

Reporte_03

Axel R

2023-03-03

Utiliza IGV para vizualizar datos de *Plasmodium chabaudi* (*P. chabaudi*)

```
cd /mnt/Timina/bioinfoII/data/IGV/  
# Pasar archivos a computadora local para utilizarlos en IGV  
rsync -rptuvl arodriguez@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/data/IGV/PccAS_v3_genome.fa .  
rsync -rptuvl arodriguez@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/data/IGV/PccAS_v3_genome.fa.fai .  
rsync -rptuvl arodriguez@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/data/IGV/PccAS_v3.gff3 .
```

El archivo `PccAS_v3_genome.fa` es el archivo con el genoma de referencia de *P. chabaudi*, mientras que `PccAS_v3.gff3` son las anotaciones.

Carga el genoma de referencia desde un archivo (bota que ha sido *indexado*)

Para cargar el genoma de referencia en IGV se va a Genomes -> Load Genome from File...". Se selecciona "PccAS_v3_genome.fa" y se da clic en "open".

Carga las anotaciones del genoma

Para cargar el archivo de anotacion, se da click en "File -> Load from File..." y se selecciona "PccAS_v3.gff3" para despues dar click en "Open".

¿Que describe el formato del archivo de anotaciones?

No sólo se requiere ver dónde han mapeado las lecturas, sino a qué genes han mapeado. Para ello, se tiene el archivo de anotación en **formato GFF3** (`PccAS_v3.gff3`), el cual contiene una lista de características, coordenadas y orientaciones correspondientes al genoma de referencia.

¿Que esta mostrando el *browser*?

Una vez cargado el archivo, se despliega una nueva pista llamada "**PccAS_v3. gff3**". La pista se muestra como un gráfico de densidad.

¿Qué pasa si te acercas o te alejas?

Se puede hacer **zoom** para ver los genes individuales, incluyendo intrones y exones. Si se hace **zoom out** las ubicaciones de los genes se superponen mostrando la abundancia relativa de genes en todo el genoma.

En la pista de anotación, los genes se presentan como cajas azules y líneas. Estas cajas representan exones, mientras que las líneas representan regiones intrónicas. Las flechas indican la dirección (o cadena) de la transcripción para cada uno de los genes. Ahora que tenemos nuestro genoma y sus características anotadas, sólo necesitamos las alineaciones de lectura para nuestras cinco muestras.

Utiliza IGV para cargar un archivo de alineamiento para un experimento de RNA-seq (MT1 y MT2)

¿Están los datos en el formato correcto?

No, los archivos a disposición se encuentran en formato **fastq**.

¿Qué tipo de datos se necesitan?

Es necesaria el alineamiento de las lecturas, del cual el resultado es un archivo **.bam**, que es el formato requerido.

¿Por qué hay dos archivos por muestra?

Porque las lecturas crudas vienen de una secuenciación *******paired-end*******.

Alineamiento de lecturas

Se trata de datos de RNA-seq así que necesitamos alineadores especiales para tener en cuenta las isoformas.

Usaremos **HISAT2** para mapear las lecturas de RNA-Seq de nuestras muestras al genoma de referencia. HISAT2 es un programa de alineación rápido y sensible para el mapeo de lecturas de secuenciación de próxima generación (tanto de DNA como de RNA) a una población de genomas humanos (así como a un solo genoma de referencia).

Indexación del genoma

```
# Cargar los modulos necesarios
module load hisat2/2.0.0-beta

# Copiar archivos necesarios a carpeta de trabajo
cd /mnt/Timina/bioinfoII/data/IGV/
cp -r ./ ../../arodriguez/GenBrowser/

# Indexar el genoma
hisat2-build PccAS_v3_genome.fa PccAS_v3_hisat2.idx
```

- `hisat2-build` construye un índice HISAT2 a partir del archivo de genoma de referencia (`PccAS_v3_genome.fa`) y el prefijo que tendrá el conjunto de archivos de salida generados por el mismo (`PccAS_v3_hisat2.idx`):

```
PccAS_v3_hisat2.idx.1.ht2  
PccAS_v3_hisat2.idx.2.ht2  
PccAS_v3_hisat2.idx.3.ht2  
PccAS_v3_hisat2.idx.4.ht2  
PccAS_v3_hisat2.idx.5.ht2  
PccAS_v3_hisat2.idx.6.ht2  
PccAS_v3_hisat2.idx.7.ht2  
PccAS_v3_hisat2.idx.8.ht2
```

Nota: estos ficheros juntos constituyen el índice. Son todo lo que se necesita para alinear las lecturas a esa referencia..