Chromatin Gene Matrix: A Proposal for Transforming Chromatin Peak Files into Gene – Centered Matrices

Vinay Swamy1, David M. McGaughey1

1 Ophthalmic Genetics and Visual Function Branch, National Eye Institute, National Institutes of Health, Bethesda, Maryland, USA

Abstract

Introduction

Transcription of genes is influenced through chromatin state changes in the DNA – histone complex and transcription factor (TF) binding. There are several popular techniques for assessing chromatin status and TF binding. Chromatin immunoprecipitation sequencing (ChIP-seq) identifies sites of the genome which interact physically with proteins; often transcription factors and histone modifications. ATAC-seq is used to identify chromatin in an open (and presumably active) state. What these different techniques share is that their primary output are peak files in the bed format. These condensed files are generated from peak callers (e.g. MACS2/SICER/HOMER) and display regions of interest (for ATAC, open chromatin, for ChIP-seq protein-DNA interaction). The bed files can be visualized in web tools like the UCSC Genome Browser and a galaxy of tools exist to annotate and ascribe function (HOMER/GREAT/ChIPSeeker/CHiPpeakAnno). Another powerful ChIP-seq data transformation is collapsing multipledatasets into a chromatin state score with Segway or chromHMM, where each base pair (or window) is given a possible function (e.g. active transcription, poised enhancer, closed chromatin).

A key downside of these approaches is that since they are peak (or genome window) based, they do not allow for direct gene to gene comparison of chromatin status. We propose the Chromatin Gene Matrix (CGM), which transforms peak bed files into gene-centered matrices, where each row *m* is a gene and each column *n* are the peak scores in windows upstream and downstream of the gene, along with exonic and intronic peak scores (Figure 1). Additional information, like gene expression from RNA-seq, can be added as additional columns.

We have written a lightweight tool (<https://github.com/davemcg/CGM>) which takes peak file along with a reference matched gene model (gtf) and in a few minutes can generate a CGM. The only dependencies are bedtools, R and the R package “tidyverse.” The process is inherently parallelizable and can create CGM from dozens of peak files in less than 20 minutes on a laptop computer.