README

Autores:

- Manuel Anglada Reina
- Gloria Álvarez Alegre
- Andrés Méndez Hurtado
- Pablo Rafael Pombero Hurtado
- Lucía De los Santos Carrascal.

ChipIpe es una herramienta especializada diseñada para el análisis de datos ChIP-seq en *Arabidopsis thaliana*. Este paquete incluye un script principal llamado chipipe.sh, que requiere una serie de parámetros especificados en un archivo .txt cada vez que se ejecuta. Un ejemplo de uso sería:

bash chipipe.sh <archivo_parametros> → *bash chipipe.sh params_chipipe.txt*

Dentro del archivo de ejemplo tarea2_definitiva/test/test_params.txt, se encuentran parámetros cruciales, como el directorio de instalación, el espacio de trabajo para los análisis, el nombre del experimento, el número de réplicas y las rutas hacia el genoma y las anotaciones del organismo en estudio. También se detallan las rutas hacia las muestras específicas de ChIP-seq y de control tipo Input.

- **installation_directory:** el directorio donde hemos instalado el paquete, en nuestro caso /home/bag2023_2/tarea2_definitiva
- **working_directory:** el directorio donde deseamos ubicar el análisis, /home/bag2023_2/tarea2_definitiva
- **experiment_name**: el nombre de las carpetas y los resultados del análisis, prr5_test
- **number_replicas:** el número de réplicas que tienes para las muestras de chip-input, en nuestro ejemplo 1.
- **path_genome**: la ruta necesaria para acceder al genoma del organismo con el que estamos trabajando, que en nuestro caso es *Arabidopsis thaliana* /home/bag2023_2/chipmuestras/chromosome1.fa
- **path_annotation**: la ruta para acceder a las anotaciones para el genoma del organismo con el que estamos trabajando /home/bag2023_2/chipmuestras/chromosome1.gtf
- **path_sample_chip_i:** siendo i un número naural, esta es la ruta necesaria para acceder a los datos de ChIP-seq de la muestra número i /home/bag2023_2/chipmuestras/chip_prr5_chr1.fq. En el caso de tener archivos pareados, deberíamos escribir ambas rutas en la misma fila, separadas por espacio.
- **path_sample_input_i**: igual que en el caso anterior, es la ruta donde se encuentra los datos de entrada relacionados con la muestra i /home/bag2023_2/chipmuestras/input_prr5_chr1.fq.gz .
- **universe_chromosomes**: es el ID, o IDs si hay varios, del cromosoma del organismo que queremos utilizar como universo genético para el enriquecimiento de términos GO y KEGG. Si empleamos más de 1, debe separarse por comas y sin espacios En caso de querer usar todos los cromosomas disponibles, se escribe "all". En nuestro caso, es 1.

- **type_of_peak**: especifica la forma de los picos que estamos buscando. Para picos estrechos, el valor debe ser 1 (en el caso del análisis de unión de factores de transcripción) y 2 para picos anchos (usados para la modificación de histonas). En nuestro Script, está indicado el número 1 para probarlo con nuestro ejemplo de chIP-seq.
- **single_or_paired**: cuando tenemos una lectura de un solo extremo, este parámetro debe ser 1. Para lecturas pareadas, debes escribir 2. En nuestro caso, es 1.
- **tss_upstream:** es el número de bases aguas arriba para definir la región del sitio de inicio de la transcripción (TSS). Indicamos 1000.
- **tss_downstream**: es el número de bases aguas abajo para definir la región del TSS. Este número debe ser positivo, ya que se considera más tarde en el código. Por ejemplo, para un análisis de la región TSS de 1000 bases aguas arriba y abajo (-1000, 1000), debemos escribir 1000 en ambos parámetros **tss_upstream** y **tss_downstream**. Indicamos de nuevo 1000.

Resumen de pasos de chipipe.sh:

1. Carga de Parámetros.
2. Generación del Espacio de Trabajo.
3. Creación del Índice del Genoma de Referencia.
4. Procesamiento de Muestras Individuales.

El último paso se realiza a través de un script auxiliar llamado **sample_proc.sh**. Este script, para cada réplica, lleva a cabo tareas como la carga de parámetros, el control de calidad de las muestras, el mapeo con el genoma de referencia, la conversión de SAM a BAM y de BAM a BAM.BAI ordenados y la llamada de picos.

Una vez procesadas todas las réplicas, se llevan a cabo otros pasos en un tercer script denominado **peak_call.sh**. Se realiza la intersección de los resultados de todas las réplicas y se identifican los motivos, siendo esta tarea realizada con HOMER. Los parámetros de HOMER pueden ajustarse según nuestras preferencias.

Para la visualización y análisis estadístico de los resultados, se utiliza un último script en R llamado **chipipe.R** que realiza las siguientes acciones:

- Carga de Parámetros.
- Definición de Regiones Promotoras.
- Cálculo de Distribución de Picos a lo Largo del Genoma.
- Anotación de Picos según Tipos de Regiones de ADN a las que se Unen.
- Almacenamiento de Picos que se Unen a Regiones Adecuadas.
- Listado de Genes Afectados por el Factor de Transcripción o la Modificación de Histonas (Reguloma).
- Enriquecimiento de Términos GO.
- Enriquecimiento de Términos KEGG.

ChipIpe define el reguloma de manera diferente para picos estrechos y anchos. Para los picos estrechos, el análisis se enfoca en genes en los que el factor de transcripción se une al promotor. En cambio, para los picos anchos, se centra en genes en los que la modificación se une al

promotor, intrones, exones o UTRs. Si el usuario desea considerar diferentes regiones, la personalización del script chipipe.R es posible. Además, este script se puede adaptar para utilizarlo con otros organismos distintos a Arabidopsis, simplemente modificando el archivo txdb y los organismos empleados para el enriquecimiento de términos GO y KEGG.

En cuanto a la salida, ChipIpe crea un directorio que contiene subdirectorios y archivos esenciales para un análisis detallado de los resultados:

- **genome**: Contiene el genoma de referencia utilizado para el análisis y su índice.
- **annotation**: Almacena la anotación de referencia empleada para el análisis.
- **samples**: Incluye un directorio para cada réplica, dividido en chip, input y replica_results.
    - **chip e input**: Contienen archivos BAM ordenados específicos para cada muestra y análisis de calidad con la función FastQC.
    - **replica_results**: Contiene archivos de picos generados por MACS2 para la réplica. Se destaca que, si solo se utiliza una réplica, el archivo narrowPeak o broadPeak se traslada al archivo de resultados y no se encuentra en este directorio.
- **results:** Contiene todos los resultados del análisis. Incluye:
    - **Ficheros de Picos Fusionados de las Réplicas**: Generados iterativamente, con n-1 ficheros para n réplicas. En caso de errores en alguna réplica, como se indica en la salida FastQC o en las estadísticas de alineación Bowtie2, simplemente eliminando los archivos fusionados pertinentes y utilizando Bedtools para intersectar el archivo fusionado anterior con los archivos de picos del resto de las réplicas, se puede ejecutar el script R. Si no se detectan errores, el análisis se realiza con el último archivo fusionado (número más alto).
    - **Motivos Detectados por HOMER**: Se encuentran en este directorio.