Genome Biology

RESEARCH HIGHLIGHT

An integrated strategy for identification of both sharp and broad peaks from next-generation sequencing data

Weiqun Peng[1]* and Keji Zhao[2]*

See research article: http://genomebiology.com/2011/12/7/r67

Abstract

A novel integrative approach has been developed by Lieb and colleagues for analyzing genome-wide datasets of different chromatin-binding factors and epigenetic states that exhibit both sharp and diffuse signals on the genome.

Keywords: ChIP-Seq; algorithms; SICER; ZINBA; Chromatin States; Epigenetics; epigenome

Transcriptional regulation plays a central role in essential biological processes, including cellular differentiation and responses to cellular and environmental signals. It is becoming increasingly clear that transcription of a large number of genes is coordinately regulated during these processes through transcriptional regulatory networks consisting of three major components: (1) *cis* regulatory elements, including promoters and enhancers; (2) *trans* factors, including transcription factors (TFs) and chromatin modifying enzymes; and (3) chromatin modification states at regulatory regions of genes. Recent development of assays, including ChIP-Seq, DNase-Seq and FAIRE-Seq, which utilize next generation high-throughput sequencing, has propelled the advance in our understanding of these three components of transcriptional regulatory networks. Critical for application of these assays are a number of different algorithms that have been developed to identify signal-enriched regions (or peaks) from these genome-wide datasets, including the latest program reported by Rashid *et al*. in this issue of *Genome Biology* [1].

*Correspondence: wpeng@gwu.edu; zhaok@nhlbi.nih.gov
[1]Department of Physics, The George Washington University, 725 21st Street NW, Washington DC, 20052, USA
[2]Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA

Increasingly numerous datasets of genome-wide profiling of epigenetic modifications and chromatin-binding proteins are being generated. Depending on the distribution pattern of sequencing reads on the genome, there are virtually two kinds of signals: sharp and diffuse signals. While TFs generally recognize specific target motifs, either in enhancers or promoters, and exhibit localized strong signals, the distribution of histone-modification signals ranges from a few nucleosomes to large chromatin domains spanning hundreds of kilobases. For example, H3K4me2 and H3K4me3, which are usually associated with enhancers and promoters, tend to exhibit localized, sharp peaks, whereas H3K27me3, associated with gene silencing, may cover entire chromatin domains [2,3]. On yet larger scales, it is known that H3K9me3 marks heterochromatic domains. In addition to chromatin modifications, some histone-modifying enzymes [4], chromatin remodeling complexes and RNA polymerase II (RNA Pol II) also exhibit extended domains of enrichment. As the cost of sequencing continues to decrease and throughput continues to increase, both at a breathtaking pace, bioinformatic and statistics tools for analyzing genome-wide datasets have become the proverbial bottleneck in garnering results from these assays. Correct and robust delineation of the variety of enrichment patterns from high-throughput sequencing data is essential in all downstream analysis, ranging from annotating the genome, identification of novel target genetic elements or biomarkers, to shedding mechanistic light on specific biological processes.

Although transcriptional regulation involves coordination of a variety of factors that exhibit different binding patterns on the genome, efforts in algorithm development have largely been focused on finding peaks in ChIP-Seq data for identification of TF-binding sites. The first generation of ChIP-Seq peak-finding programs has been summarized and evaluated by Wilbanks and Facciotti [5]. Common in many of these peak-finding programs is a coverage threshold approach based on a variety of statistics, which vary from algorithm to algorithm. Novel

peak finding programs continue to be developed, among which is a topology-based approach that takes into account the shape of the peaks and uses a tree-based statistic for significance determination [6]. Compared with programs suitable for sharp peaks of TF-binding events, computational tools in identifying the diffuse signals spanning large chromatin domains in ChIP-Seq data have been rather limited, due to high noise level, insufficient sequencing coverage, and lack of objective standards for evaluation. The first program designed for identifying such signals was SICER [7]. Motivated by the known mechanisms of domain formation of histone modifications, SICER uses a spatial clustering approach backed by novel statistics for cluster extension and considers the context of enrichment explicitly. It has been successfully applied to ChIP-Seq for histone modifications and chromatin-binding enzymes. Another program, RSEG [8], employs a hidden Markov model approach that takes read mappability into account and provides a statistical approach for domain boundary determination. Although most of these programs provide satisfactory results when applied to datasets with a specific type of peak patterns, particular attention is needed to match the designing feature of the algorithms with the peak pattern of the datasets. When peak-finding programs designed primarily for TF-binding site identification are used on dispersed signals, sensitivity is low and integrity of the domains is compromised. When domain-finding approaches are used on data with sharp peaks, the information on the precise location of peaks may not be fully recovered. Thus, integrative strategies for analyzing the genome-wide datasets of different factors and epigenetic modifications that exhibit both sharp and diffuse signals are needed. ZINBA [1] is among the first to address this need.

ZINBA uses a novel mixture regression approach for its statistical framework. It partitions the genome into non-overlapping windows and each window is probabilistically classified into three components: background, enrichment and zero. The component zero is introduced to account for the many windows with zero tags due to insufficient sequencing coverage. The probability attributed to each component in each window is assigned via an expectation-maximization algorithm, using a negative binomial distribution to parameterize the enrichment and background components. The parameters of this distribution include several covariates, which serve to assist modeling of each component. Known important covariates include read mappability, copy number variations and, in some situations, GC content. As this list suggests, systematic biases of the ChIP-Seq assay are among the most relevant covariates. Indeed, the authors showed using simulated data that inclusion of relevant covariates improves the performance of the algorithm.

This distribution-plus-covariate approach enables incorporation of systematic biases in enrichment identification, a feature especially appealing when the matching control library is not available (for example, FAIRE-seq), or the control library is of poor sequencing coverage. The authors demonstrated that, by considering the non-input covariates, the performance of ZINBA in the absence of an input control library rivals that of the algorithm with an input control library. Additonally, these covariates provide important information for systematic understanding of the compositions of enrichment and background. In general, the ZINBA framework allows any covariate to be included and its relevance estimated, making it easily adaptable to additional covariates or new systems.

Rashid *et al*. [1] evaluated the performance of ZINBA using datasets that cover enrichment length-scales ranging from punctate signals produced by sequence-specific binding events of insulator binding protein CTCF to diffuse signals of the histone modification mark H3K36me3, associated with elongation of RNA Pol II. For punctate signals, ZINBA performs as well as MACS [9], which is one of the most popular peak finders. For diffuse signals such as H3K36me3, ZINBA-designated domains are shown to correlate well with expression. As an example for integrative analysis involving both punctate and diffuse signals, the authors examined the stalled and elongating RNA Pol II. Interestingly, by using ZINBA and a refined definition of stalling score that adjusts for lengths, a significant improvement of the anti-correlation between the stalling score and gene expression is achieved.

ZINBA is a much welcomed addition to the arsenal of bioinformaticians and computational biologists working alongside biologists. Future refinements may focus on resolving or improving the following issues with the method: (1) the assumption of independence of read counts of each window, which is mitigated at an empirical level when merging of enrichment is enforced; and (2) the intensive computational resources needed in model selection, which presumably is due to the need to carry out this operation for each window.

Understanding the diverse modes of functional organization of the genome has been a major theme of genome biology and continues to evolve given the recent developments in what we learn about the three-dimensional structure of genomes [10]. We expect to see even more efforts in development of algorithms like ZINBA, allowing multiple resolutions and integrating diverse data types. As the authors mentioned in the paper, a major factor affecting all ChIP-Seq analysis tools, especially those dealing with diffuse signals, remains the sequencing depth. Most of the libraries displaying dispersed signals for large genomes, such as human and mouse, are far

from saturated. They will be likely to remain so in the near future, even with the continuing increase in sequencing throughput. Yet there is only limited evaluation of methods from the perspective of sequencing coverage. The scaling behavior of analysis methods (that is, how robust the analysis results are as the sequencing depth is changed) should receive more attention in development and evaluation of new methods.

Abbreviation
RNA Pol II, RNA polymerase II; TF, transcription factor.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We thank Brian Abraham for critical reading of the manuscript. This work was supported by the Division of Intramural Research Program of National Heart, Lung and Blood Institute, NIH (KZ).

Published: 25 July 2011

References

1. Rashid N, Giresi P, Sun W, Ibrahim J, Lieb J: **ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions**. *Genome Biol* 2011, **12**:r67
2. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40:**897-903.
3. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nat Biotechnol* 2010, **28:**817-825.
4. Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K: **Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes.** *Cell* 2009, **138:**1019-1031.
5. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-Seq peak detection.** *PLoS One* 2010, **5:**e11471.
6. Hower V, Evans S, Pachter L: **Shape-based peak identification for ChIP-Seq.** *BMC Bioinformatics* 2011, **12:**15.
7. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.** *Bioinformatics* 2009, **25:**1952-1958.
8. Song Q, Smith AD: **Identifying dispersed epigenomic domains from ChIP-Seq data.** *Bioinformatics* 2011, **27:**870-871.
9. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9:**R137.
10. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289-293.

doi:10.1186/gb-2011-12-7-120
Cite this article as: Peng W, Zhao K: An integrated strategy for identification of both sharp and broad peaks from next-generation sequencing data. *Genome Biology* 2011, **12**:120.