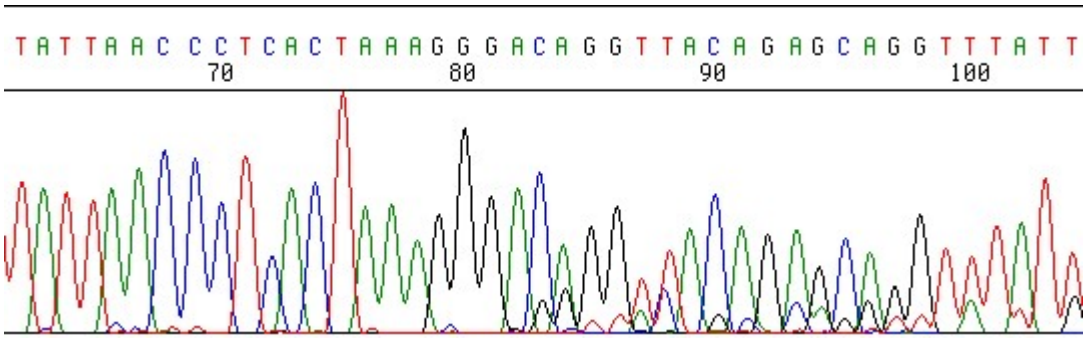
Documentación proceso

Automatización entrega secuencias sanger

v.2.3



Bioinformatics Unit , Instituto de Salud Carlos III
Majadahonda (Madrid)

Contenido

1 INTRODUCCIÓN

La Unidad de Genómica lleva a cabo el servicio de Secuenciación Sanger del Instituto de Salud Carlos III. A través de este servicio los investigadores pueden solicitar la secuenciación de sus muestras a partir de un formulario. Una vez concluido el servicio la Unidad de genómica pone a disposición del investigador los resultados de la secuenciación utilizando un recurso compartido situado en Cibeles.

Esta situación presenta una serie de problemas:

- Obligatoriedad de realizar la copia de ficheros/resultados a la carpeta del investigador solicitante correspondiente. Una misma carrera de secuenciación contiene secuencias de varios usuarios, por lo que la separación de cada grupo de ficheros a cada investigador tiene que hacerse de manera manual.
- Los permisos a las carpetas de los distintos laboratorios que se utilizan en Cibeles se mantienen de manera difícil en el tiempo debido a la gran movilidad del personal en el centro.
- Mediante este mecanismo no hay una notificación al usuario de que tiene disponibles las secuencias, siendo éste un requisito solicitado para cumplir la norma ISO por la que está acreditada la Unidad de Genómica para el ensayo de secuenciación Sanger.

Para solucionar estos problemas se ha desarrollado un proceso de automatización que permite la compartición personalizada de los resultados de cada carrera a cada investigador solicitante. Además se realiza una notificación por mail a la persona o personas solicitantes con un link a la descarga de los archivos necesarios.

Por último se especifica que las muestras serán borradas automáticamente al cabo de 7 días, siendo los datos almacenados para su backup y archivo por la Unidad de Genómica con un proceso ajeno a este procedimiento.

2 PASOS A REALIZAR POR LA UNIDAD DE GENÓMICA

2.1 ACTUALIZACIÓN DEL FICHERO DE CONFIGURACIÓN DE LA CARRERA

Este fichero contiene todas las muestras de la secuenciación que han de ser asignadas a los investigadores. Para realizar esta separación se ha de rellenar en el campo “Comment” la dirección de correo del investigador para poder enviarle un email con el enlace de la carpeta donde se puede descargar las muestras.

Ejemplo:

Well	Sample Name	**Comment**	Results Group 1	Instrument Protocol 1	Analysis Protocol 1
A01	1992PBS_R	**jperez@.isciii.es**	Secuenciacion	LongSeq50_POP7_Z_00518	3730P7BD3_1_seq

Rellenando todas las filas de la columna “Comment” asignando cada una de ellas con el correo del investigador se podrá asignar todas las muestras pertenecientes a un investigador y permitir con ello que pueda acceder a ellas.

En el caso de que una muestra o conjunto de muestras hayan de compartirse con más de 1 investigador, se separaran los correos electrónicos con 2 puntos “:”.

Ejemplo:

Well	Sample Name	**Comment**	Results Group 1	Instrument Protocol 1	Analysis Protocol 1
A11	44587SAR_R	**igarcia@.isciii.es:pruiz@isciii.es**	Secuenciacion	LongSeq50_POP7_Z	37OP7B9_1_seq

2.2 Copia de ficheros a recurso compartido

Una vez finalizada la carrera y disponibles las secuencias a compartir, se copiarán al recurso compartido sanger_seq al que tiene acceso la Unidad de genómica:

- Fichero “.txt” generado para la configuración de carrera con la información del solicitante. Ej. GN18-190INFA.txt
- Carpeta con las secuencias obtenidas de la carrera. Ej. GN18-190INFA

2.3 (Opcional) Re-compartición de ficheros no copiados por el investigador

En el caso de que haya pasado el periodo establecido para que los resultados hayan sido copiados por el investigador y el recurso compartido ya no esté disponible; se ha creado un mecanismo para poder volver a compartir unos datos específicos de una carrera y un investigador.

Para ello debe rellenarse el fichero reshare.csv en el recurso sanger estableciéndose el nombre de la carrera y el email del investigador separado por una coma. Sólo un investigador está permitido (UN SOLO MAIL). En el caso de que se quiera compartir con varios investigadores debe hacerse en distitnas líneas.

GN18-176A	smonzon@isciii.es

3 Recepción del correo con el enlace

El proceso de automatización detectará cuando los resultados de una nueva carrera estén disponibles y separará los resultados por investigador generando una carpeta compartida donde sólo tendrán acceso los usuarios establecidos en el archivo de configuración de la carrera.

Por último enviará un correo electrónico como el que se muestra a continuación informando de que los resultados están disponibles, junto con el link necesario para acceder a las secuencias.

Sanger sequencing run GN18-190AINFA has finished.

Your samples has been successfully sequenced!

You can retrieve your samples from:

\\barbarroja\20181030_GN18-190AINFA_m.jimenez_smonzon

The samples will be available for 7 days, afterwards the folder will be deleted.

If you have any question/issue please contact us:

Contacto	Email
Bioinformatics Unit	bioinformatica@isciii.es
Genomics Unit	azaballos@isciii.es

BU-ISCIII
Bioinformatics Unit , Instituto de Salud Carlos III
Majadahonda
Madrid

NOTA:

Para una misma carrera, el usuario podrá recibir más de un correo, dependiendo de las muestras que ha compartido con otros usuarios.

Es decir, si en el fichero configuración de la carrera fuese como en este ejemplo:

Well	Sample Name	**Comment**	Results Group 1	Instrument Protocol 1	Analysis Protocol 1
A10	44587SAR_R	**paco@.isciii.es:pedro@isciii.es**	Secuenciacion	LongSeq50_POP7_Z	37OP7B9_1_seq
C01	45787SAR_R	**paco@.isciii.es:eva@isciii.es**	Secuenciacion	LongSeq50_POP7_Z	37OP7B9_1_seq
D05	44857SAR_R	**paco@.isciii.es**	Secuenciacion	LongSeq50_POP7_Z	37OP7B9_1_seq

El usuario Paco recibiría:

- Un correo con el enlace a las muestras compartidas con Pedro
- Otro correo con el enlace de las muestras compartidas con Eva
- Otro correo con el enlace de las muestras que son sólo de él.

4 REQUISITOS E INSTALACIÓN DEL SCRIPT DE AUTOMATIZACIÓN

4.1 DESCRIPCIÓN DEL SCRIPT

Para el desarrollo de este proceso de automatización se ha desarrollado un script en bash que realiza los siguientes pasos:

- Separación de las secuencias por investigador.
- Compartición de las carpetas mediante CIFS/SAMBA.
- Envío de correo electrónico por carpeta compartida.

La ejecución del script es controlado por un proceso crontab, configurado para visualizar la carpeta compartida sanger_seq de la Unidad de genómica, y esperar a que una nueva tanda de secuencias estén disponibles. Cuando esto ocurre el proceso crontab lanza el script de compartición de secuencias.

4.2 REQUISITOS

Para su funcionamiento el script necesita para funcionar:

1) Un servidor de ficheros samba con S.O Linux.
2) Un servidor con S.O Linux que ejecute el script mediante un proceso crontab.

Nota: el programa está testado y desarrollado en centos 6.10 por lo que se recomienda su utilización.

El programa tiene las siguientes dependencias:

- crontab
- sendmail
- ssh
- rsync

Además se tienen en cuenta estos requisitos:

- En el servidor donde se ejecutará el script estará montado de forma permanente el directorio "**sanger**" donde se situarán los ficheros que servirán de input al script de Sanger.
- Los ficheros resultantes de la ejecución del script se copiarán en el servidor de ficheros remoto en **/srv/genomica/sanger_users**

Para poder ejecutar y copiar los ficheros al servidor de ficheros se ha de cumplir con las siguientes condiciones:

1. La ejecución del script de Sanger requiere que se creen los ficheros primeramente en dicho servidor para posteriormente copiarse, por ello el script debe tener acceso de escritura en la carpeta donde se está ejecutando.

2. Debe estar abierta en los firewalls la conexión por ssh entre los 2 servidores.
3. Se utilizará un usuario con permisos para reiniciar el servicio de “samba” así como la copia de las carpetas compartidas.
4. La conexión por ssh requiere la utilización de introducir la contraseña en el momento que el script va a realizar la copia de los ficheros. Esto implica que el script se quedará esperando a que de forma manual se introduzca la contraseña para establecer la conexión. Se recomienda que se copie la clave de ssh en el servidor de ficheros para evitar teclear la contraseña cada vez que se ejecuta el script (se describe en el **punto 4.3.3**)

4.3 NSTALACIÓN

El personal de sistemas decidirá en que directorio se situará los ficheros que componen el script de Sanger:

La solución está compuesta por los siguientes ficheros:

1. sanger_crontab_script.sh
2. sanger_script.sh
3. sanger_remove_old_files.sh
4. template_mail.htm
5. template.conf
6. sanger_configuration

Los ficheros con extensión ”.sh” son los scripts de bash, los ficheros “template” contiene la plantilla que se modificará durante la ejecución del script para enviar el correo al usuario y el otro para configurar la compartición de los ficheros mediante samba. El último fichero “sanger_configuration” contiene la configuración de las variables del script para determinar la localización de los ficheros, así como los parámetros para la conexión hacia el gestor de ficheros.

4.3.1 Copia de los ficheros del script de Sanger.

Lo primero que se ha de hacer es crear la carpeta donde se va a alojar el script.

 # mkdir /opt/<nombre_carpeta>

El código se puede obtener del recurso svn de bioinformática y descomprimirlo

 # tar -xvf sanger_script_v2.0.tar.gz

Alternativamente se puede clonar el repositorio de github:

 # git clone https://github.com/BU-ISCIII/sanger_script.git .

Nota.- El punto del final es obligatorio para que copie los ficheros dentro de esa carpeta. De no hacerlo git crea una subcarpeta con el nombre del proyecto en git hub.

4.3.2 Modificación de los parámetros de configuración

Para establecer las variables de configuración del script se ha creado un fichero de configuración ***"sanger_configuration"*** con el objeto de que no tenga que modificar los scripts.

El fichero de configuración se ha dividido en 3 bloques:

1. Ajustes para los parámetros en el servidor donde se ejecuta el script denominado "local server".

2. Ajustes para la conexión hacia el servidor de ficheros y las carpetas donde se situarán los ficheros compartidos

3. Ajustes para el borrado de ficheros antiguos.

CONFIGURATION FILES ON LOCAL SERVER

En este bloque se configurarán los parámetros usados en el servidor donde están los scripts

- **SANGER_SCRIPT**. Path donde está instalado el script, incluyendo el nombre del script "sanger_script.sh" (E.g /opt/sanger_script/sanger_script.sh)
- **PROCESSED_FILE_DIRECTORY**. Path donde se guardarán información de que ficheros han sido procesados. (E.g /opt/sanger_script/logs). **Nota:** el directorio debe existir.
- **PROCESSED_FILE_NAME**. Nombre del fichero que va a contener los ficheros procesados. (E.g run_processed)
- **PATH_SANGER_FOLDER**. Path donde se ha montado la carpeta de samba en el servidor donde se ejecuta el script. (E.g /srv/sanger)
- **TMP_SAMBA_SHARE_DIR**. Path temporal donde se guradarán temporalmente las carpetas antes de ser copiadas al servidor de ficheros. (Eg. /opt/sanger_script/tmp/share)
- **SAMBA_SHARE_TEMPLATE**. Path donde se haya la plantilla de configuración de las carpetas de Samba, incluyendo el nombre del fichero "template.conf" (E.g /opt/sanger_script/template.conf)
- **SAMBA_TRANSFERED_FOLDERS.** Path donde se guardarán la información de las carpetas creadas en el servidor de ficheros para su posterior eliminación. (E.g /opt/sanger_script/transfered_folder)
- **TEMPLATE_EMAIL**. Path donde se haya la plantilla para el envío de correos incluyendo el nombre del fichero "template_mail.htm" (E.g /opt/sanger_script/template_mail.htm)

CONFIGURATION FILES ON REMOTE SERVER

En este bloque se configurarán los parámetros relacionados en el servidor de ficheros

DIRECTORY ON THE REMOTE SERVER, WHERE THE SHARED FILES WILL BE COPY

- **REMOTE_SAMBA_SHARE_DIR**. Path donde estarán los ficheros de configuración de samba . (E.g /etc/samba/smb.conf.d)

- **REMOTE_SAMBA_SHARED_FOLDER**. Path donde se copiarán las carpetas compartidas. (E.g /srv/genomica/sanger_users)
- **REMOTE_USER.** Usuario para la conexión al servidor de ficheros. (E.g bioinfoadm)
- **REMOTE_SAMBA_SERVER**. Nombre del servidor de ficheros (E.g neptuno)

########## Configuration settings for deleting old folders ########

Este último bloque contiene los parámetros usados en el script de borrado de ficheros antiguos.

1. **RETENTION_TIME_SHARED_FOLDERS**. Tiempo en días que se han de guardar las carpetas compartidas antes de borrarse. (E.g +181)
2. **RETENTION_TIME_CONF_FILES**. Tiempo en días que se han de guardar las carpetas compartidas antes de eliminar la compartición. (E.g +15)

4.3.3 Copia de la clave pública al servidor de ficheros (Opcional)

Como comentábamos anteriormente a la hora de copiar los ficheros al servidor de ficheros el script se parará a la espera de que se introduzca la contraseña.

Para evitar esto se recomienda que se copie la clave pública existente donde se está ejecutando el script al servidor de ficheros.

Para ello se ejecutará el siguiente comando en el servidor donde se ejecuta el script:

 # ssh-copy-id -i ~/.ssh/<mykey> root@<host>

Se supone que la clave ya está creada en el servidor.

4.4 Configuración en el crontab

El último paso será crear 2 entradas en el crontab del usuario que ejecuta el script.

- Añadiendo el fichero “sanger_crontab_script.sh”, para que se ejecute el script de sanger.
- Añadiendo una segunda línea “sanger_remove_old_files.sh” para el borrado de ficheros compartidos, una vez pasado el periodo de retención.

4.5 Logs

Se generan tres logs diferentes:

1. **Crontab log:** información sobre las operaciones que realiza el script.
2. **Run_processed:** nombre de las carreras de secuenciación sanger procesadas.
3. **Samba_folders:** información sobre las carpetas que se van compartiendo indicando carrera, investigador/es, fecha y número de ficheros.

Además de los logs se enviará un correo en la detección de un error tanto al usuario si es un problema de formato en los ficheros de entrada, como a los desarrolladores y/o administradores ante cualquier error de ejecución.

VERSION CONTROL

Version	Author	Date	Reasons for the Document update
V.1.0	Luis Chapado	20/11/2018	Primera guía de usuario y de instalación del proceso de automatización de entrega de resultados sanger.
V. 1.1	Sara Monzón	26/11/2018	Revisión del texto. Introducción añadida e información de requisitos de instalación,
V2.0	Sara Monzón y Luis Chapado	25/04/2019	Actualización de instalación de la nueva versión del script.
V2.2	Sara Monzón	22/08/2019	Inclusión del mecanismo de re-compartición. Información sobre emails de error.
V2.3	Sara Monzón	02/01/2020	Aclaración en recompartición. Actualizado índice.