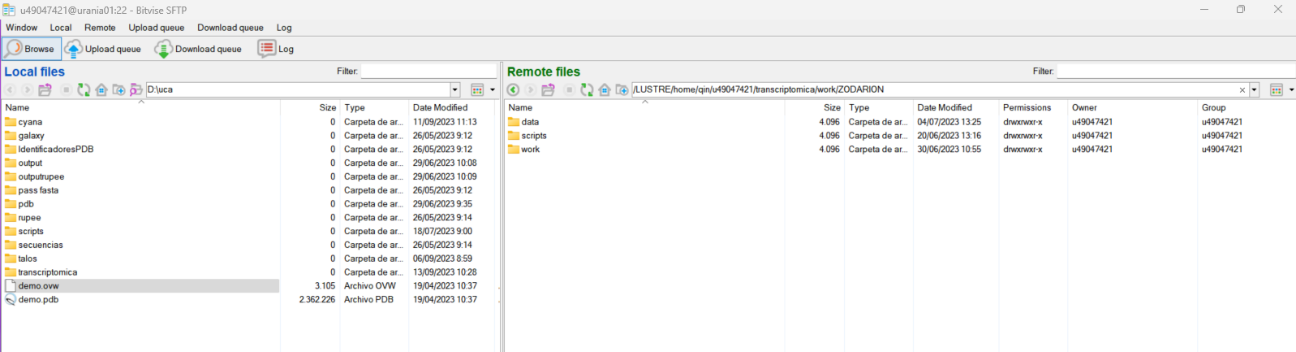
Una vez conectado, aparecen dos botones a la izquierda: “New terminal console” y “New SFTP window”.

Desde la consola podremos navegar a través del sistema de ficheros del clúster utilizando los comandos de Linux.

Texto

Descripción generada automáticamente

Una opción más cómoda es utilizar el servicio SFTP, donde a la izquierda tendremos el sistema de archivos de nuestro PC y a la derecha el del clúster. Podremos navegar a través de ellos y arrastrar archivos de un lado a otro para copiarlos, crear carpetas, etc.



Podemos modificar directamente los ficheros desde aquí abriéndolos con un editor de texto (Notepad++) o un editor de código (Visual Studio Code).

Para ejecutar los scripts debemos hacerlo desde la consola.

Utilizamos el comando “cd nombreDelDirectorio” para acceder a un directorio. Con el comando “cd ..” volvemos atrás.

Visualizamos el contenido del directorio con el comando “ls”.

Ejecutamos el script mediante el comando “sbatch nombreDelScript”.

Para comprobar el estado de la ejecución, utilizamos el comando “squeue”. Aparecerá una lista de los trabajos que se están ejecutando o están pendientes de hacerlo. Cada trabajo tiene asociado un ID.

Para cancelar la ejecución de un script debemos usar el comando “scancel” seguido del ID correspondiente.

**Paso 2: Preparación del directorio de trabajo**

Abre tanto la ventana SFTP como la consola, ya que se trabajará desde ambos lados.

Dentro de tu directorio personal del clúster, crea un directorio llamado “transcriptomica” si aún no lo tienes:

mkdir transcriptomica (También puedes crear carpetas desde la ventana SFTP)

Entramos en él:

cd transcriptomica

Copia el directorio “template” con el siguiente comando:

cp -r /LUSTRE/home/qin/u49047421/transcriptomica/template .

Nota: el punto “.” final de la línea anterior sirve para indicar el directorio actual, es decir, a donde se copiará “template”.

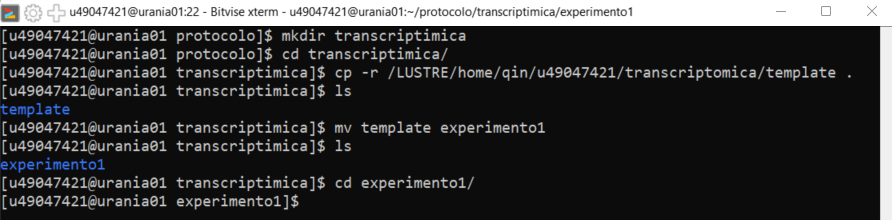
Este directorio es una plantilla de trabajo preparada para ejecutar todos los scripts.

Cambia el nombre del directorio por otro más adecuado para tu experimento:

mv template experimento1

Entramos dentro:

cd experimento1

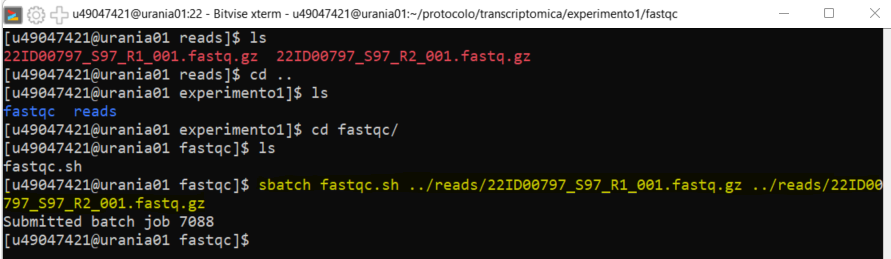


Copia al directorio “reads” los dos ficheros de partida procedentes de la secuenciación. Estos ficheros tienen un formato fastq, aunque probablemente estén comprimidos. No es necesario descomprimirlos.

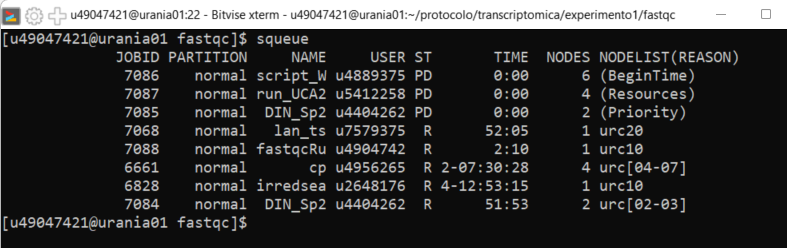


**Paso 3: Análisis de calidad con FastQC**

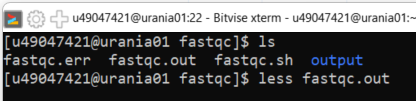
Entra en el directorio “fastqc” y ejecuta el script con el comando “sbatch” pasando como argumentos los dos ficheros de partida.

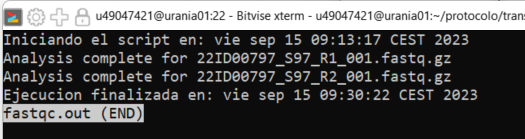


Con el comando “squeue” puedes comprobar el estado de la ejecución del script desde la cola de trabajo del clúster.



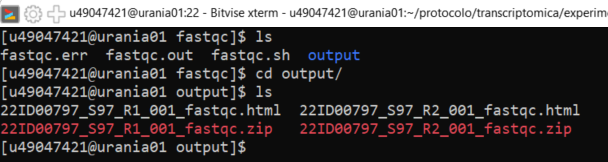
Para comprobar si la ejecución ha terminado, además de consultar la cola de trabajo del clúster, puedes visualizar con el comando “less” el fichero “fastqc.out” que se ha generado.



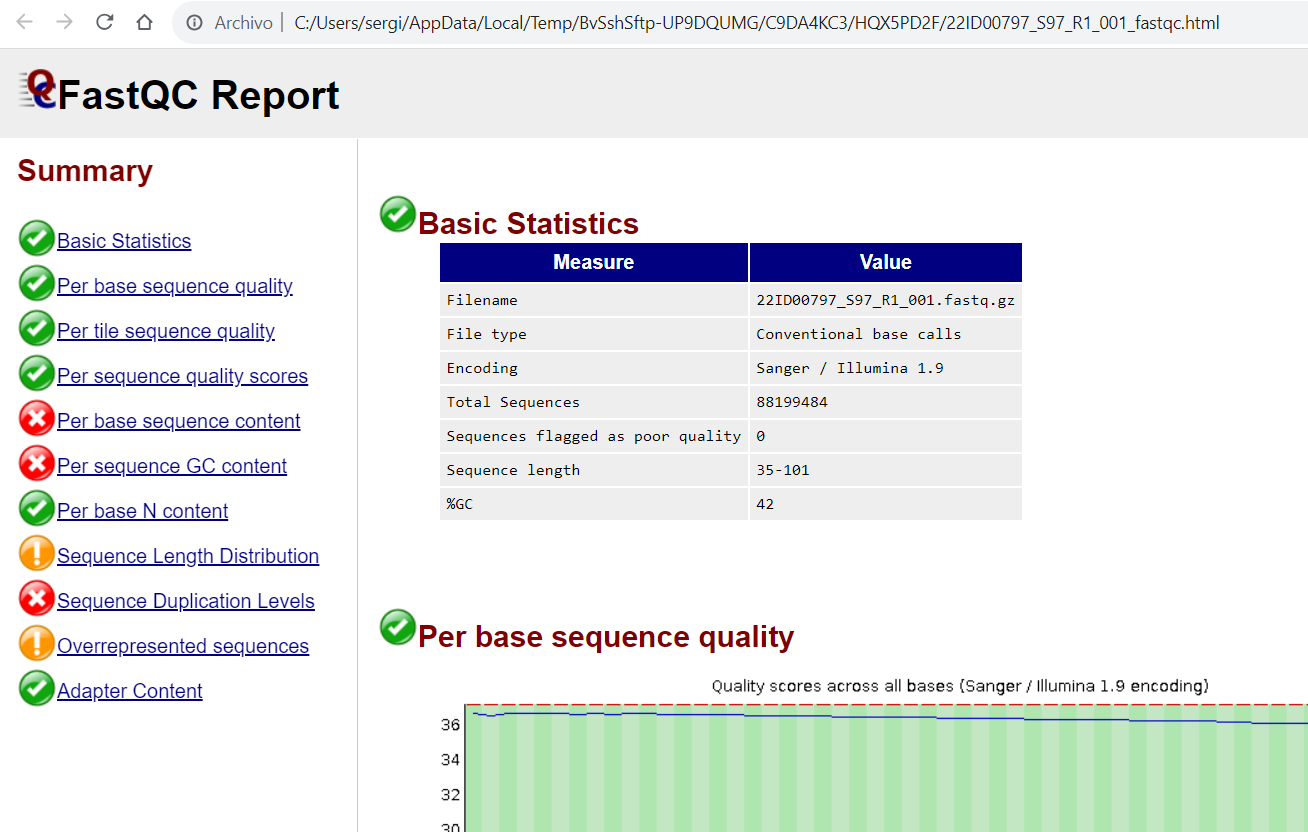


Para salir pulsa la tecla q.

En el directorio “output” generado aparecerán los ficheros con el análisis de calidad de los reads.



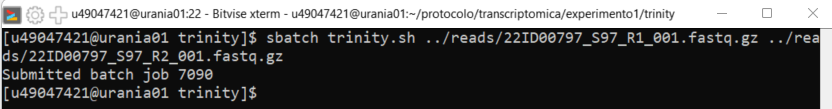
Puedes descargarlos desde la ventana SFTP o abrirlos directamente.



**Paso 4: Ensamblaje con Trinity**

Una vez comprobada la calidad de los reads, el siguiente paso es el ensamblaje para obtener un fichero fasta con todas las secuencias.

Entramos en el directorio “trinity” y ejecutamos con el comando “sbatch” el script “trinity.sh” pasándole como argumentos los dos ficheros de partida del directorio “reads”.



Se generarán los ficheros “trinity.err” y “trinity.out” que podrás visualizar para hacer un seguimiento de la ejecución.

Al finalizar la ejecución satisfactoriamente se creará un fichero con los resultados llamado “trinity\_out\_dir.Trinity.fasta”.

**Paso 5: Control de calidad del ensamblaje con BUSCO**

Para ejecutar BUSCO desde un ambiente offline necesitamos una base de datos de linaje como referencia. En nuestro caso: metazoa\_odb10.

Puedes descargarla manualmente desde <https://busco-data.ezlab.org/v5/data/lineages/> y guardarla en un nuevo directorio dentro de “transcriptomica”.

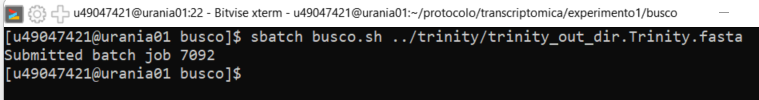
Por ejemplo, en “transcriptomica/data/BUSCO\_DB/ metazoa\_odb10”

Este fichero normalmente será común para todos los experimentos. Existe una copia en:

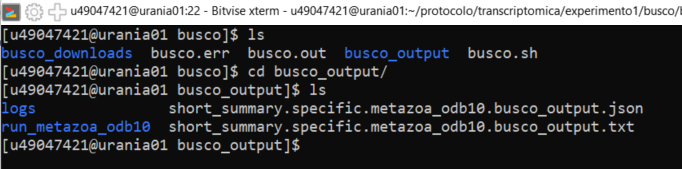
/LUSTRE/home/qin/u49047421/transcriptomica/data/BUSCO\_DB/metazoa\_odb10

Esta copia ya está referenciada dentro del script “busco.sh”. En caso de querer utilizar otra, es necesario modificar el script.

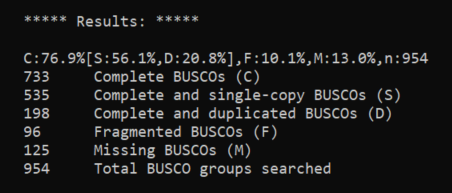
Accede al directorio “busco” y dentro de él ejecuta con “sbatch” el script “busco.sh” pasando como argumento el fichero fasta obtenido con Trinity.



Tras la ejecución se creará un directorio llamado “busco\_output”.



Dentro de él podemos consultar el *short summary* con el comando “less” para visualizar los resultados.



**Paso 6: BLASTX**

El siguiente paso es utilizar Blastx para comparar las secuencias con una base de datos de péptidos.

Si se trata de una nueva base de datos, primero necesitamos construirla. Para ello necesitamos guardar el fichero fasta con la base de datos en un directorio común para todos los experimentos.

Por ejemplo, si vamos a utilizar el fichero “CTX-May23bis\_completeseq.fasta” como base de datos. Podemos guardarlo en:

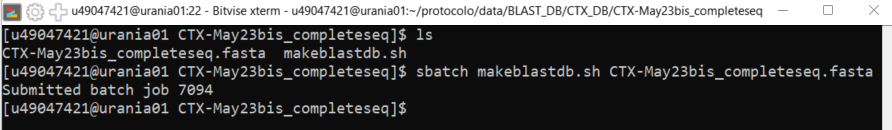
transcriptomica/data/BLAST\_DB/CTX\_DB/CTX-May23bis\_completeseq/ CTX-May23bis\_completeseq.fasta

Para construir la base de datos copia el script “makeblastdb.sh” que se encuentra en el directorio “blastx” al directorio donde se encuentra el fichero fasta.

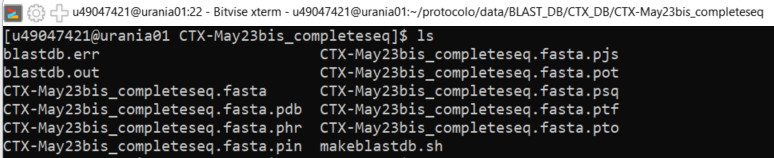
Imagen que contiene Texto

Descripción generada automáticamente

Una vez copiado y situado dentro del directorio, ejecuta el script con “sbatch” pasándole el fichero fasta como argumento.



Se generarán diferentes ficheros con los que Blastx trabajará automáticamente.



Una vez que ya disponemos de una base de datos construida, podemos entrar en el directorio “blastx” y ejecutar el script “blastx.sh” con el comando “sbatch”. Es necesario pasarle 2 argumentos en el orden correcto.

1. QUERY: Fichero fasta obtenido en Trinity.
2. DB: Fichero fasta de la base de datos de péptidos previamente construida. La base de datos mencionada anteriormente ya se encuentra construida en: /LUSTRE/home/qin/u49047421/transcriptomica/data/BLAST\_DB/CTX\_DB/CTX-May23bis\_completeseq/CTX-May23bis\_completeseq.fasta

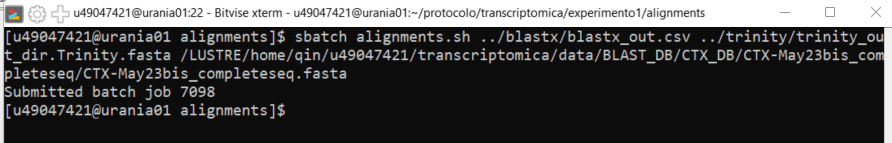


Tras la ejecución se generará un fichero “blastx\_out.csv”.

**Paso 7: Extracción de alineamientos**

Accede al directorio “alignments” y ejecuta con “sbatch” el script “alignments.sh”. Es necesario pasarle 3 argumentos en el orden correcto.

1. BLAST\_CSV: Fichero csv obtenido en Blastx.
2. TRINITY\_FASTA: Fichero fasta obtenido en Trinity.
3. DB: Fichero fasta de la base de datos de péptidos previamente construida.



Para verificar que la ejecución ha finalizado correctamente consulta el fichero “alignments.out”.

Se creará un directorio llamado “Alineamientos\_mafft” en donde se encontrarán todos los ficheros con la extensión “mafft.fasta” que contendrán los alineamientos detectados.

También puede ser útil el fichero “extracted\_sequences.fasta”.

Descarga los resultados desde la ventana SFTP a tu PC para continuar trabajando con los alineamientos detectados.

**Paso 8: Filtrado de los alineamientos**

Accede al directorio “curation\_filter”. Antes de ejecutar el script es necesario revisar los parámetros de configuración del filtro. Éstos se encuentran en el fichero “config.txt” dentro del directorio “python\_scripts”. Para modificar la configuración puedes editar el fichero o sustituirlo por otro con el mismo nombre y formato.

Una vez revisado los parámetros, ejecuta con “sbatch” el script “filter.sh”, no es necesario indicar ningún argumento. La ejecución puede durar alrededor de 20 minutos. Tras finalizar accede al directorio “Alineamientos\_filtrados”.

Dentro de este directorio se encuentran todos los alineamientos que han pasado el filtro y un fichero llamado “informe.tsv”. Puedes descargar este fichero para consultar toda la información del filtrado.

El fichero contiene las siguientes columnas:

* **Seq\_file**: El número del fichero original mafft.fasta analizado.
* **Frame\_ID**: El identificador del marco de lectura analizado.
* **Ref\_ID**: El identificador de la secuencia de la base de datos con la que se ha intentado alinear el marco de lectura.
* **Pasa\_filtro**: Indica si el alineamiento entre las dos secuencias comparadas ha pasado el filtro.
* **Hay\_segmento\_de\_subsecuencias\_validas**: Criterio de filtrado 1 (debe ser “True”). Hace referencia a que existen partea de las dos secuencias que alinean correctamente.
* **longitud\_minima\_total\_subseqs\_superada**: Criterio de filtrado 2 (debe ser “True”). Hace referencia a la cantidad mínima de aminoácidos que deben alinearse.
* **ratio\_minimo\_longitud\_superado**: Criterio de filtrado 3 (debe ser “True”). Hace referencia al porcentaje mínimo que la parte alineada debe abarcar en la secuencia de la base de datos.
* **Stop\_codon\_en\_mitad\_de\_dos\_segmentos\_subsecuencias\_validas**: Criterio de filtrado 4 (debe ser “False”). Hace referencia a la presencia de un codón de parada en mitad de la secuencia analizada.
* **Vector\_alineamiento**: Representación binaria del alineamiento entre las dos secuencias: 1(aa alineado), 0 (aa no alineado), \* (codón de parada).