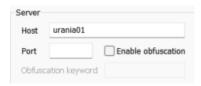
Protocolo de obtención de alineamientos de péptidos a partir de reads

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.



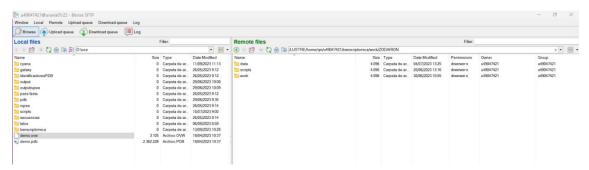
Luego pulsamos el botón "Log in" para conectarnos.

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.



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 "1s".

Ejecutamos el script mediante el comando "sbαtch 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/transcriptofilter/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 "sbαtch" pasando como argumentos los dos ficheros de partida.

```
[u49047421@urania01 fastqc]$ ls
fastqc.sh
[u49047421@urania01 fastqc]$ sbatch fastqc.sh ../reads/22ID00797_S97_R1_001.fastq.gz ../reads/22ID00
797_S97_R2_001.fastq.gz
Submitted batch job 7088
[u49047421@urania01 fastqc]$
```

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

```
🗾 👸 占 u49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~/protocolo/transcriptomica/experimento1/fastqc
u49047421@urania01 fastqc]$ squeue
             JOBID PARTITION
                                 NAME
                                           USER ST
                                                          TIME NODES NODELIST(REASON)
                      normal script_W u4889375 PD
                                                         0:00
                                                                   6 (BeginTime)
                                                                    4 (Resources)
              7087
                      normal run_UCA2 u5412258 PD
                                                         0:00
                      normal DIN_Sp2 u4404262 PD
              7085
                                                         0:00
                                                                    2 (Priority)
                               lan_ts u7579375 R
                                                         52:05
              7068
                      normal
                                                                    1 urc20
              7088
                      normal fastqcRu u4904742 R
                                                         2:10
                                                                    1 urc10
                                   cp u4956265 R 2-07:30:28
              6661
                      normal
                                                                    4 urc[04-07]
                      normal irredsea u2648176 R 4-12:53:15
              6828
                                                                    1 urc10
              7084
                      normal DIN_Sp2 u4404262
                                                 R
                                                         51:53
                                                                    2 urc[02-03]
[u49047421@urania01 fastqc]$
```

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.

```
□ □ □ u49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~

[u49047421@urania01 fastqc]$ ls

fastqc.err fastqc.out fastqc.sh output

[u49047421@urania01 fastqc]$ less fastqc.out

□ □ □ □ u49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~/protocolo/trans

Iniciando el script en: vie sep 15 09:13:17 CEST 2023

Analysis complete for 22ID00797_S97_R1_001.fastq.gz

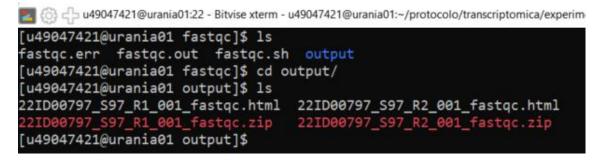
Analysis complete for 22ID00797_S97_R2_001.fastq.gz

Ejecucion finalizada en: vie sep 15 09:30:22 CEST 2023

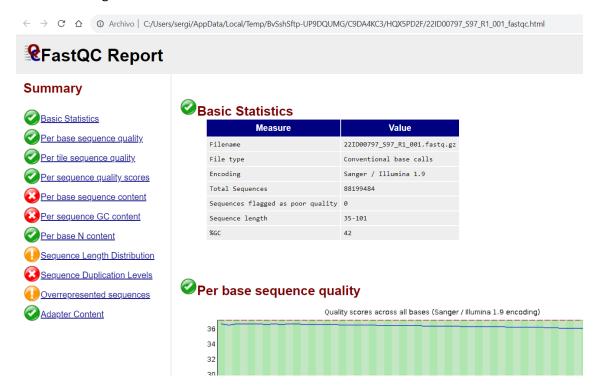
fastqc.out (END)
```

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/gin/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.

```
[u49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~/protocolo/transcriptomica/experimento1/busco —
[u49047421@urania01 busco]$ sbatch busco.sh ../trinity/trinity_out_dir.Trinity.fasta
Submitted batch job 7092
[u49047421@urania01 busco]$
```

Tras la ejecución se creará un directorio llamado "busco output".

```
wide description of the summary of the summar
```

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.

```
☑ ﴿ ← u49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~/protocolo/transcriptomica/experimento1/blastx

[u49047421@urania01 blastx]$ ls
blastx.sh makeblastdb.sh
[u49047421@urania01 blastx]$ cp makeblastdb.sh ../../../data/BLAST_DB/CTX_DB/CTX-May23bis_completese
q/
[u49047421@urania01 blastx]$
```

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

```
W49047421@urania01:22 - Bitvise xterm - u49047421@urania01:~/protocolo/data/BLAST_DB/CTX_DB/CTX-May23bis_completeseq — 
[u49047421@urania01 CTX-May23bis_completeseq]$ ls
CTX-May23bis_completeseq.fasta makeblastdb.sh
[u49047421@urania01 CTX-May23bis_completeseq]$ sbatch makeblastdb.sh CTX-May23bis_completeseq.fasta
Submitted batch job 7094
[u49047421@urania01 CTX-May23bis_completeseq]$
```

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 "sbαtch". Es necesario pasarle 2 argumentos en el orden correcto.

- 1. QUERY: Fichero fasta obtenido en Trinity.
- 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

```
Without the state of the s
```

Tras la ejecución se generará un fichero "blastx_out.csv".

Paso 7: Extracción de alineamientos

Accede al directorio "alignments" y ejecuta con "sbαtch" 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).

Paso 9: Filtro de Metionina y clasificación de alineamientos

Accede al directorio "metionine filter" y ejecuta con "sbαtch" el script "filter.sh"

Una vez finalizada la ejecución, en el directorio "resultados" se encontrarán los alineamientos clasificados.

- Alineamientos_Perfectos: La metionina inicial en las secuencias de referencia coincide con la metionina inicial del frame.
- Alineamientos_Limpios: Se trata de los alineamientos perfectos ya formateados sin guiones, comenzando por la metionina inicial y finalizando en el codón de stop correspondiente.
- Alineamientos_M_Previa: La metionina inicial en las secuencias de referencia coincide con la metionina inicial del frame, pero existe una metionina anterior que podría ser el verdadero inicio de la secuencia.
- Alineamientos_Multiframe: Son los ficheros en donde más de un frame alinea con las secuencias de referencia.
- Alineamientos_Revision_Manual: Necesitan ser revisados manualmente.

Paso 10: Clasificación por superfamilias de conotoxinas y otros péptidos

Accede al directorio "superfamily" y ejecuta con "sbatch" el script "SF-filter.sh". Es necesario pasarle 3 argumentos en el orden correcto.

- 1. DB1: Fichero fasta de la base de datos de péptidos señales de conotoxinas.
- 2. DB2: Fichero fasta de la base de datos de superfamilias de conotoxinas.
- 3. DB3: Fichero fasta de la base de datos de péptidos que hemos usado en el paso 6.

[u49047421@urania01 superfamily]\$ sbatch SF-filter.sh /LUSTRE/home/qin/u49047421/transcriptomica/dat
a/BLAST_DB/CTX_DB/CTX_SF_signal/CTX_SF_signal.fasta /LUSTRE/home/qin/u49047421/transcriptomica/data/
BLAST_DB/CTX_DB/conotoxinas-virroDB/conotoxinas-virroDB.fasta /LUSTRE/home/qin/u49047421/transcripto
mica/data/BLAST_DB/CTX_DB/CTX-Oct23_cleanseq/CTX-Oct23_cleanseq.fasta

Una vez finalizada la ejecución tendremos principalmente dos directorios de interés.

En el directorio "resultados" tendremos todos los ficheros csv con las secuencias clasificadas. Debemos descargarnos este directorio y ejecutar en nuestro PC la herramienta "csvToExcel.exe".

Se abrirá una ventana para seleccionar un directorio y deberemos seleccionar el que hemos descargado.

Una vez ejecutado, obtendremos un fichero Excel en ese mismo directorio donde estarán el resumen de los resultados en diferentes hojas.

- Perfectos_SF: Lista de secuencias (del directorio "Alineamientos_Perfectos") en donde se ha identificado un péptido señal de la base de datos de péptidos señal y posteriormente ha alineado con la base de datos de superfamilias. En la columna "SF_signal" se indica al péptido señal con el que alinea con un umbral del 70% de coincidencia (pident_signal). En la columna "match" nos aparece el mejor resultado y en la columna "SF-hormone" la superfamilia a la que pertenece.
- Perfectos_NoSF: Lista de secuencias (del directorio "Alineamientos_Perfectos") en donde no se ha identificado un péptido señal de la base de datos de péptidos señal pero posteriormente ha alineado con la base de datos de superfamilias.
- Perfectos_noMatch: Lista de secuencias (del directorio "Alineamientos_Perfectos") que no han alineado con ninguna secuencia de la base de datos de superfamilias.
- Mprevia_SF: Lista de secuencias (del directorio "Alineamientos_M_Previa") en donde se ha identificado un péptido señal de la base de datos de péptidos señal y posteriormente ha alineado con la base de datos de superfamilias.
- Mprevia_NoSF: Lista de secuencias (del directorio "Alineamientos_M_Previa") en donde no se ha identificado un péptido señal de la base de datos de péptidos señal pero posteriormente ha alineado con la base de datos de superfamilias.
- Mprevia_noMatch: Lista de secuencias (del directorio "Alineamientos_M_Previa") que no han alineado con ninguna secuencia de la base de datos de superfamilias.
- Revision_manual_Match: Lista de secuencias (del directorio "Alineamientos_Revision_Manual") que han alineado con la base de datos de superfamilias.
- Revision_manual_noMatch: Lista de secuencias (del directorio "Alineamientos_Revision_Manual") que no han alineado con la base de datos de superfamilias.

El directorio "alineamientos_NoSF" contiene los alineamientos en formato "mafft.fasta" de las secuencias de "Perfectos_NoSF", "Mprevia_NoSF" y "Revision_manual_Match" frente a todas las secuencias de la base de datos de superfamilias con las que ha habido un match. En estos alineamientos se puede comprobar en la descripción de cada secuencia a qué superfamilia pertenecen.

Tras la ejecución de este script también podemos realizar la misma comprobación de las descripciones en los alineamientos "mafft.fasta" del directorio "metionine_filter" del paso 9.

Esta revisión será útil para las secuencias que no han obtenido un match con la base de datos de superfamilias (noMatch) pero sí lo obtuvieron con la base de datos utilizada en el paso 6.

Las secuencias del directorio "Alineamientos_Multiframe" del paso 9 deberán ser revisadas siempre, ya que no son procesadas en el paso 10.

Modo automático: Ejecución de autolauncher

Todo el flujo de trabajo descrito anteriormente se puede realizar de forma automática. Existen dos puntos de partida:

- Punto de partida inicial: Indicando los dos reads y la base de datos de referencia.
- Punto de partida después del ensamblaje: Indicando el fichero fasta ensamblado con Trinity y la base de datos de referencia.

Se realizan los pasos 1, 2 y 3 descritos anteriormente para preparar nuestro directorio de trabajo con los dos reads.

Si disponemos del fichero fasta ensamblado debemos copiarlo al directorio "trinity".

Desde el directorio de trabajo ejecutamos "sbatch autolauncher.sh" seguido de las opciones con los argumentos necesarios.

Las opciones son las siguientes:

- --r1 ruta/del/read1
- --r2 ruta/del/read2
- --db ruta/de/la/base/de/datos
- --trinity ruta/del/fichero/ensamblado
- --db2 ruta/de/la/base/de/datos/de/péptidos/señal
- --db3 ruta/de/la/base/de/datos/de/superfamilias

Las opciones db2 y db3 son opcionales. Debemos incluirlos si queremos que se haga la clasificación por superfamilias. En ese caso, también debemos descargarnos el directorio de resultados dentro de "superfamily" y ejecutar en nuestro PC la herramienta "csvToExcel.exe".

En caso contrario, el filtro acabará con la ejecución del paso 9 y deberemos revisarlos alineamientos obtenidos en "metionine_filter".

Ejemplos de uso:

@urania01 imperialis_optimizado]\$ sbatch autolauncher.sh --r1 reads/SRR2609542_1.fastq.gz --r2 reads/SRR2609542_2.fastq.gz --db /LUSTRE/home/qin/u49047421/transcriptomica/data/BLAST_DB/CTX_DB/CTX_DB/CTX-Oct23_cleanseq/CTX-Oct23_cleanseq.fasta

[u49047421@urania01 pruebaAuto_SF_11-4-24]\$ sbatch autolauncher.sh --db /LUSTRE/home/qin/u49047421/transcriptomica/data/BLAST_DB/CTX_DB/CTX-Oct23_cleanseq/CTX-Oct23_cleanseq.fasta --trinity ./trinity/trinity_out_dir.Trinity.fasta --db2 /LUSTRE/home/qin/u49047421/transcriptomica/data/BLAST_DB/CTX_DB/CTX_SF_signal/CTX_SF_signal.fasta --db3 /LUSTRE/home/qin/u49047421/transcriptomica/data/BLAST_DB/CTX_DB/CTX_DB/conotoxinas-virroDB/conotoxinas-virroDB.fasta

Los ficheros "t-filter.err" y "t-filter.out" contienen detalles de la ejecución del script.