En este repositorio se encuentran varios scripts interconectados para la realización de un análisis de ChIP-seq: chipepe.sh, chip_proc.sh, input_proc.sh, peak_calling.sh y chip.R. Chipepe.sh es el script principal a partir del cual se redirige al resto de los scripts. En él se comienza el análisis de ChIP-seq, con la descarga y asignación de los distintos parámetros requeridos para el estudio, la creación del espacio de trabajo y la construcción del índice del genoma de referencia. Este script redirigirá las muestras inputs y chips a los script input_proc.sh y chip_proc.sh, respectivamente, para el procesamiento de las mismas. El procesamiento de las muestras incluye el análisis de la calidad y el mapeado al genoma de referencia. Los scripts input_proc.sh y chip_proc.sh, una vez completado el procesamiento de las muestras, redirigirán al script peak_calling.sh, script en el cual se analizarán los picos y se realizará una intersección entre picos comunes. Además, se identificarán los motivos de unión a proteínas mediante la herramienta HOMER. Finalmente, este script termina redirigiendo a un script de Rstudio. En este último script se va a realizar la determinación del reguloma, el análisis de la distribución de los picos en el genoma y de los TSS mediante diferentes gráficos y los análisis de enriquecimiento funcional y metabólico.

La ejecución de los scripts requiere únicamente un fichero .txt de parámetros (params.txt). Este fichero de parámetros debe contener la siguiente información:

- La ruta de localización del directorio de trabajo (working_directory), es decir, el lugar donde se va a realizar el análisis.
- La ruta de localización del directorio de instalación (installation_directory), que indica el lugar elegido para la descarga del contenido.
- La ruta de localización del genoma (path_genome), que representa el lugar donde se encuentra el genoma que va a utilizarse en el estudio.
- La ruta de localización de la anotación (path_annotation), que representa el lugar donde se encuentra la anotación del genoma que va a utilizarse en el estudio.
- La ruta de localización de las muestras input (path_input), que indica el lugar donde se encuentran los inputs que van a utilizarse en el análisis.
- La ruta de localización de las muestras chip (path_chip), que indica el lugar donde se encuentran las muestras chips que van a utilizarse en el análisis.
- El nombre que va a dársele al experimento (experiment_name).
- El universo de cromosomas escogido (universe_chromosome).
- El número tanto de inputs (number_of_inputs) como de chips (number_of_chips) que tiene el estudio.
- Si la muestra es o no un factor de transcripción (transcription_factor).

Estos datos son específicos para cada análisis de ChIP-seq, de manera que deberán ser modificados en función del análisis que vaya a realizarse.

Para una mayor claridad, en este repositorio también se incluye un fichero de ejemplo contenido en la carpeta test (params.txt).

Una vez se tenga este fichero y los scripts descargados, simplemente es necesario ejecutar el script chipepe.sh y utilizar el fichero de parámetros como dato de entrada. Ejemplo: ./chipepe.sh test/params.txt

Explicación de los scripts de manera más concreta

chipepe.sh: Al ejecutar chipepe.sh, se llevará a cabo la creación integral del entorno de trabajo, incluyendo la generación de todas las carpetas necesarias, la creación del índice del genoma y la anotación. Posteriormente, se ejecutarán los scripts chip_proc.sh e input_proc.sh.

chip_proc.sh: Este script se encarga del procesamiento de las muestras de chip. Realiza análisis de calidad, el mapeo de las lecturas cortas con el genoma de referencia y genera archivos .sam y .bam. Proporcionamos un script ajustado para un número de muestras (1-9). En caso de tener más muestras, es necesario ajustar el script, específicamente el límite superior del intervalo en la primera estructura de control "if".

input_proc.sh: Este script se encarga del procesamiento de las muestras de input. Realiza análisis de calidad, mapeo de lecturas con el genoma de referencia y genera archivos .sam y .bam. Además, realiza la fusión (merge) de los archivos de input. Proporcionamos un script para un número de muestras entre 1 y 9, y en caso de tener más muestras, es necesario ajustar el script, especialmente el límite superior del intervalo en la primera estructura de control "if". En el último "if", donde se especifica $NUMPROC -eq 2 * $NUMTOTAL, este número debe ser el doble de la suma total de todas las muestras disponibles (combinando chip e input, considerando réplicas) por la duplicación de estos en el peaklist.txt, que es donde se origina el valor de NUMPROC. Por lo tanto, este parámetro también debe ajustarse según las características específicas del estudio.

peak_calling.sh: Este script tiene la responsabilidad de llevar a cabo el proceso de determinación de picos, también conocido como "peak calling". Este procedimiento es habitual tanto para las muestras chip como para las de input. En este contexto, hemos facilitado un script preparado para un rango de muestras que va de 1 a 9. En caso de tener más muestras, se recomienda ajustar el límite superior de este intervalo según sea necesario.

chip.R: Lleva a cabo el análisis matemático-computacional de los datos en R