FindER: A Sensitive Analytical Tool to Study Epigenetic Modifications and Protein-DNA Binding from ChIP-Seq data

M Bilenky[1], S Gakkhar[1], S Jones[1], M Hirst[1,2]

1. *Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada*
2. *Michael Smith Laboratories, Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada*

Abstract

We present a versatile analysis tool developed to Find Enriched Regions (FindER v2.0.0) in ChIP-Seq datasets. FindER is intended to overcome sequence depth limitations of many existing tools and provides a common mechanism for identifying enrichment from localized (e.g. H3K4me3 histone modification, or DNA-protein binding) as well as dispersed (e.g. H3K27me3, H3K36me3) ChIP-Seq signal profiles, or mixed of two signal types (e.g. H3K4me1). After reading aligned IP-signal data and DNA Input control data in the BAM format, FindER computationally segments genome into nucleoseome-scale size bins defined by the local read density distribution in the IP data. Using these adaptive bins FindER first blacklists alignment artifacts (that are typically appear as spurious enriched regions) using Input DNA control (with an option to use external list of blacklisted regions as well).

FindER v2.0.0 uses a novel approach to find enrichment using DNA sequence specific classification of the identified genomic bins into enriched and non- enriched applying Otsu's method widely used in the image processing. Both the local sequence coverage of individual bins and the clustering of the bins along the genome are taken into account to assign significance. The final list of enriched regions is determined by applying a False Discovery Rate control process.

The approach is free from assumptions about underlying distributions of read density and common for different histone modifications. This allows for integrative multi-sample analysis for data with comparable signal-to-noise ratio, and is especially important in examining the effect of relative signal strength on the biology. FindER accepts aligned reads in a standard alignment format as input and generates a list of enriched genomic locations at a given significance (FDR) threshold genome wide. Crucially, FindER is a production grade application. It is a user friendly tool implemented in Java, and it has been tested on terabyte scale ChIP-Seq data. We present examples of application of FindER to the various ChIP-seq data.