The statistical framework for identifying True REGulatory (TREG) transcription factor binding events by concordance analysis of ChIP-seq and differential gene expression data.

Zhen Hu1, Johannes M Freudenberg1, Siva Sivaganesan2, and Mario Medvedovic§

1Laboratory for Statistical Genomics and Systems Biology, Department of Environmental Health, University of Cincinnati College of Medicine, 3223 Eden Ave. ML 56, Cincinnati OH 45267-0056, USA
2Mathematical Sciences Department, University of Cincinnati, Cincinnati, OH 45221, USA
§Corresponding/Presenting author
Email addresses: medvedm@ucmail.uc.edu

Abstract:

Inferring condition-dependent transcriptional regulation requires assessing the relationship between transcriptional factor (TF) binding events and changes in the expression levels of targeted genes. The complex nature of the transcriptional regulation by TF-binding events, as well as the noise in the TF binding data produced by next-generation sequencing technologies introduces numerous challenges in performing such analyses. The specificity of transcriptional initiation in eukaryotic genomes is maintained through regulatory programs entailing a complex interaction of transcription factors (TF), epigenetic modification of the regulatory DNA regions and associated histones, chromatin-remodeling proteins and the basal transcriptional machinery (Locker, 2001).

High-throughput sequencing of immuno-precipitated DNA fragments (ChIP-seq) provides means to assess genome-wide expression regulatory events (GERE) such as TF-DNA interactions and epigenetic modifications of the chromatin (Park, 2009). Sophisticated statistical methodologies have been developed for identifying GEREs in terms of “peaks” in the distributions of ChIP-seq data (Choi et al. 2009; Ji et al. 2008; Pepke et al. 2009; Rozowsky et al. 2009; Spyrou et al. 2009; Zhang et al. 2008). However, binding of a transcription factor in a gene’s promoter alone does not always result in regulation of its transcription. In the case of highly studied pleitropic regulator ERα, the transcriptional regulation depends on the presence of specific co-factors as well as the type of the activating ligand (Welboren et al. 2009). In general, whether or not ChIP-seq data indicates that a gene’s expression is regulated by the TF is a function of the number of peaks, their intensity and proximity to the transcription start site (TSS). Another analytical challenge in identifying true regulatory TF-gene interactions for pleitropic regulators like ERα is caused by non-genomic regulatory mechanisms (Silva et al. 2010) and transcriptional changes in secondary targets (Bourdeau et al. 2008). Therefore, the identification of true regulatory TF-gene relationships requires effective per-gene summaries/scores measuring the totality of evidence in ChIP-seq data that this is the case, as well as measurements of the gene expression level as a proxy for transcriptional intensity. Integrating information about TF binding from ChIP-seq experiments and changes in gene expression with the goal of identifying true regulatory TF-gene interactions represents another analytical challenge. The simple approach of identifying genes with both significant TF binding summary scores and significant changes in expression is problematic due to well documented problems in selecting optimal significance thresholds.
We developed the comprehensive statistical framework for assessing True REGulatory (TREG) TF-gene interactions by integrated analysis of ChIP-seq and gene expression data. TREG consists of two distinct modules. The first module is a novel two-stage mixture generative statistical model for “peaks” in large genomic regions around genes TSS (1 million bases up and down from TSS). The model is motivated by the two basic assumptions about the nature of gene expression regulation and examination of the distribution of chip-seq data. Our statistical models encode assumption that the number and intensity of “peaks” in ChIP-seq data as well as their proximity to TSS are all positively correlated with the likelihood of a functional TF-gene interaction (Ouyang et al. 2009). By postulating a statistical model we are able to estimate all parameters needed for calculating scores directly from the data. Fitting this two-stage model results in the scores and associated probabilities of regulation based on ChIP-seq data alone. We show that our approach produces effective summaries for both TFs with binding sites clustered in a close proximity of TSS (eg E2F1) and TFs known to exhibit regulation through binding to distant enhancers (eg ERα).

The second TREG module is the integrative statistical model for assessing the overall concordance between probabilities of regulation produced by the first module and probabilities of differential gene expression. Our method provides both, the global assessment of concordance, and gene-level statistical significance of concordances (p-values) which can be used to identify “truly” regulated genes from such integrative analysis. The effectiveness of TREG framework is demonstrated by characterizing primary and secondary ERα targets. Using then the regulatory signatures consisting of probabilities of regulation we dissect the transcriptional program of ER-negative breast tumors into primary ER-targets and secondary proliferation-related genes regulated by E2F1 and MYC genes. Our results also resolve some controversy regarding the role of c-Myc regulated genes in the effects of E2 on proliferation of MCF-7 cells line (Bourdeau et al. 2008). We clearly show the statistically significant concordance between E2 up-regulated genes and the c-Myc binding pattern in ChIP-seq data.

This research was supported by grants from the National Human Genome Research Institute (R01 HG003749), National Library of Medicine (R21 LM009662) and NIEHS Center for Environmental Genetics grant (P30 ES06096).

Reference List