Bioinformatics Advance Access published August 12, 2013

BIOINFORMATICS APPLICATIONS NOTE 2013, pages 1–2
doi:10.1093/bioinformatics/btt428

Genome analysis

Advance Access publication August 8, 2013

ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq

Alper Kucukural[1,*,†], Hakan Özadam[1,†], Guramrit Singh[1], Melissa J. Moore[1] and Can Cenik[1,2,*]

[1]Department of Biochemistry and Molecular Pharmacology, Howard Hughes Medical Institute, University of Massachusetts Medical School, Worcester, MA 01605 and [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Associate Editor: Michael Brudno

ABSTRACT

Summary: Unlike DNA, RNA abundances can vary over several orders of magnitude. Thus, identification of RNA–protein binding sites from high-throughput sequencing data presents unique challenges. Although peak identification in ChIP-Seq data has been extensively explored, there are few bioinformatics tools tailored for peak calling on analogous datasets for RNA-binding proteins. Here we describe ASPeak (abundance sensitive peak detection algorithm), an implementation of an algorithm that we previously applied to detect peaks in exon junction complex RNA immunoprecipitation in tandem experiments. Our peak detection algorithm yields stringent and robust target sets enabling sensitive motif finding and downstream functional analyses.

Availability: ASPeak is implemented in Perl as a complete pipeline that takes bedGraph files as input. ASPeak implementation is freely available at https://sourceforge.net/projects/as-peak under the GNU General Public License. ASPeak can be run on a personal computer, yet is designed to be easily parallelizable. ASPeak can also run on high performance computing clusters providing efficient speedup. The documentation and user manual can be obtained from http://master.dl.sourceforge.net/project/as-peak/manual.pdf.

Contact: alper.kucukural@umassmed.edu or ccenik@stanford.edu

Received on March 8, 2013; revised on July 17, 2013; accepted on July 20, 2013

1 INTRODUCTION

High-throughput sequencing of short RNA fragments directly associated with RNA-binding proteins enables transcriptome-wide mapping of protein binding sites on RNAs. Such binding sites can be enriched by either immunopurification (IP) of RNA–protein complexes (RIP) or cross-linking followed by IP (CLIP) (König *et al.*, 2012). These approaches are analogous to ChIP experiments for DNA-binding proteins. Although identification of peaks in ChIP-Seq data has been addressed by numerous approaches (Ji *et al.*, 2008; Rashid *et al.*, 2011; Rozowsky *et al.*, 2009; Zhang *et al.*, 2008), development of specific bioinformatics tools for RNA–protein binding site identification has lagged behind with few alternatives (Kishore *et al.*, 2011, Li *et al.*, 2013, Uren *et al.*, 2012). With the recent explosion of high-throughput sequencing data for dozens of RNA-binding proteins, there is an urgent need to develop efficient and user-friendly applications to better address the computational challenges of identifying peaks in sequences with variable starting abundances. Here, we describe a new peak detection algorithm that is sensitive to differential expression levels of target transcripts. This approach enables robust peak detection even in low abundance transcripts. We previously used our algorithm to define binding sites for the exon junction complex (EJC) (Singh *et al.*, 2012). Here we describe an open-source and a much-improved implementation of this abundance sensitive peak detection algorithm (ASPeak).

2 IMPLEMENTATION AND ALGORITHM

RIP and CLIP are enrichment strategies where IP with an RNA-binding protein-specific antibody leads to efficient and specific pull-down of target RNAs. Consider an example where an RIP protocol enriches the targets of an RNA-binding protein by 1000-fold. Let us assume that there are two RNA species with different abundances: RNA-X at 1 molecule/cell and RNA-Y at 1000 molecules/cell. If we assume that an RNA-binding protein binds all molecules of RNA-X and 1 molecule/cell of RNA-Y, we expect to obtain an equal number of sequence fragments from both RNA-X and RNA-Y after IP. Consistent with this, we previously observed a Spearman correlation of 0.87 between RIP-Seq and RNA-Seq read numbers for individual genes (Fig. 1A, Singh *et al.*, 2012).

For accurate peak calling in transcripts of varying abundance, it is essential to consider the dependence between background levels and expression. Consequently, the observed number of sequencing reads in a RIP-Seq library is expected to correlate significantly with the parallel RNA-Seq library. The user is free to choose between RIP-input or any other appropriate RNA-Seq data to estimate expression levels. For simplicity, we used RNA-Seq as our expression measure.

Our algorithm computes expression-sensitive backgrounds with respect to user-defined genomic intervals. Most commonly, users are interested in peaks within specific transcripts. In this case, the intervals can be from an annotation source such as RefSeq, and can include coding exons, 3′ UTRs, 5′UTRs and introns. Additionally, intergenic regions or any interval discovered using *ab initio* methods such as Cufflinks (Roberts

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

et al., 2011) can be used. Specifically, we modeled read counts for each interval as a negative binomial (NB) distribution, parameterized by interval-specific p and r.

To estimate p and r for each interval, we used the method of moments approach as described in the user manual.

Peaks on each interval were detected using its specific p and r and an NB test whose distribution function is given by

$$F(k,p,r)=\sum_{i=0}^{k}\binom{r+k-1}{k}p^r(1-p)^k$$

where k is the number of observed reads. This test is applied to each nucleotide such that the probability of observing k or more reads at any given position is $1-F(k,p,r)$. If a $P<10^{-2}$ is found, the next L nucleotides are considered for inclusion into the peak. We considered only the center of each aligned read. Consequently, read counts from distinct nucleotide positions are modeled as independent random variables. The P-value upon peak extension is calculated as a sum of N NB distributions as background.

We required the extended peak P-value to be also $<10^{-2}$. When no such extension is possible, a final P-value for the entire peak is calculated using all contiguous positions (see User Manual). For intervals with no detectable expression in the RNA-Seq data, a local window approach is used. Three nucleotide windows (default: 1k, 5k and 10k, as in Zhang *et al.*, 2008) centered on the position of interest were used to estimate the NB distribution parameters p and r. The combination that maximizes the expected value of the distribution was used to call the peaks as above (see User Manual for details).

3 CASE STUDY

To compare the performance of ASPeak, we used two popular alternative peak calling approaches [MACS v.1.4.1 (Zhang *et al.*, 2008) and Piranha v.1.2.0 (Uren *et al.*, 2012)]. Ideally, we would want to test the performance on a gold standard dataset. However, no such information is currently available for either RIP-Seq or CLIP-Seq. Therefore, we relied on prior biological knowledge about one RNA-binding protein complex, EJC, for evaluation. Specifically, previous work showed that the multi-protein EJC is deposited ~24 nt upstream of exon–exon junctions. Recently, Singh *et al.* (2012) and Saulière *et al.* (2012) generated both RIP-Seq and CLIP-Seq data for EJC. We called peaks using ASPeak, MACS or Piranha and compared the fraction of called peaks that fall into the expected position of ~24 nt upstream of the exon–exon junction. We found that ASPeak reported a higher fraction of such peaks overall despite strong overlap between the three programs. We additionally tested performance on two RIP-Seq and two CLIP-Seq data (see User Manual).

4 CONCLUSION

ASPeak is a fast and efficient expression-sensitive peak caller for CLIP- and RIP-Seq data. ASPeak implementation has the ability to run on multiple processors resulting in a significant speedup when used on high performance computing centers. Compared with MACS or Piranha, ASPeak provides comparable or better results when used in an expression-sensitive mode. In addition, ASPeak is supported with extensive documentation that allows experienced bioinformaticians to customize their analyses using detailed parameter files. Finally, a streamlined mode is available for biologists with limited knowledge of Linux systems to easily run ASPeak using default settings.

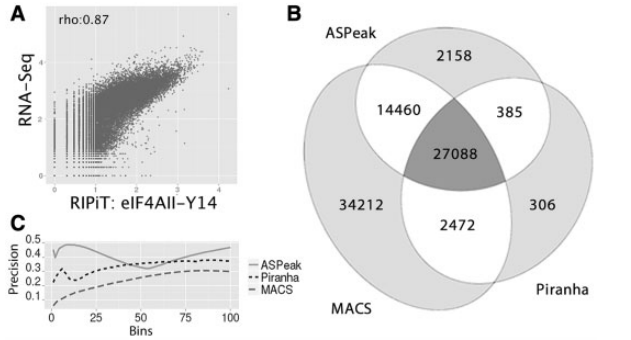

Fig. 1. (**A**) Scatter plot for RNA-Seq versus RIPiT data (Singh *et al.*, 2012). Each data point indicates the log10 of the number of reads mapped to a transcript. (**B**) ASPeak, MACS 1.4.1 or Piranha v.1.2.0 was used to call peaks as detailed in the user manual. A Venn diagram showing the overlap between three approaches with respect to the number of peaks intersecting the exon junction region (15–35 nt upstream of the exon–intron boundary) was shown. (**C**) We ranked called peaks by program-specific scores from the highest to lowest confidence. Then, we separated each list into 100 equal size bins and plotted the ratio of EJC peaks to all peaks (precision) in each bin cummulatively

Our implementation, ASPeak, is an open-source command-line program. Input files are RNA-Seq, CLIP-/ RIP-Seq data in BED/SAM/BAM/BOWTIE format and region annotation files such as RefSeq in BED format. Peak regions are output in a tab-separated format. ASPeak can be run using a single command in a streamlined and simple manner. Additionaly, ASPeak can be parallelized and allows advanced users to run different steps separately for effective debugging and specialized applications.

Funding: NIH grant # GM53007 (to M.J.M.). M.J.M. is an HHMI Investigator.

Conflict of Interest: none declared.

REFERENCES

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Kishore,S. *et al.* (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.

König,J. *et al.* (2012) Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.*, **13**, 77–83.

Li,Y. *et al.* (2013) RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments. *Nucleic Acids Res.*, **41**, e94.

Rashid,N.U. *et al.* (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.

Roberts,A. *et al.* (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.

Rozowsky,J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **1**, 66–75.

Saulière,J. *et al.* (2012) CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat. Struct. Mol. Biol.*, **19**, 1124–1131.

Singh,G. *et al.* (2012) The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell*, **151**, 750–764.

Uren,P.J. *et al.* (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, **28**, 3013–3020.

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