

Proyecto Colaborativo de Desarrollo de Software

Nombre del proyecto:

Nina Valley Scan

Responsable principal:

Dra. Alejandra Medina-Rivera y Gabriel Balsa Cantú

Breve planteamiento del problema:

Se desarrolló un nuevo método de predicción de sitios de pegado de factores transcripcionales (TFBS) que pueda utilizar datos de ChIP-Seq actuales.

Este script está basado en trabajo anterior de Stephen A. Ramsey, el cual se tradujo de Matlab a C++ y R, con el propósito de crear una herramienta basada en tecnología de código abierto.

El método busca identificar fenotipos de valle en datos obtenidos por secuenciación y alineación. El input de datos son los picos obtenidos por *bedtools coverage* en cada base par en formato BED.

El input tiene el formato [Chr] [PeakStart] [PeakEnd] [index] [signal], donde index tiene base 1 y marca cada base par desde el inicio del pico hasta el final (e.g.: **chr12 67720 67925 1 1.5**)

El script correrá un filtro de señal, convolución y analizará el fenotipo de la señal para obtener una “calidad de valle” en forma de número decimal.

Objetivo del proyecto:

Obtener el valor de valles en 5 diferentes rangos de bases par (e.g.: 40, 80, 120, 150, 200), los cuales deben ser agrupados en dos reportes (rangos grandes, rangos chicos) en formato BED y compatibles con genome browser, que muestren la región por cromosoma de dichos valles.

Las regiones del cromosoma que presenten una “calidad de valle” alta deberían correlacionarse con sitios de pegado de factores transcripcionales (TFBS).

Datos con los que se cuenta:

Metodología (v.g. RNA-seq):

ChIP-Seq

Plataforma (v.g. Illumina):

Illumina Genome Analyzer Iix, Illumina Genome Analyzer II

Condiciones experimentales (v.g. tejido enfermo, tejido sano): H3K27ac ChIP-seq on human endothelial cell of umbilical vein

Réplicas (v.g. 3 réplicas por condición):

2 réplicas

Controles:

Control ChIP-seq on human endothelial cell of umbilical vein

Información extra sobre los sets de datos:

Los siguientes datos de ENCODE se emplean actualmente para pruebas del script:

Replica: <https://www.encodeproject.org/experiments/ENCSR000ALB/>

Control: <https://www.encodeproject.org/experiments/ENCSR000ALG/>

Tomar un subset de estos datos, un bloque de varios picos del Cromosoma 12, realizando las pruebas en mucho menor tiempo.

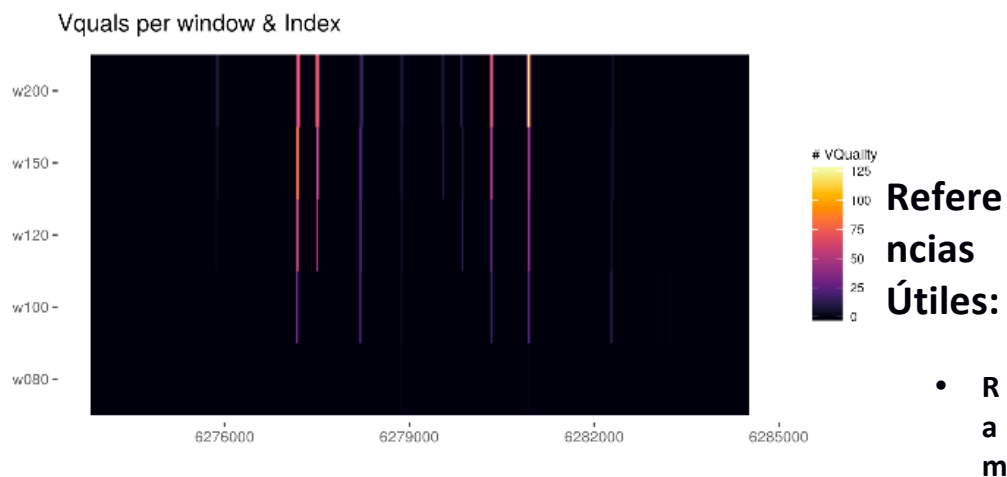
Resultado ideal que debe generar el software:

(v.g. una gráfica comparando tales variables, un archivo con tales columnas)

- Dos archivos de reporte según el rango usado en el análisis (Small, Big) con las columnas: *[chrom, chromStart, chromEnd, Vlog, Vqual]*

chrom	chromStart	chromEnd	Vlog	Vqual
chr12	6274028	6274055	0	1
chr12	6274427	6274536	0.035848	4.998394
chr12	6275013	6275156	0.050023	4.767338
chr12	6275828	6275936	2.082783	38.841927
chr12	6277113	6277250	26.141425	140.554396
chr12	6277454	6277558	24.240254	132.43
chr12	6278155	6278260	9.548239	114.384969
chr12	6278809	6278936	2.314635	34.796328

- Gráficas de calor de los primeros 10 picos analizados



- **Trabajo anterior de Stephen A. Ramsey (Matlab):**
<http://magnet.systemsbiology.net/hac/>
- ENCODE - Bradley Bernstein (Datos Replica):
<https://www.encodeproject.org/experiments/ENCSR000ALB/>
- ENCODE - Bradley Bernstein (Datos Control):
<https://www.encodeproject.org/experiments/ENCSR000ALG/>