

Differential binding

DiffBind, THOR

Dóra Bihary

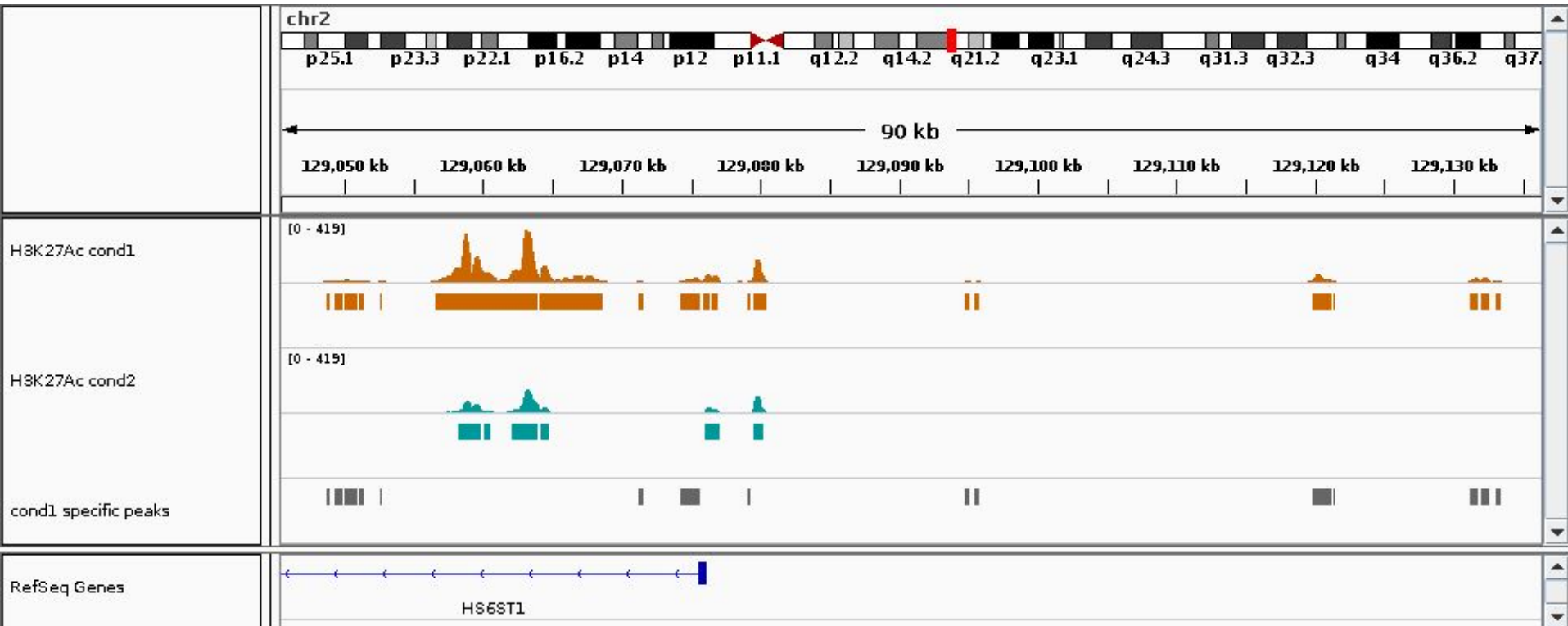
MRC Cancer Unit, University of Cambridge

CRUK CI Bioinformatics Summer School
July 2018

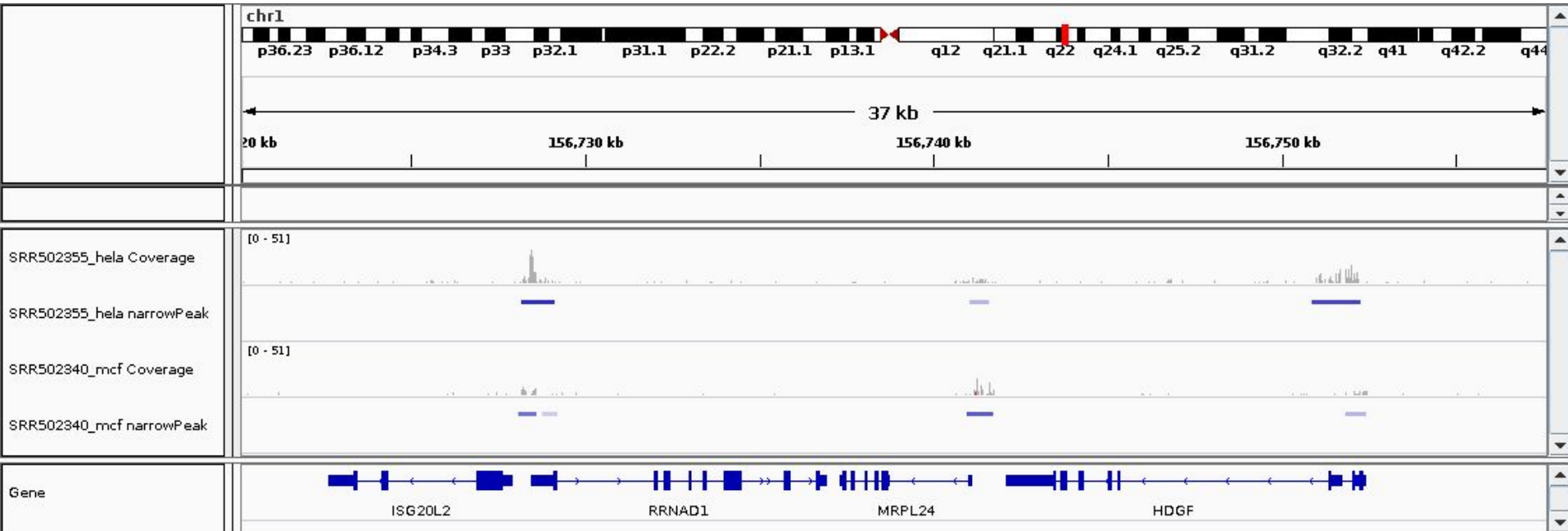
Overview

- Introduction to differential binding
- DiffBind - Bioconductor package
- THOR and ODIN - standalone tools

Peak intersection vs. differential binding



Differential binding



Differential binding

- The aim of differential binding analysis is to compare changes in protein-DNA interactions measured by ChIP-seq
- Two main types:
 - Two-stage methods (DiffBind):
 - Identify candidate peaks using peak callers like MACS2
 - Apply methods tailored for differential expression analysis like DESeq2 and edgeR
 - One-stage methods (THOR):
 - Based on segmentation methods like hidden Markov models (HMM) or sliding window approaches

Overview

- Introduction to differential binding
- DiffBind - Bioconductor package
- THOR and ODIN - standalone tools

DiffBind

- <https://bioconductor.org/packages/release/bioc/html/DiffBind.html>

Main steps:

- Reads in peak sets generated by peak callers like MACS2
- Quality control
- Identifies consensus peak set for further analysis
- Counts reads based on .bam files provided
- Generates affinity matrix: a normalised read count matrix
- Differential binding affinity analysis
- Generates plots and reports results

DiffBind - Differential binding with DESeq2

- DESeq2 is a Bioconductor package for differential expression analysis
- <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>
- It normalises raw read counts for the number of reads in peaks (default) or the total library size
- Calculates dispersion
- Uses negative binomial generalised linear models (GLMs) to model counts
- P-values and differentially expressed genes are calculated by Wald test

DiffBind - Differential binding with edgeR

- edgeR is also a Bioconductor package for differential expression analysis
- <http://bioconductor.org/packages/release/bioc/html/edgeR.html>
- edgeR stands for Empirical analysis of Digital Gene Expression data in R
- It normalises raw read counts for the total library size (default) or the number of reads in peaks using TMM (trimmed mean of M-values)
 - Define gene-wise log-fold changes compared to samples which have expression closest to average
 - Trim highest and lowest expressed genes (genes with higher read counts have lower variance on log scale)
- Builds a model based on your experimental setup
- Calculates dispersion based on an empirical Bayes method
- Uses negative binomial GLMs to identify differentially expressed/bound sites

Overview

- Introduction to differential binding
- DiffBind - Bioconductor package
- THOR and ODIN - standalone tools

ODIN - One-stage Differential peak caller

- Capable of detecting differential peaks (DPs) in pairs of ChIP-seq data
 - Performs
 - Genomic signal processing
 - Peak calling
 - Post processing
 - P-value calculation
1. Main steps of genomic signal processing
 - a. Fragment the DNA into bins, count reads in each bin
 - b. Ignore reads with poor mappability
 - c. Determine fragment size (based on cross correlation)
 - d. Input subtraction
 - e. Normalize based on sequencing depth
 - f. Filter bins with low number of reads

ODIN - One-stage Differential peak caller

2. HMM-based peak calling
 - a. Three state HMM model:
 - i. DP gained in first sample
 - ii. DP gained in second sample
 - iii. Background
3. Post processing
 - a. Ignore DPs with size smaller than the estimated fragment size
 - b. Merge DPs that have distance less than the estimated fragment size
4. P-value calculation

THOR

- Extension of ODIN:
 - Allows the analysis of multiple replicates of two conditions based on a negative binomial distribution
 - Two additional normalisation of ChIP libraries
 - TMM approach
 - Housekeeping genes based normalisation
- Python package
- Inputs:
 - .bam files for the two conditions
 - Chromosome sizes
- Outputs:
 - Post-processed, normalised bigWig files
 - DPs in .bed and .narrowPeak format

Comparison of DiffBind and THOR

- Both tools are capable of handling replicated ChIP-seq peak sets
- The methods used by DiffBind
 - Were originally designed for differential expression analysis on RNA-seq data that assumes that most of the genes between conditions are not differentially expressed - this might not be true for differential binding
- DiffBind is usually more stringent resulting in less DPs
 - This might as well be a limitation in finding true differences between conditions
- It is usually a good practice to use multiple tools and come up with a consensus solution by combining those results
- Other tools you can use for differential binding analysis:
 - SICER
 - MACS2
 - RSEG
 - MAnorm
 - HOMER
 - MMDiff
 - etc.