Curso: ANALISIS DE DATOS DE NGS (18-22 noviembre 2019)

Servicio de Genómica y Secuenciación Masiva

Centro de Biología Molecular Severo Ochoa (CBMSO) CSIC-UAM

PROGRAMA

Módulo 1. Introducción a tecnologías NGS. Ensamblaje de novo y anotación de genomas.

Las tecnologías NGS han revolucionado el campo de la genómica, permitiendo la secuenciación de un gran número de genomas en muy poco tiempo. La metodología implica fragmentar el ADN, reparar los extremos de las moléculas e incorporar en ellos adaptadores universales que permiten la secuenciación paralelizada de un número enorme de estas moléculas en cada experimento. El genoma brinda, en principio, el catálogo completo de genes que un organismo puede expresar, pero la interpretación de esta información, a todos los niveles (desde el enorme volumen de datos crudos originados por el secuenciador, hasta los miles de genes identificados asociados a diversas funciones) constituye un **gran reto desde el punto de vista computacional**.

El **ensamblaje *de novo*** es uno de los procedimientos empleados para **reconstruir genomas desconocidos**. A partir de un gran número de secuencias nucleotídicas proporcionadas por los secuenciadores, y recurriendo a potentes algoritmos, se puede reconstruir el genoma del organismo estudiado. La **calidad de esta reconstrucción dependerá de varios factores**, como es la cantidad total de bases secuenciadas, longitud de las secuencias obtenidas, la existencia o no de zonas repetitivas en el genoma original, el contenido en GC del organismo, etc..

En este módulo se presentará en primer lugar una **visión general de las tecnologías** de NGS actuales. Seguidamente se procederá a explicar los distintos **formatos de archivos usados** en NGS, así como la exploración de los repositorios FTP para la **descarga de los genomas de referencia** (EBI, NCBI, UCSC). Por último, se tratará la técnica de **ensamblaje *de novo***.

18 noviembre 2019 (lunes)	**Introducción a tecnologías NGS. Ensamblaje *de novo* y anotación de genomas.**	
Horario	Actividad	Docente
09:00 – 09:15	Presentación y breve introducción al curso.	Begoña Aguado
09:15 – 10:45	Tecnologías de NGS y ensamblaje *de novo*	Fernando Carrasco
10:45 – 11:30	Formatos de archivos usados en NGS. Exploración de repositorios FTP para descarga de genomas de referencia (EBI, NCBI, UCSC).	Ramón Peiró Eva Castillo
11:30 – 12:00	*Pausa.*	
12:00 – 12:30	Descarga de la referencia de *Escherichia coli K-12 MG1655.* Exploración de ENA y SRA.	Ramón Peiró Eva Castillo
12:30 – 14:00	Ensamblaje *de novo* con SPAdes. Anotación del genoma de *E. coli*.	Sandra González Eva Sacristán

Módulo 2. Resecuenciación. Análisis de cobertura y variantes.

La resecuenciación consiste en la **secuenciación de nuevas muestras de organismos previamente secuenciados**, como técnica para detectar potenciales cambios como **SNPs, deleciones o inserciones**, **etc**, en conjunto **denominados variantes**. La resecuenciación puede ser de **genoma completo o dirigida**. Esta última consiste en el aislamiento, enriquecimiento y secuenciación de regiones específicas de interés del genoma en una muestra. La resecuenciación permite la detección sistemática tanto de variantes comunes como variantes raras o poco frecuentes. Actualmente, la combinación entre sistemas de enriquecimiento en solución y la secuenciación masiva se ha convertido en el método de elección para caracterizar de forma selectiva un gran número de genes de manera simultánea gracias a su alta precisión, reproducibilidad y rendimiento. Por otro lado, la **resecuenciación del exoma** es una técnica novedosa que permite la captura, el enriquecimiento y la secuenciación de regiones genómicas codificantes. La resecuenciación del exoma completo en humano permite la identificación de nuevos genes asociados tanto a enfermedades raras como comunes.

En este módulo se explicará en qué consiste la **resecuenciación**, así como los **análisis de cobertura y variantes**. Se procederá a enseñar cómo realizar descargas de genomas de referencia y de las lecturas, a realizar controles de calidad de las lecturas y alineamientos de las lecturas contra el genoma de referencia. Finalmente, se tratará el análisis de cobertura y la búsqueda de variantes (Variant Calling).

19 noviembre 2019 (martes) **Resecuenciación. Análisis de cobertura y variantes.**		
Horario	Actividad	Docente
09:00 – 9:30	Resecuenciación. Análisis de cobertura y variantes.	Fernando Carrasco
9:30 – 10:00	Descarga de la referencia de *Mycobacterium tuberculosis* H37R y de lecturas. Control de calidad de las lecturas. Recorte.	Eva Castillo Ramón Peiró
10:00 – 11:00	Alineamiento de lecturas contra el genoma de referencia con BWA.	Eva Castillo Ramón Peiró
11:00 – 11:30	*Pausa.*	
11:30 – 13:00	Análisis de cobertura con *genomeCoverageBed*. Visualización con IGV.	Eva Sacristán Sandra González
13:00 – 14:00	Análisis de variantes con *GATK*. Anotación de variantes con *SnpEff*.	Eva Sacristán Sandra González

Módulo 3. Detección de picos (ChIP-Seq).

La estructura de la cromatina desempeña un papel fundamental en la función del ADN. Regula procesos que se dan sobre la estructura nucleosomal como la transcripción, la replicación y la recombinación. Determinar la distribución de las modificaciones específicas de las histonas y sus variantes, así como la de otros componentes de la cromatina, sobre secuencias específicas del ADN, puede proporcionar información valiosa acerca de cómo funcionan estas proteínas (y sus modificaciones) dentro del contexto de la cromatina. La **Inmunoprecipitación de Cromatina (ChIP)** es un método bioquímico usado principalmente para determinar la localización en el genoma de histonas modificadas y de otras proteínas *in vivo*. También se emplea para estudiar la unión de factores de transcripción al ADN. Esta técnica consiste en el uso de un anticuerpo que reconozca la proteína de interés no solamente en disolución sino también en la cromatina. La técnica de ChIP-seq consta básicamente de dos pasos, entrecruzamiento con formaldehído del ADN a las proteínas unidas a éste *in vivo* seguido de la inmunoprecipitación de los complejos proteína-ADN con anticuerpos específicos. Las secuencias específicas de ADN inmunoprecipitadas son,mediante un protocolo específico, secuenciadas y alineadas contra el genoma de referencia, dando resultados de acumulación en los sitios de unión proteína-ADN, mostrando picos de cobertura.

En este módulo, además de explicar los **conceptos de Secuenciación de cromatina inmunoprecipitada (ChIP-Seq)**, se procederá a realizar además de tratamientos previos, **detección de picos, mapeo de los picos** y **extracción de las secuencias de DNA**, así como **detección de motivos y generación de los logos** de los motivos.

20 noviembre 2018 (miércoles) **Detección de picos (ChIP-Seq).**		
Horario	Actividad	Docente
09:00 – 9:30	Secuenciación de cromatina inmunoprecipitada (ChIP-Seq).	Fernando Carrasco
9:30 – 11:00	Descarga de secuencias y tratamiento previo. Control de calidad. Alineamiento de secuencias contra el genoma de referencia de *Escherichia coli str. K-12 substr. MG165* con Bowtie2.	Sandra González Ramón Peiró
11:00 – 11:30	*Pausa*.	
11:30 – 13:00	Detección de picos con *MACS2*. Mapeo de picos y extracción de secuencias de DNA.	Sandra González Ramón Peiró
13:00 – 14:00	Detección de motivos con *MEME*. Generación de logos de motivos. Búsqueda de motivos obtenidos en un archivo de secuencias con *FIMO*. Descarga de los archivos de interés.	Sandra González Ramón Peiró

Módulo 4. Expresión diferencial (RNA-Seq).

El **RNA-Seq** (secuenciación de ARN), también llamado **secuenciación de transcriptoma**, utiliza tecnología NGS para revelar **la presencia y la cantidad de ARN en una muestra biológica en un momento dado en el tiempo**. El RNA-Seq se utiliza para analizar el transcriptoma celular que cambia continuamente. Específicamente, el RNA-Seq permite estudiar las transcripciones de genes alternativos, modificaciones post-transcripcionales, fusión génica, mutaciones/SNP y cambios en la expresión génica a lo largo del tiempo, o diferencias en la expresión génica en diferentes grupos o tratamientos. Además de transcripciones del **mRNA**, el RNA-Seq permite estudiar las diversas poblaciones de RNA para incluir el RNA total, pequeño RNA, tal como **miRNA, tRNA, RNA ribosomal**. El RNA-Seq se puede también utilizar para determinar límites exón / intrón y verificar o corregir los límites de los genes previamente anotados, así como detectar las distintas formas de **splicing**. En esta metodología el RNA debe ser retrotranscrito a cDNA para luego fragmentarlo y preparar la correspondiente librería de forma similar a lo descrito en la introducción al módulo 1.

En este módulo se mostrará en qué consiste la **expresión diferencial mediante RNA-Seq** y como después de los tratamientos previos se puede llevar a cabo un análisis de la expresión diferencial, **splicing alternativo** y filtrado de datos.

21 noviembre 2019 (jueves) **Expresión diferencial (RNA-Seq).**		
Horario	Actividad	Docente
09:00 – 9:30	Expresión diferencial mediante RNA-Seq.	Fernando Carrasco
9:30 – 10:00	Descarga de secuencias y tratamiento previo. Control de calidad.	Eva Castillo Ramón Peiró
10:00 – 11:00	Alineamiento de lecturas contra el genoma de referencia de *Mycobacterium smegmatis str. MC2 155* con *Hisat2*. Conteo con HTseq-count.	Eva Castillo Ramón Peiró Eva Sacristán
11:00 – 11:30	*Pausa*.	
11:30 – 13:00	Análisis de expresión diferencial con *Deseq2.* Filtrado de datos.	Eva Castillo Ramón Peiró Eva Sacristán
13:00 – 14:00	*Splicing* alternativo con *rMats*.	Eva Castillo Ramón Peiró Eva Sacristán

Módulo 5. Metagenómica (16S).

La metagenómicase define como el **estudio del material genético, el cual es obtenido directamente de muestras ambientales**. Para ello, se amplifican mediante PCR genes específicos (para bacterias normalmente el gen del ARNr **16S**, para hongos el **ITS**, y para Eukariotas el gen **18S**) y se secuencian para producir un perfil específico de la diversidad en una muestra natural procedente de diversos ecosistemas (desde ambientes extremos a microbiota humana). La metagenómica ahora permite investigar la ecología microbiana a mayor escala y con mejor detalle que antes.

En este módulo se mostrará en qué consiste la metagenómica mediante **amplicones de la región 16S** y cómo analizar los datos obtenidos mediante **Qiime2**.

22 noviembre 2019 (viernes) **Metagenómica (16S).**		
Horario	Actividad	Docente
09:00 – 9:30	Metagenómica mediante amplicones de la región 16S.	Fernando Carrasco
9:30 – 10:00	Descarga de secuencias y tratamiento previo. Control de calidad.	Eva Sacristán Eva Castillo Sandra González
10:00 – 11:00	Metagenómica con Qiime2.	Eva Sacristán Eva Castillo Sandra González
11:00 – 11:30	*Pausa*.	
11:30 – 13:45	Metagenómica con Qiime2 (continuación).	Eva Castillo Eva Sacristán Sandra González
13:45 – 14:00	Conclusiones y cierre del curso.	Begoña Aguado

Equipo docente:

Eva Castillo Rosa	ecastillo@cbm.csic.es
Ramón Peiró Pastor	rpeiro@cbm.csic.es
Sandra González de la Fuente	sandra.g@cbm.csic.es
Eva Sacristán Horcajada	esacristan@cbm.csic.es
Fernando Carrasco Ramiro	fcarrasco@cbm.csic.es

Equipo organizador:

Ramón Peiró Pastor	rpeiro@cbm.csic.es
Fernando Carrasco Ramiro	fcarrasco@cbm.csic.es
Begoña Aguado Orea	baguado@cbm.csic.es

Administración/contacto:

Almudena Hernando	ahernando@cbm.csic.es