BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 492–493
doi:10.1093/bioinformatics/bts722

Genome analysis

Advance Access publication January 7, 2013

BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets

Jianrong Wang[1], Victoria V. Lunyak[2] and I. King Jordan[1,3,*]

[1]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA, [2]Buck Institute for Age Research, Novato, CA 94945, USA and [3]PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia

Associate Editor: Michael Brudno

ABSTRACT

Summary: Although some histone modification chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) signals show abrupt peaks across narrow and specific genomic locations, others have diffuse distributions along chromosomes, and their large contiguous enrichment landscapes are better modeled as broad peaks. Here, we present BroadPeak, an algorithm for the identification of such broad peaks from diffuse ChIP-seq datasets. We show that BroadPeak is a linear time algorithm that requires only two parameters, and we validate its performance on real and simulated histone modification ChIP-seq datasets. BroadPeak calls peaks that are highly coincident with both the underlying ChIP-seq tag count distributions and relevant biological features, such as the gene bodies of actively transcribed genes, and it shows superior overall recall and precision of known broad peaks from simulated datasets.

Availability: The source code and documentations are available at http://jordan.biology.gatech.edu/page/software/broadpeak/.

Contact: king.jordan@biology.gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 22, 2012; revised on November 30, 2012; accepted on December 21, 2012

1 INTRODUCTION

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) technology has been used to produce genome-wide maps for a variety of histone modifications in a number of different cell types (Barski *et al.*, 2007; Ernst *et al.*, 2011; Wang *et al.*, 2008). One of the most critical ChIP-seq data processing steps is peak calling, i.e. the identification of contiguous genomic regions that are significantly enriched with ChIP-seq tags when compared with the genomic background tag count distribution (Park, 2009). Although numerous computational methods that can reliably identify narrow histone modification peaks have been developed (Laajala *et al.*, 2009; Park, 2009), there are fewer methods for calling broad peaks (Song and Smith, 2011; Zang *et al.*, 2009). Here, we describe BroadPeak, a distinct approach for calling broad peaks in diffuse ChIP-seq datasets, and apply the algorithm to both real and simulated datasets to evaluate its performance.

*To whom correspondence should be addressed.

2 METHODS

Algorithm overview: The basic idea of BroadPeak is to assign appropriate positive scores to high-tag sites and negative scores to low-tag sites (i.e. gaps) and then to model the broad peaks as segments with maximal cumulative scores (maximal scoring segments) along chromosomes (Supplementary Figs S1 and S2). This is an adoption of the maximal-segment algorithm, which has most often been applied for sequence comparison (Karlin and Altschul, 1993). In addition to providing an optimal solution for peak calling, the use of the maximal-segment algorithm has the advantages of requiring fewer parameters and only linear-time complexity for computation (Supplementary Methods and Supplementary Fig. S2).

Problem formulation: The genome under consideration is divided into small non-overlapping bins of a user-defined size (e.g. 200 bp), and each bin is assigned with a ChIP-seq tag count. The bins are first classified into high-tag and low-tag bins based on a tag count threshold derived from the standard tag count *Poisson* distribution (which is parameterized by the genomic average bin tag count λ). Each high-tag bin is then assigned with a positive score s_1, and each low-tag bin is assigned with a negative score s_2. The cumulative score from bin i to bin j is the sum of the scores of individual bins between i and j. Maximal scoring segments are segments with maximal cumulative scores, i.e. the cumulative scores will decrease if the segments extend to longer segments or shrink to shorter segments. Thus, identifications of maximal scoring segments are equivalent to setting the boundaries of broad peaks with the locally highest spatial densities of high-tag bins (Supplementary Fig. S2).

Scoring and parameters: The positive and negative scores described earlier in the text (s_1 and s_2) need to be carefully designed to obtain reasonable peaks. Based on the theorems proved by Karlin and Altschul (Karlin and Altschul, 1990), the optimal scoring scheme consists of the log likelihood ratios: $s_1 = ln(p/q)$ and $s_2 = ln[(1-p)/(1-q)]$, where p is the estimated spatial density of high-tag bins in real broad peaks and q is the genomic background spatial density. Thus, p and q are the only parameters needed for BroadPeak. One important feature of this scoring scheme is that, when the segment lengths are large, the spatial densities within the resulted maximal scoring segments will approximate the real target density p (Karlin and Altschul, 1990). This feature theoretically supports the validity of the final identified broad peaks, as their compositions of high-tag bins will resemble real peaks, and it also suggests that the gaps will be adaptively allowed based on the data, namely the target and background densities.

To accurately estimate the target density p, BroadPeak provides two options: supervised and unsupervised estimations (Supplementary Fig. S1). For supervised estimation, the user needs to provide a list of regions that are enriched with broad peaks based on *a priori* knowledge (e.g. highly transcribed gene bodies for H3K36me3 parameter estimation). For unsupervised estimation, BroadPeak first uses a sliding window approach to obtain an initial set of regions showing spatial density changes and model the occurrence of high-tag bins as non-homogeneous Poisson processes with change points (Raftery and Akman, 1986). Conjugate gamma prior distributions are built, and a

Gibbs sampling algorithm (Robert and Casella, 2004) is applied to estimate p and q (Supplementary Methods).

Peak identification: BroadPeak applies the linear time Ruzzo–Tompa algorithm (Ruzzo and Tompa, 1999) to search for all maximal scoring segments (Supplementary Figs S1 and S2). For each maximal scoring segment, the observed spatial density of high-tag bins is compared with the background using a z-test, and only the segments with significantly higher densities ($P<0.05$) are added to the final broad peak list.

3 PERFORMANCE EVALUATION

To evaluate the performance of BroadPeak, we first applied the algorithm to the analysis of ChIP-seq datasets of histone modifications from human CD4$^+$ T cells (Barski *et al.*, 2007). Visual inspection of the called broad peaks shows that they are much consistent with the underlying diffuse ChIP-seq tag count distributions (Fig. 1A and B, Supplementary Fig. S3). For H3K36me3 and H3K79me2, the called broad peaks are also highly coincident with the locations of gene bodies (Fig. 1A), consistent with their known roles as marks of transcriptional elongation (Barski *et al.*, 2007). Closer inspection reveals that the locations of the called peaks for these two marks relative to gene bodies are distinct; H3K79me2 peaks are enriched at and downstream of transcriptional start sites, whereas H3K36me3 peaks are enriched upstream and at transcriptional termination sites (Fig. 1C). These results are also consistent with the known biological roles of these modifications.

Additional biologically relevant results can be seen for the called H3K27me3 peaks (Fig. 1B). H3K27me3 is thought to mark repressive chromatin domains, and the edges of the broad peaks called for this mark are enriched for binding sites of the CTCF protein (Fig. 1B and D, Supplementary Fig. S4), which is known to be related to the activity of chromatin insulators and barriers (Cuddapah *et al.*, 2009). Weaker enrichments are also observed for called broad peaks of the H3K9me3 (Supplementary Fig. S5). Similar performances are shown when the algorithm is run in supervised and unsupervised modes for H3K36me3 (Supplementary Methods and Supplementary Fig. S6).

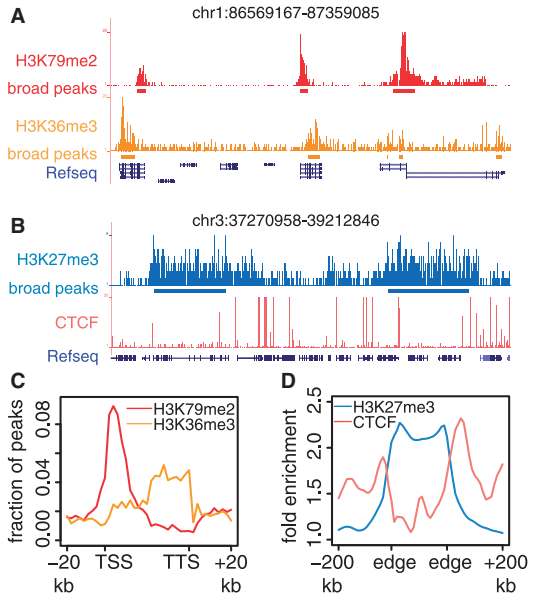

Fig. 1. Performance evaluation of BroadPeak. (**A**) Examples of broad peaks of H3K79me2 and H3K36me3 identified by BroadPeak compared with the underlying ChIP-seq tag counts. (**B**) Examples of broad peaks of H3K27me3 identified by BroadPeak and CTCF binding profiles. (**C**) Relative distributions of broad peaks of H3K79me2 and H3K36me3 around gene bodies. (**D**) Enrichments of CTCF bindings around the edges of large H3K27me3 broad peaks (>200 kb)

We further evaluate the performance of BroadPeak using three separate simulated ChIP-seq tag datasets, which correspond to a set of predefined (i.e. known) broad peaks (Supplementary Methods). This approach was used to allow for a controlled comparison of results obtained from BroadPeak versus results from the existing methods MACS, SICER and RSEG (Song and Smith, 2011; Zang *et al.*, 2009; Zhang *et al.*, 2008). For each method tested, the ability to identify known broad peaks from simulated tag datasets was quantified using precision and recall. For all three simulated datasets, BroadPeak achieves substantial improvements on recall (Supplementary Table S1) while maintaining good precision (albeit slightly lower than seen for MACS and SICER). Overall, BroadPeak shows the highest value for the harmonic mean of recall and precision (F score). Finally, BroadPeak appears particularly well suited for calling large peaks as can be seen from the more contiguous characterization of known large peaks from the simulated datasets (Supplementary Fig. S7) and the relatively broader distribution of called peak sizes (Supplementary Fig. S8).

Funding: Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [BR-4839 to J.W. and I.K.J.]; Buck Institute Trust Fund (to V.V.L.).

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Cuddapah,S. *et al.* (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90**, 5873–5877.

Laajala,T.D. *et al.* (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**, 618.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

Raftery,A.E. and Akman,V.E. (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika*, **73**, 85–89.

Robert,C.P. and Casella,G. (2004) Monte Carlo statistical methods. In *Springer Texts in Statistics*. Springer, New York, pp. 454–455.

Ruzzo,W.L. and Tompa,M. (1999) A linear time algorithm for finding all maximal scoring subsequences. *Proc. 7th Int. Conf. Intell. Syst. Mol. Biol.*, 234–241.

Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human xgenome. *Nat. Genet.*, **40**, 897–903.

Zang,C. *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.