

>X_BU-ISCIII

MANUAL ENVÍO DE SECUENCIAS Y METADATA ASOCIADO A LA PLATAFORMA RELECOV

WP6 RELECOV 2.0



Título del Proyecto (Grant Agreement number)	Consolidation of WGS and RT-PCR activities for SARS-CoV-2 in Spain towards sustainable use and integration of enhanced infrastructure and processes in the RELECOV Network (101113109)
Acrónimo Proyecto	RELECOV 2.0
Convocatoria	EU4H-2022-DGA-MS-IBA-1
Organismo Financiador	European Health and Digital Executive Agency (HaDEA)
Work Package. Nº - Título	WP6 - Relecov Platform: Data management, analysis, visualisation and sharing.
Work Package Leader	Isabel Cuesta de la Plaza
Fecha	21/11/2023
Autores	Pablo Mata Aroco - Emi Arjona Bolaños
Revisores	Sara Monzón Fernández, Isabel Cuesta de la Plaza

Registro de Versiones

Fecha	Versión	Autor	Descripción
21/11/2023	v.0	Pablo Mata Aroco	Manual envío de metadata y secuencias a la plataforma RELECOV
07/02/2024	v.1	Pablo Mata Aroco	Modificación de datos obligatorios en el archivo de metadata
01/10/2024	v.2	Pablo Mata Aroco	Subida al SFTP de archivos BAM secuenciados desde Ion Torrent
06/05/2025	v.3	Pablo Mata Aroco	Adaptación al nuevo documento de metadata v3.0.0

ÍNDICE

01 CONTEXTO	1
02 COMPARTIR LOS METADATOS	1
Opción 1 - Plantilla excel	1
Opción 2 - Relecov Tools	3
03 REPOSITORIO SFTP	3
Estructura de carpetas dentro del SFTP	3
Preparación de los datos	4
Comprobación de integridad de los ficheros	4
Opción 1. Sistema Operativo Windows	5
Opción 1.a línea de comandos de Windows	5
Opción 1.b Utilizando software externo (Fsum FrontEnd)	5
Opción 2. Sistema Operativo Linux	6
Subida de ficheros al recurso sftp	6
Manual para subir los datos usando filezilla	6
Bajada de ficheros BAM Secuenciación con Ion Torrent	8
04 FAQs	9

01 CONTEXTO

El 22 de enero de 2021, la Ponencia de Alertas y Planes de Preparación y Respuesta, junto con la Comisión de Salud Pública del Consejo Interterritorial, dio luz verde a la creación de la red de vigilancia genómica para el SARS-CoV-2, que ahora conocemos como RELECOV. Esta iniciativa ha sido respaldada por la financiación de dos proyectos europeos sucesivos destinados a establecer y fortalecer la vigilancia genómica a nivel nacional del SARS-CoV-2 (Subvención ECDC/HERA 2021/2021/PHF/23776 y EU4H-2022-DGA-MS-IBA-1).

En este contexto, uno de los principales objetivos de la red es desarrollar una plataforma que facilite la vigilancia genómica a nivel nacional a través de la colaboración entre los laboratorios. Dicha plataforma permitirá el almacenamiento y el análisis conjunto de datos de secuenciación de genomas completos de virus, ofreciendo herramientas de visualización y análisis, mostrando la evolución viral en el territorio nacional. Además, esta plataforma facilitará la compartición centralizada de datos a nivel nacional e internacional a través de bases de datos públicas, cumpliendo así con los requisitos europeos.

Este manual proporciona una guía detallada de los pasos que deben seguirse para incluir en la plataforma las muestras secuenciadas en cada laboratorio perteneciente a la red, junto con la información de metadatos necesaria para un seguimiento y análisis adecuados de las muestras.

02 COMPARTIR METADATOS

Pasos para incluir muestras en la plataforma

Opción 1. Plantilla excel

Los metadatos serán proporcionados en tablas de formato excel (xlsx), para ello se dispone de la siguiente plantilla que debe ser rellenada siguiendo las instrucciones a continuación:

Relecov_metadata_template_v3.0.0.xlsx

La primera hoja (Overview) muestra la descripción de cada campo con un ejemplo de cómo debe completarse, indicando también qué campos son obligatorios. Los campos obligatorios son en los que existe una “Y” en la columna “Mandatory”.

La segunda hoja (METADATA_LAB) es la hoja que se debe cumplimentar. La primera línea incluye la indicación si el campo es requerido (YES). La segunda línea incluye un ejemplo, la tercera línea una descripción del campo y la cuarta línea es el nombre del campo. No debes modificar ni eliminar ninguna de estas cuatro líneas, solo debes comenzar a completar la información de la primera muestra en la quinta línea, creando una línea para cada muestra. Existen campos con opciones limitadas que tienen que seleccionarse de la lista desplegable, en la que puedes seleccionar el valor apropiado para tu muestra. Si no encuentras la opción que necesitas, por favor escríbenos a bioinformatica@isciii.es

La tercera hoja (DATA VALIDATION) incluye las opciones disponibles para cada campo. Si necesitas opciones adicionales, por favor contáctanos en bioinformatica@isciii.es

Instrucciones para rellenar la plantilla Excel

1. La cumplimentación de la tabla ha de ser llevada a cabo utilizando los campos que vienen predefinidos en sendos desplegables que aparecen para cada columna. Aquellas muestras con campos indebidamente rellenos serán automáticamente descartadas durante el proceso de validación de los metadatos.
2. Sólo para aquellos campos que no cuenten con este desplegable, como es el caso de los identificadores o la fecha se permitirá el texto libre.
3. El conjunto mínimo de datos a recoger se ha establecido en función de discusiones con el grupo de trabajo de RELECOV y basándose también en los requerimientos mínimos de bases de datos públicas como ENA o GISAID. Los siguientes campos son obligatorios, y cualquier muestra que no los contenga no podrá ser incluida en la plataforma:
 - Organism
 - Identificadores de la muestra (pueden ser repetidos, según el caso):
 - Sample ID given by originating laboratory*
 - Sample ID given if multiple rna-extraction or passages
 - Sample ID given for sequencing**
 - Originating Laboratory
 - Submitting Institution
 - Sample Collection Date (Formato: YYYY-MM-DD. Ej: 2020-08-29)
 - Host
 - Librería utilizada: En este caso se podrá rellenar una de las dos siguientes ya que son mutuamente excluyentes.
 - Commercial All-in-one library kit
 - Library Preparation Kit
 - Enrichment panel/assay
 - Enrichment panel/assay version

- Sequencing Instrument Model
- Source material
- Sequencing technique
- Library layout
- Sequence file R1 fastq
- Sequence file R2 fastq (Requerido en Paired-end)

Nota: En el caso de que se quieran utilizar valores no recogidos en el desplegable, por favor, remita este caso a la unidad de bioinformática del ISCIII (bioinformatica@isciii.es) para que sea revisado.

*Los campos "Sample ID given by the submitting laboratory" y "Sample ID given in the microbiology lab" pueden ser iguales al campo previo "Sample ID given by originating laboratory".

**Es muy importante que este código sea único para que el proceso de automatización funcione correctamente. Si la muestra es re-secuenciada el número de la carrera debe ser precedida por un punto. [SEQ_220624 (SEQ_220624.2)].

Opción 2. Uso de relecov-tools***

Relecov-tools es un paquete de herramientas desarrollado en Python con el propósito de ayudar a los laboratorios de redes a procesar y cargar sus metadatos en la base de datos. Este paquete cuenta con un módulo que permite la homogeneización de los metadatos a un formato estándar y su subida a la plataforma. <https://github.com/BU-ISCIII/relecov-tools>

***pronto disponible

03 SUBIDA DE DATOS DE SECUENCIACIÓN AL REPOSITORIO REMOTO (SFTP)

Un repositorio SFTP (Secure File Transfer Protocol) es un sistema de almacenamiento remoto de archivos que utiliza el protocolo seguro SSH para transferir y administrar archivos de manera segura a través de una conexión cifrada.

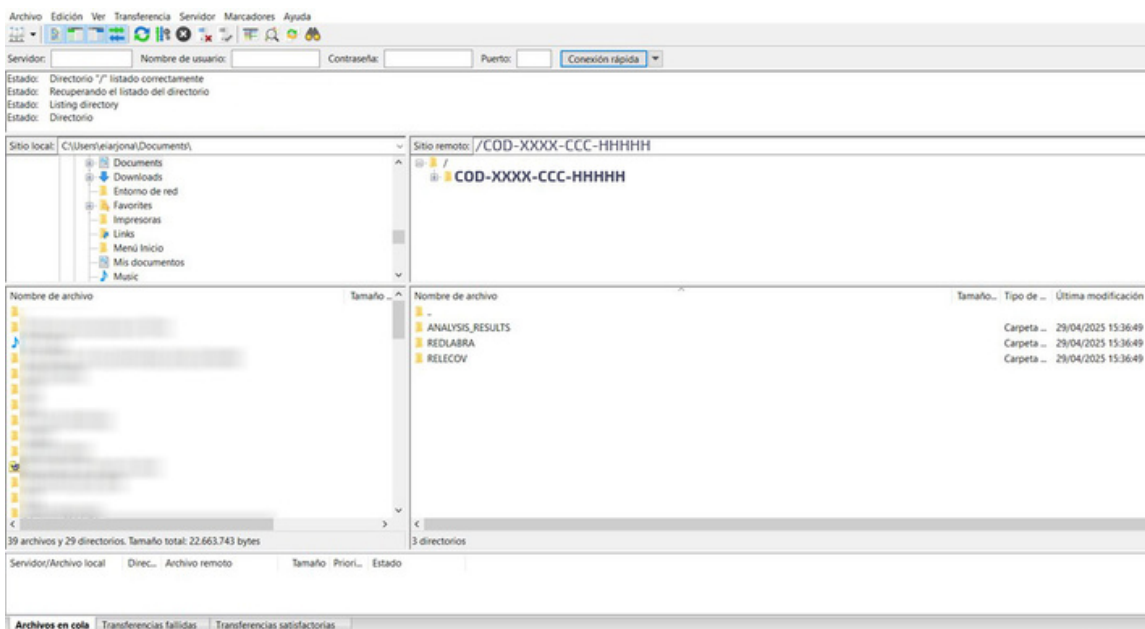
Una vez rellenado el excel de metadatos, este se subirá al repositorio sftp en conjunto con los archivos de las secuencias, que deberán estar en formato fastq comprimido (*.fastq.gz).

Estructura de carpetas dentro del SFTP

Cada laboratorio contará con una carpeta propia en el repositorio sftp con el siguiente nombre (COD-XXXX-CCC-HHHHH). Dentro de esa carpeta existen tres carpetas con los nombres siguientes:

- RELECOV
- REDLABRA
- ANALYSIS_RESULTS

Carpeta RELECOV: Carpeta para la subida de datos de RELECOV. No es posible realizar descargas de esa carpeta.



Preparación de los datos

En la carpeta RELECOV deben estar juntos el excel de metadatos, los fastq.gz y el md5sum. Todas las muestras que se quieran subir al repositorio han de estar contenidas en el excel de metadatos para su correcta validación.

Recomendamos mantener una nomenclatura común entre los nombres de los ficheros de las muestras, en especial para la referencia a la orientación de la lectura. Se ruega evitar la utilización de espacios en los nombres ya que dificulta enormemente el procesamiento de los archivos, recomendando sustituir éstos por guiones o guiones bajos. Por ejemplo: Para “muestra sars 257 forward” y “muestra sars 257 reverse” algunas sugerencias serían:

- **muestrasars257_R1.fastq.gz / muestrasars257_R2.fastq.gz**
- **muestra-sars-257_R1.fastq.gz / muestra-sars-257_R2.fastq.gz**
- **muestra_sars_257_R1.fastq.gz / muestra_sars_257_R2.fastq.gz**

Comprobación de integridad de los ficheros

La copia y/o transferencia de archivos no es perfecta y lleva consigo un porcentaje de error por cada byte que se transfiere, que puede dar lugar a la corrupción del archivo transferido. Esta probabilidad, aunque mínima, aumenta proporcionalmente con el tamaño de los archivos. Para garantizar la integridad de los archivos, cada archivo deberá contar con un hash md5 para verificar que la transferencia ha sido correcta y que el archivo no ha sido modificado ni ha quedado corrupto durante el proceso.



Opción 1. Sistema Operativo Windows

Opción 1.a línea de comandos de Windows

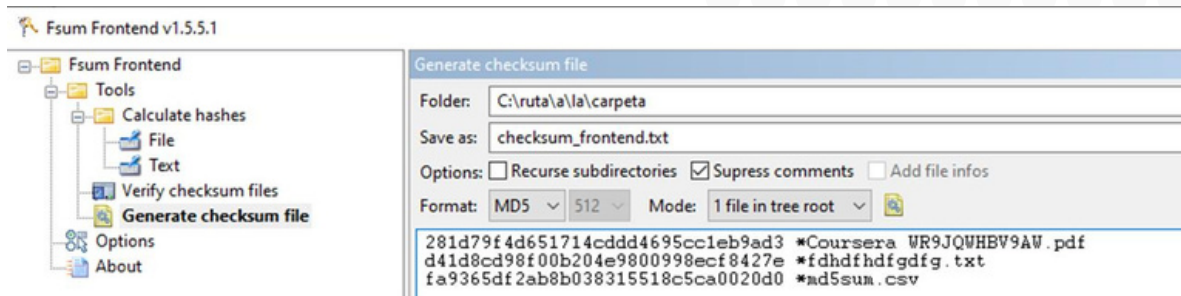
1. Abrir una consola/terminal en Windows (En inicio, buscar PowerShell ISE y ejecutar)
2. Una vez abierta la terminal:
 - a. Ir a la carpeta donde estén guardados los ficheros que van a ser subidos al repositorio (mantener comillas “ ”): `cd "C:\ruta\ejemplo\carpeta"`
 - b. Pegar el siguiente comando en la terminal para generar los hash md5 de todos los archivos en la carpeta.

Get-FileHash -Algorithm MD5 -Path (Get-ChildItem . -Recurse) | Select-Object Hash, Path | export-csv -Path "md5sum.csv" -Delimiter "`t" -NoTypeInfoInformation

```
1 "Hash" "Path"
2 "30DFA94E08451977CC0E6DE0155ED2CD" "C:\Users\ruta\ejemplo\carpeta\secuencia_R1.fastq.gz"
3 "D41D8CD98F00B204E9800998ECF8427E" "C:\Users\ruta\ejemplo\carpeta\secuencia_R2.fastq.gz"
4 "281D79F4D651714CDD4695CC1EB9AD3" "C:\Users\ruta\ejemplo\carpeta\excel_metadatos.xlsx"
5 "FA9365DF2AB8B038315518C5CA0020D0" "C:\Users\ruta\ejemplo\carpeta\md5sum.csv"
```

Opción 1.b Utilizando software externo [Fsum FrontEnd]

1. Descarga Fsum FrontEnd: <https://sourceforge.net/projects/fsumfe/>
2. Extraerlo en una carpeta cualquiera
3. Ejecutar FsumFrontEnd.exe
4. Hacer click en “Generate checksum file”
5. Elegir la carpeta con los archivos y el nombre del archivo
6. Cambiar el algoritmo de encriptación a MD5 en “Format”.
7. Hacer click en “Supress comments” pues esto impediría la verificación automática.
8. Hacer click en generar o pulsar “Ctrl + G”



Opción 2. Sistema Operativo Linux

En el caso de linux, se pueden utilizar los siguientes comandos de consola desde la carpeta en la que tengamos los archivos listos para ser subidos al repositorio:

1. Abrir una terminal/consola de linux
2. Acceder a la carpeta: `cd /ruta/a/la/carpeta/`
3. Generar los hash md5 con el siguiente comando: `md5sum * > md5sum.txt`

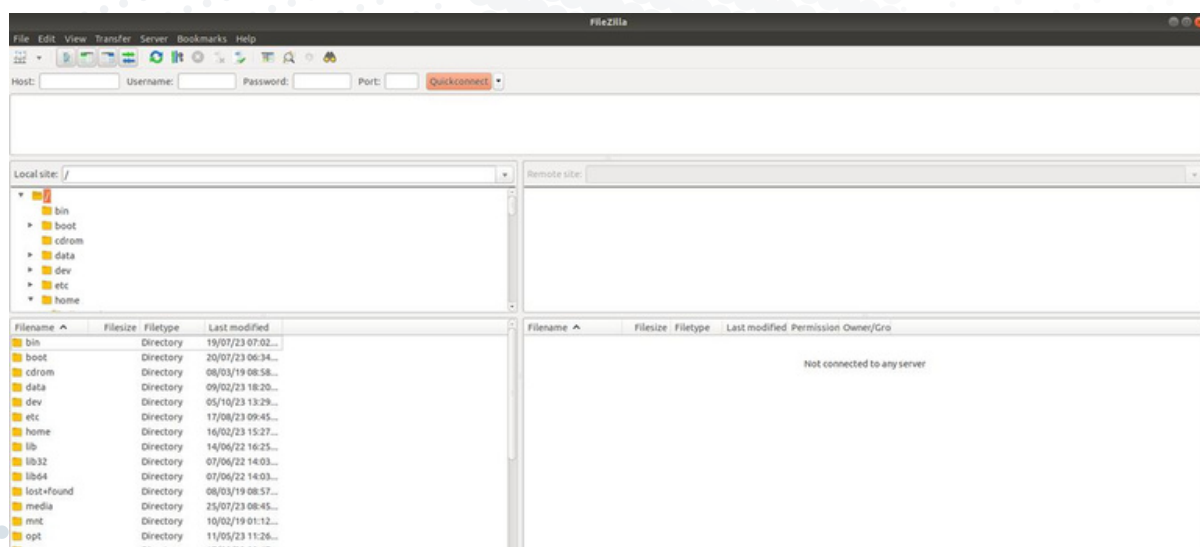
Esto generará automáticamente un archivo de texto en el que vendrán recogidos todos los archivos de la carpeta seleccionada seguidos de su hash único para su validación.

Subida de ficheros al recurso sftp

Para facilitar el proceso se recomienda el uso de la herramienta **Filezilla®** ya que cuenta con interfaz gráfica y utiliza el protocolo sftp por defecto para la transferencia de datos, garantizando la seguridad en el proceso.

Manual para subir los datos usando Filezilla®

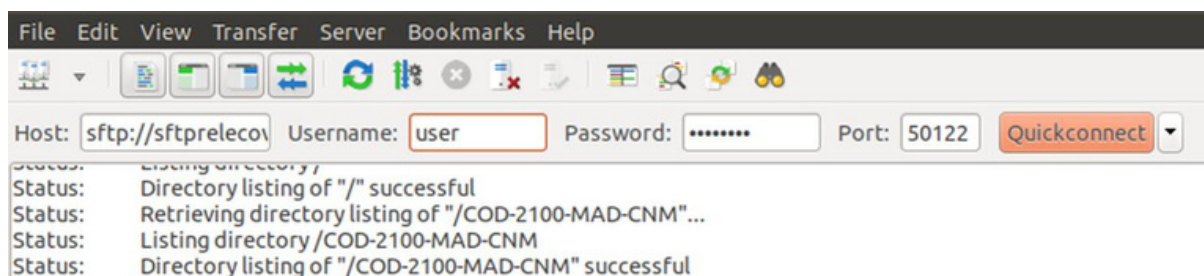
Una vez abramos Filezilla, nos encontraremos con una pantalla como la siguiente:



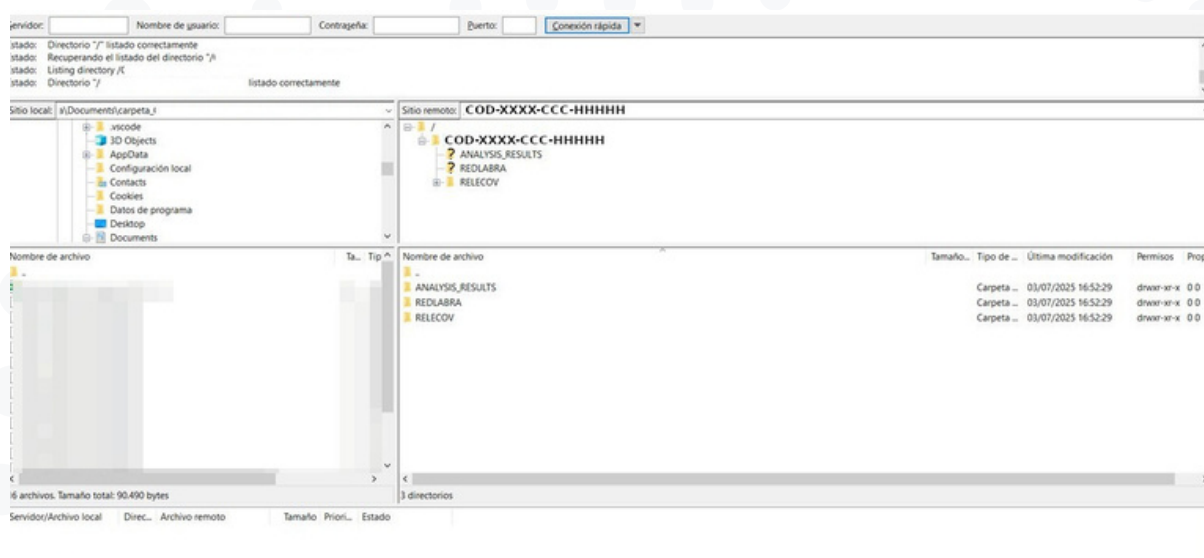
La configuración de acceso remoto en protocolo sftp/ssh-ftp es la siguiente:

- **URL acceso:** sftp://sftpgenvigies.isciii.es
- **Autenticación:** Usuario + Contraseña
- **Puerto:** 50122

Tendremos que introducir la URL de acceso al repositorio así como las credenciales provistas previamente y hacer click en **Quickconnect**.



Si la conexión se ha realizado correctamente, aparecerán en la parte derecha nuestra carpeta, asociada a nuestro laboratorio. Debemos hacer click en la carpeta correspondiente a nuestro centro para poder acceder a ella:



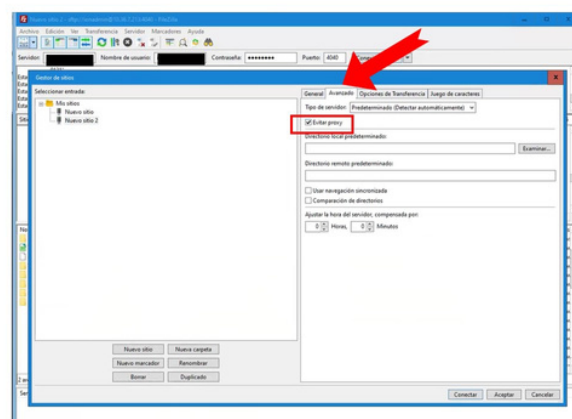
Después, tan solo quedará arrastrar los archivos que queramos subir desde nuestra carpeta local (izquierda) a la del repositorio (derecha) y esperar a que termine la transferencia de datos.

Bajada de ficheros BAM del recurso sftp de la Suite de Ion Torrent.

Los laboratorios que utilicen la secuenciación con Ion Torrent deberán subir al SFTP de RELECOV los archivos llamados BAM que se generan desde el plugin “VariantCaller” que está presente en la mayoría de las suites de análisis de Ion Torrent. En este documento se puede encontrar la forma detallada de conseguir los BAM.

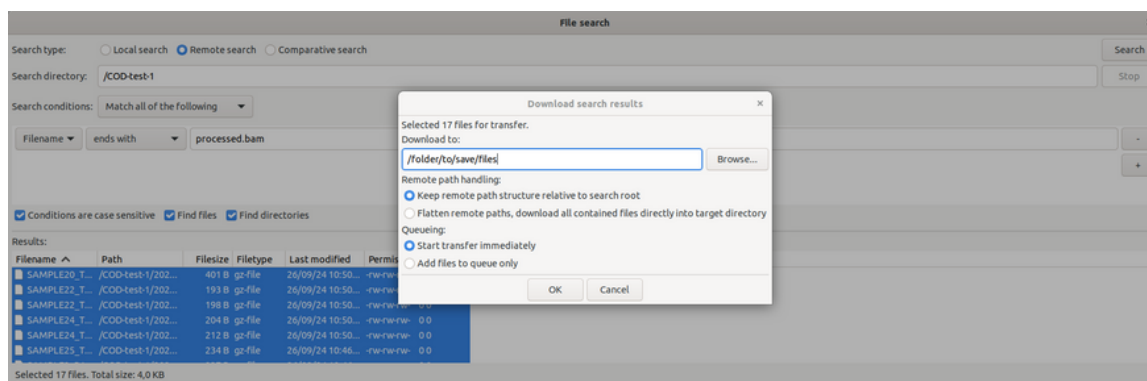
Una vez se han generado los **TVC-Processed Reads** (en los que ya se han eliminado los primers), los archivos BAM se pueden descargar manualmente desde el experimento, o a través de el SFTP del secuenciador, a donde se puede acceder con el usuario y contraseña de cada laboratorio (para conocer estos datos tendrá que ponerse en contacto con la casa comercial).

Al SFTP del secuenciador puede accederse también por medio del software **Filezilla®**. Es importante tener en cuenta que en la pestaña Avanzado del gestor de sitios, hay que marcar “Evitar Proxy”



Una vez ubicados en la carpeta correspondiente, puede seguirse la secuencia de pasos siguiente, para hacer una descarga automática de todos los archivos BAM, con el nombre “**processed.bam**” que serían los archivos que nos interesan.

1. En la barra de herramientas, click en "server" (O pulsar F3 para buscar directamente)
2. Click en "Search remote files"
3. Elegir la carpeta e indicar "**ends with**" y "**processed.bam**"
4. Seleccionar todos los archivos en la parte de abajo
5. Click derecho y "download"
6. Elegir la carpeta de destino y hacer click en "ok"



Una vez descargados los archivos en local, pueden subirse al SFPT de RELECOV, siguiendo el paso previo de este punto.

04 FAQs

Problemas con la conexión al repositorio sftp

1. No puedo conectarme al recurso sftp con Filezilla®.

- Asegúrate de que el protocolo de transferencia de archivos seleccionado es sftp. En el caso de Filezilla esta opción está por defecto pero puedes asegurarte añadiendo un nuevo sitio
- En ocasiones el puerto de entrada 50122 está cerrado por el firewall de tu centro. Es posible que tengas que pedir acceso a la unidad de gestión de sistemas informáticos de tu centro.
- Si aparece el error “mala resolución del hostname” existe un error en la dirección de acceso al repositorio. En algunos casos, sustituir la dirección de Host “sftp://sftpgenvigies.isciii.es” por la ip “212.128.123.109” es suficiente.

2. He olvidado mi usuario y/o contraseña de acceso al repositorio sftp.

- Por favor, contacte con la unidad de bioinformática del ISCIII (bioinformatica@isciii.es) para poder facilitarles sus credenciales de nuevo.