Bioinformatics, 33(13), 2017, 2029–2031
doi: 10.1093/bioinformatics/btx087
Advance Access Publication Date: 22 February 2017
Applications Note

OXFORD

Genome analysis

MotifMap-RNA: a genome-wide map of RBP binding sites

Yu Liu[1], Sha Sun[2], Timothy Bredy[3], Marcelo Wood[3], Robert C. Spitale[4,*] and Pierre Baldi[1,*]

[1]Department of Computer Science and Institute for Genomics and Bioinformatics, [2]Department of Developmental and Cell Biology, [3]Department of Neurobiology and Behavior and [4]Department of Pharmaceutical Sciences and the Chao Family Comprehensive Cancer Center, University of California, Irvine, CA 92697, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 24, 2016; revised on January 20, 2017; editorial decision on February 6, 2017; accepted on February 7, 2017

Abstract

Motivation: RNA plays a critical role in gene expression and its regulation. RNA binding proteins (RBPs), in turn, are important regulators of RNA. Thanks to the availability of large scale data for RBP binding motifs and *in vivo* binding sites results in the form of eCLIP experiments, it is now possible to computationally predict RBP binding sites across the whole genome.
Results: We describe MotifMap-RNA, an extension of MotifMap which predicts binding sites for RBP motifs across human and mouse genomes and allows large scale querying of predicted binding sites.
Availability and Implementation: The data and corresponding web server are available from: http://motifmap-rna.ics.uci.edu/ as part of the MotifMap web portal.
Contact: rspitale@uci.edu or pfbaldi@uci.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA serves not only as a messenger between DNA and protein, but also as a regulator of important processes such as genome organization and gene expression (Morris and Mattick, 2014). RNA itself is regulated by a diverse collection of RNA binding proteins (RBPs), which are responsible for an array of functions such as alternative splicing, RNA modification, polyadenylation, mRNA transport and translational regulation (Glisovic *et al.*, 2008; Mercer *et al.*, 2009). RBPs typically bind to their targets via one or more RNA binding domains (RBDs) which are thought to have specific binding motifs (Lunde *et al.*, 2007). Due to the large number of known and predicted RBPs and their important role in RNA regulation, there has been much interest in systematically understanding their behavior.

Recently, large scale *in vivo* surveys have been carried out to discover the binding motifs of a large number of RBPs (Cook *et al.*, 2011; Ray *et al.*, 2013). At the same time, high throughput *in vivo* eCLIP experiments have been effective in identifying the RBP bindings to RNAs in human immortalized and primary cells (Tollervey *et al.*, 2011; Van Nostrand *et al.*, 2016). Together, these biological data provide a foundation for systematically predicting RBP binding sites across the whole genome and validating them. Previously computational methods have been described to predict motif specific RBP binding sites for a given sequence or a range of sequences (Paz *et al.*, 2014; Zhang *et al.*, 2013). However, to our knowledge, there is no service that allows systematic, genome-wide binding site querying.

Here we describe MotifMap-RNA, a novel extension of MotifMap (Daily *et al.*, 2011; Xie *et al.*, 2009), a system for transcription factors binding site prediction, to RBP binding sites. MotifMap-RNA predicts *z*-score based binding sites specific to RBP motifs across the human and mouse genomes. It also allows the user to filter and sort the results based on clustering of local binding sites, represented by weighted *z*-scores, or evolutionary conservation, quantified by Bayesian branch length scores (BBLS). Furthermore, we organized genomic sequences into 4 major classes: UTRs, intronic regions, lncRNAs and miRNAs, for all of which we generated class specific model parameters. Finally, we implemented a web server which allows the user to interact with MotifMap-RNA results through a friendly interface.

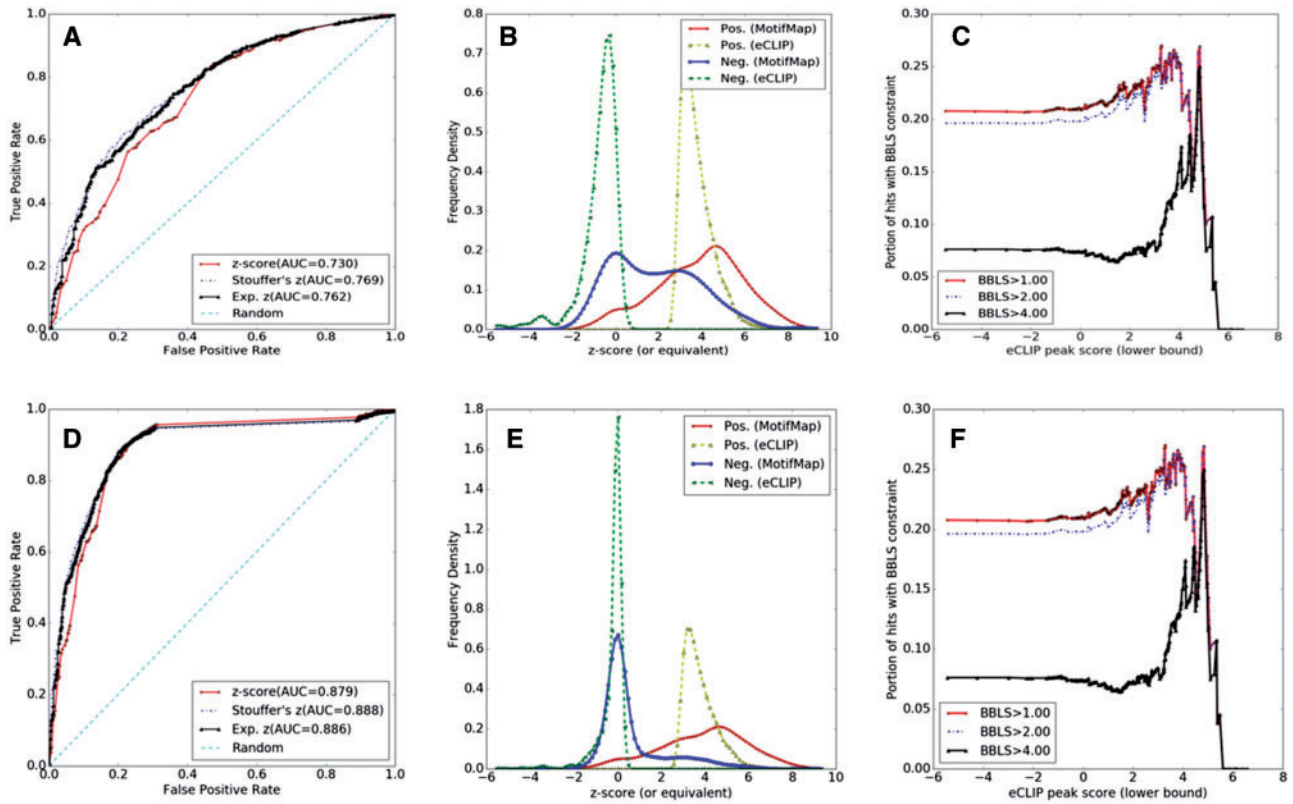

Fig. 1. (**A**) ROC curve of a representative RBP (HNRNPK) from 3 types of *z*-scores: raw, exponentially weighted *z*-score and Stouffer's *z*-scores using rank as weights (details in supplement). Results show that aggregating *z*-scores improves the AUC performance. (**B**) Distribution of ground truth peak scores for both negative and positive examples used in ROC calculation, and their corresponding aggregate *z*-scores from MotifMap (exponentially weighted, corresponds to black solid line in A). (**C**) Portion of MotifMap hits with BBLS scores greater or equal to the marked thresholds from hits corresponding to (positive) peaks at different peak score cutoffs. This shows the relative amount of highly conserved hits increases as the peak score increases, i.e. the most positively enriched peaks tend to overlap highly conserved MotifMap hits. However, the total portion of conserved hits is low (<0.3). (**D**). Same as (A) but with random sequences added as negative examples. This shows that random sequences improves the AUC performance but retains the same trend. (**E**). Same as (B) but with random sequences as negatives. MotifMap-RNA is able to identity random negatives effectively. (**F**). Same as (C) but with random sequences as negatives. BBLS score is not sensitive to the addition of random negatives, since high BBLS scores are concentrated on a portion of the most positive peaks (Color version of this figure is available at *Bioinformatics* online.)

2 Materials and methods

2.1 Motif and genomic data collection

We obtained experimental and computational RBP binding motif data, in the form of positional weight matrices (PWMs), from RBPDB and CISBP (Cook *et al.*, 2011; Ray *et al.*, 2013). In total, we curated 371 PWMs, 266 of which are from human, 22 from mouse and 83 from other sources. We estimated that these motifs correspond to approximately 235 unique RBPs. We also downloaded the latest human and mouse genome assemblies (hg38, mm10) and their multiple species alignments from the UCSC genomic browser (http://genome.ucsc.edu/). We filtered the genomic sequences into 4 classes by annotations: untranslated regions (UTRs), intronic regions, long non-coding RNA (lncRNA) and miRNA (sources in supplement).

2.2 Scoring

We scanned each class of filtered genomic sequences using curated motifs, calculating *z*-scores with class-specific mean and variance to achieve better specificity. For each sequence in a class (e.g. 5'UTR of a particular gene), we filtered the top scoring motif binding sites in terms of *z*-scores (only positive *z*-scores were considered), up to 3 on each strand. These were considered hits from the motif. Per sequence hits were chosen over hits with the highest absolute *z*-scores to maximize the coverage across the entire genome.

We also incorporated additional metrics to measure the hits. Some RBP bindings tend to be locally clustered (Ule *et al.*, 2006). As such, Two forms of weighted *z*-scores were used to reflect local clustering of high *z*-score hits from the same motif. In addition, RBP binding sites can be less conserved than TF binding sites (Gerstberger *et al.*, 2014; Vaquerizas *et al.*, 2009). Conservation scores in the form of BBLS were also generated using method described in the original MotifMap (Xie *et al.*, 2009). Details about scoring and filtering of hits are described in the Supplementary Material.

3 Results

Overall we generated binding predictions for 371 motifs in 4 classes of human and mouse genomic sequences. The total number of hits is typically between 100 000 and 200 000. While the amount of hits can be enormous due to the short and degenerate nature of some motifs, which may produce lower quality hits, the user can effectively filter out a small set of hits of interest using a combination of aforementioned metrics through the web portal.

3.1 Validating the quality of *z*-score and BBLS hits

In order to validate the quality of the predictions, we downloaded eCLIP results for 12 RBPs from the ENCODE project (https://www.encodeproject.org/; Van Nostrand *et al.*, 2016) and generated ROC

curves for matching MotifMap-RNA results. As an example, UTR results from the RBP HNRNPK are shown in Figure 1A–C (from HepG2 tissue, Replicate 1). Notably, aggregating *z*-scores improves the AUC performance while high BBLS scores tend to concentrate on highly positive peaks.

Due to the fact that many eCLIP results lack sufficient negative examples for ROC curve estimation, we included random sequences not overlapping any positive eCLIP peaks as extra negatives. Their effect on HNRNPK results are shown in Figure 1D–F. AUC performance generally improves while BBLS performance remains consistent.

Overall, with random sequences added, we obtained an average AUC of 0.76 for *z*-scores in the UTR region, and 0.68 for lncRNA region. For details on the validation method, see the supplement.

Additionally, Fisher's exact test was applied to MotifMap-RNA results which overlap positive or negative peaks. In all tested cases, MotifMap-RNA hits significantly overlap more positive peaks. Comparison to existing method (RBPmap) also shows favorable results (details in Supplementary Tables S1 and S2).

3.2 Web server

We constructed a database to host the results and implemented the MotifMap-RNA web portal, which provides the user a friendly interface to effectively find, filter, sort and navigate the binding site results in two different modes: motif search and gene search. In motif search, the user can obtain an interactive table containing results from all of the hits of the selected motif, filtered and sorted by a variety of parameters. In gene search, instead of selecting one motif, the user can input a gene symbol or an annotation ID (e.g. miRNA accession), and search for hits from all motifs to that target (details in Supplementary Material).

4 Conclusion

In conclusion, MotifMap-RNA is a novel system for genome-wide querying of RBP binding sites. Together with its friendly interface, it will assist users in their investigations of RBPs and RNA regulation, and the fundamental roles they play across multiple biological processes.

Conflict of Interest: none declared.

References

Cook,K.B. *et al.* (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, 1–8.

Daily,K. *et al.* (2011) MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics*, **12**, 495.

Gerstberger,S. *et al.* (2014) Evolutionary Conservation and Expression of Human RNA-Binding Proteins and Their Role in Human Genetic Disease. In: Yeo,G.W. (ed.), *Advances in Experimental Medicine and Biology*, Vol. **825**. Springer, New York, NY. pp. 1–55.

Glisovic,T. *et al.* (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.

Lunde,B.M. *et al.* (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.

Mercer,T.R. *et al.* (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.

Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev.. Genet.*, **15**, 423–437.

Paz,I. *et al.* (2014) RBPmap: A web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **42**, 1–7.

Ray,D. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.

Tollervey,J.R. *et al.* (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.*, **14**, 452–458.

Ule,J. *et al.* (2006) An RNA map predicting Nova-dependent splicing regulation. *Nature*, **444**, 580–586.

Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 1–9.

Vaquerizas,J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Xie,X. *et al.* (2009) MotifMap: A human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, **25**, 167–174.

Zhang,C. *et al.* (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res.*, **41**, 6793–6807.