Bioinformatics, 31(22), 2015, 3703–3705
doi: 10.1093/bioinformatics/btv442
Advance Access Publication Date: 29 July 2015
Applications Note

OXFORD

Gene expression

BackCLIP: a tool to identify common background presence in PAR-CLIP datasets

P. H. Reyes-Herrera[1,*,†], C. A. Speck-Hernandez[2,†], C. A. Sierra[2] and S. Herrera[3,4]

[1]Colombian Corporation for Agricultural Research (CORPOICA), 250047 Bogotá, Colombia [2]Universidad Antonio Nariño, 110311 Bogotá, Colombia, [3]Woods Hole Oceanographic Institution, 02543 Massachusetts, USA and [4]Massachusetts Institute of Technology, 02139 Massachusetts, USA

*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.
Associate Editor: Ziv Bar-Joseph

Received on February 17, 2015; revised on June 28, 2015; accepted on July 19, 2015

Abstract

Motivation: PAR-CLIP, a CLIP-seq protocol, derives a transcriptome wide set of binding sites for RNA-binding proteins. Even though the protocol uses stringent washing to remove experimental noise, some of it remains. A recent study measured three sets of non-specific RNA backgrounds which are present in several PAR-CLIP datasets. However, a tool to identify the presence of common background in PAR-CLIP datasets is not yet available.
Results: We used the measured sets of non-specific RNA backgrounds to build a common background set. Each element from the common background set has a score that reflects its presence in several PAR-CLIP datasets. We present a tool that uses this score to identify the amount of common backgrounds present in a PAR-CLIP dataset, and we provide the user the option to use or remove it. We used the proposed strategy in 30 PAR-CLIP datasets from nine proteins. It is possible to identify the presence of common backgrounds in a dataset and identify differences in datasets for the same protein. This method is the first step in the process of completely removing such backgrounds.
Availability: The tool was implemented in python. The common background set and the supplementary data are available at https://github.com/phrh/BackCLIP.
Contact: phreyes@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-binding proteins (RBPs) have important roles in RNA regulation. The first step to understand RBPs' specific functions is to identify the RNA targets for each RBP. The introduction of protocols combining CLIP (UV crosslinking and immunoprecipitation) and high-throughput sequencing (commonly known as CLIP-seq protocols) have made it possible to obtain sets of binding sites for RBPs at a transcriptome-wide scale (Licatalosi *et al.*, 2008). However, each CLIP-seq protocol introduces distinct modifications to reduce the presence of background (non-crosslinked RNA).

PAR-CLIP, a frequently used CLIP-seq protocol, uses photo activatable nucleosides to label the transcripts in addition to an enhanced crosslinking (Hafner *et al.*, 2010). These modifications induce specific nucleotides transitions that facilitate the recognition of the crosslinked sites. The presence of a common background in PAR-CLIP datasets has been noted (Sievers *et al.*, 2012). This non-specific RNA background must be taken into account when processing PAR-CLIP data because it can interfere with the distinction of the specific characteristics recognized by the RBPs, and therefore the identification and understanding of binding targets and protein function.

A recent study (Friedersdorf and Keene, 2014) experimentally measured three background sets and demonstrated that background RNA is common in several PAR-CLIP datasets. This background RNA mainly originates from false binding sites. It is worth noting

that PAR-CLIP induced transitions were also present in several sites from the measured background sets, thus it is difficult to distinguish background RNA.

Although computational tools exist specifically for CLIP-seq data (Reyes-Herrera and Ficarra, 2014), only a few proposals address the background issue (Comoglio *et al.*, 2015; Uren *et al.*, 2012; Wang *et al.*, 2014). These computational tools use mathematical models to distinguish binding sites from the RNA background based on characteristics such as read counts or the number of induced transitions, but these characteristics are measured from whole datasets (both background and binding sites). Friedersdorf and Keene (2014) present an alternative strategy for background correction, it consists on removing sites from the PAR-CLIP dataset that overlap by one or more nucleotides with the sites present in one of three known background RNA sets. However, a limitation of this strategy is that it may remove sites that incidentally overlap with the background.

The datasets of non-specific RNA background constitute a valuable reference resource for the quantification of the amount of background present in a PAR-CLIP dataset. Here, we build on this resource to develop a computational tool, BackCLIP, to identify the presence of common background RNA in PAR-CLIP datasets.

2 Materials and Methods

2.1 Common background

We used the three background RNA sets (45, 35 and 20 kDa) obtained in (Friedersdorf and Keene, 2014) to build an initial background set, as illustrated in Figure 1(a). Each site in the initial background set was assigned a score, *s*, equal to the number of the background sets that contained the specific site.

Then we used several PAR-CLIP datasets to refine the scores and build a common background set as illustrated in Figure 1(b). First, we found common sites between the initial background set and a PAR-CLIP dataset by using Pybedtools (Dale *et al.*, 2011). Then, we increased the scores of the sites present in both sets and updated the common background set. The score is directly associated with the number of datasets that contain a site present in the common background set.

We used data publicly available to build and test the common background. These datasets have several similarities in experimental conditions. All data are from human cell lines (the majority from HEK-293); 4-SU is the photoactivatable ribonucleoside used to derive most of the datasets. We used a random partition to select the datasets used to build the common background and the datasets used for testing. We reviewed several considerations to build a common background that is representative of several PAR-CLIP datasets and an unbiased test set.

We used the aforementioned approach and built the common background set using 19 PAR-CLIP datasets from seven RBP. This common background was used to quantify the background RNA in 30 PAR-CLIP datasets from nine proteins (details in Supplementary Data). Two common background sets were identified: (i) the common background set used to test this proposal (built on 19 PAR-CLIP datasets from seven RBP), and (ii) a common background set built on all the used PAR-CLIP datasets (built on 49 PAR-CLIP datasets from 16 RBP).

2.2 Measure background presence

We propose to use the common background set to quantify the amount of background RNA present in any PAR-CLIP dataset, as illustrated in Figure 1(c). This method suggests a threshold for the scores, which is the 99 percentile learned from the common background scores (discarding sites with a score equal to 1). To find the intersection between a dataset and the common background, we obtained the proportion of sites with a score higher than the threshold in the intersection compared with the number of sites in the intersection. Then we examined two indicators (i) the aforementioned proportion and (ii) the number of sites in the intersection set compared with the number of sites in the PAR-CLIP dataset. We use these two parameters as a quantitative measure of the background presence. The first indicator shows the amount of common background in the intersection. The second indicator provides the number of intersection sites in the evaluated dataset. If the first indicator is >50% and the second indicator is >25% then, the presence of background in the dataset must be considered. The user is then alerted and asked whether the sites with a score higher than the threshold can be removed.

Using Tophat (Trapnell *et al.*, 2009) and Bowtie (Langmead *et al.*, 2009), the examined raw PAR-CLIP datasets were aligned to hg19 to obtain clusters of overlapping sequences. We excluded reads of <20 nt and clusters with less than five reads, and used the clusters to measure the background RNA presence.

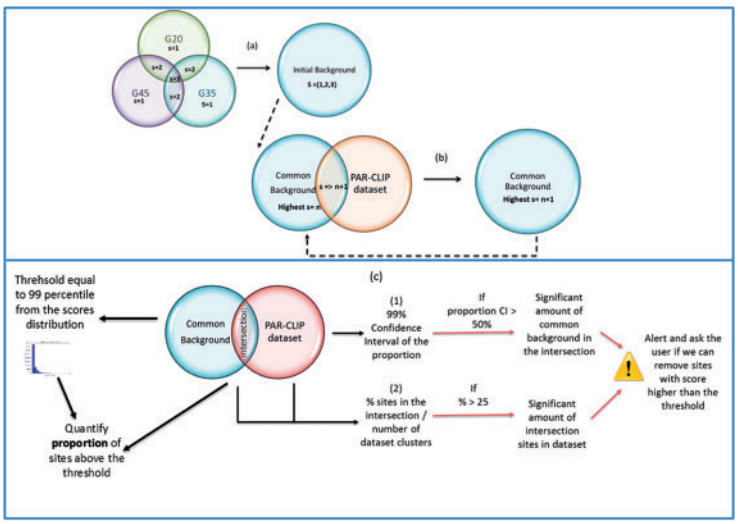

Fig. 1. (**a**) Set-up for initial background and score (**b**). Set-up for the common background. To refine the sites score using the initial background [from (**a**)] and several PAR-CLIP datasets (**c**) Set-up to measure background presence

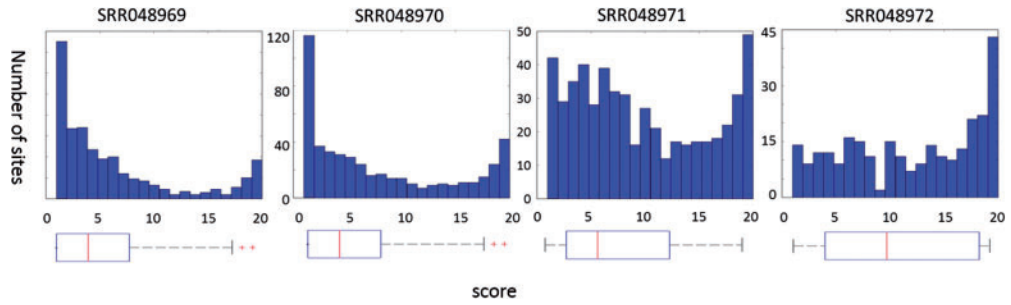

Fig. 2. Histogram and boxplot of the scores in the intersection between the common background and four QKI PAR-CLIP datasets (Tophat alignment)

Table 1. QKI datasets results for the PAR-CLIP dataset and for the background intersection

Dataset	Clusters	Clusters with motif (%)	Background intersection Sites	Background intersection sites with motif (%)	Background intersection sites / Clusters (%)	motifs in intersection / motifs in dataset (%)	Proportion of sites above threshold (score = 8)	BackCLIP sites identified (score ≤ 8)	BackCLIP sites with motif (%)
SRR048969	5286	44	654	11	12	3.1	[32%, 41%]	260	2
SRR048970	5091	48	479	14	9	2.7	[29%, 39%]	176	2
SRR048971	1688	23	539	5	32	7.0	[41%, 52%]	294	2
SRR048972	590	13	276	3	47	10.8	[55%, 59%]	178	1

3 Results

As an example, we selected Quaking (QKI) protein from the nine RBPs (30 PAR-CLIP datasets, details in the Supplementary Information), and four of its PAR-CLIP datasets. These were selected because the proteins motif is known, and these four datasets are from the same study (Hafner *et al.*, 2010). However, there are marked differences.

Figure 2 shows the distributions for the *s* scores in the intersection dataset (boxplot and histogram). We observed significant differences in the two indicators among the four QKI datasets. For the first two datasets, the number of sites in the intersection is <20% of the corresponding number of dataset clusters, whereas for the last two datasets the number was higher than 30%. Moreover, for the last two datasets, 50% or more of the sites in the intersection had a score over the threshold (proportion CI).

The information in Table 1 indicates that the percentage of motifs in the intersection compared with the original dataset is low. The two smaller datasets have a greater proportion of sites in the intersection with the background, compared with the number of clusters. This shows that the amount of common background RNA is different in each dataset even for the same protein. It is also evident that the percentage for the QKI motif (AYUAAY) is higher in the clusters than in the intersection with the background confirming that the background dataset has non-specific sites. However, the motif relative count (motifs in intersection/motifs in dataset) is higher than 5% in the last two datasets because the motif count in the original dataset is small, and any motif in the intersection makes a difference. Table 1 shows the number of sites and corresponding sites with motifs identified as the background using only the data (column fourth and fifth) (Friedersdorf and Keene, 2014), and the number of sites and corresponding sites with motifs identified as the background using BackCLIP (last two columns). Our proposal identifies noisy sites without losing true positive sites.

In conclusion, BackCLIP is a useful tool to identify the amount of common background in any PAR-CLIP dataset.

Acknowledgements

We are grateful to Ana Maria Ortega Villa (Virginia Tech.) for her statistical assessment.

Funding

Support for this research was provided by Universidad Antonio Nariño, project grant 2012221.

Conflict of Interest: none declared.

References

Comoglio,F. *et al.* (2015) Sensitive and highly resolved identification of RNA-protein interaction sites in par-clip data. *BMC Bioinformatics*, **16**.

Dale,R.K. *et al.* (2011) Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, **27**, 3423–3424.

Friedersdorf,M. B. and Keene,J. D. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol.*, **15**, R2.

Hafner,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microrna target sites by PAR-CLIP. *Cell*, **141**, 129–141.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

Licatalosi,D.D. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 46–49.

Reyes-Herrera,P.H. and Ficarra,E. (2014) Computational methods for CLIP-seq data processing. *Bioinform. Biol. Insights*, **8**, 199–207.

Sievers,C. *et al.* (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160.

Trapnell,C. *et al.* (2009) Tophat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**, 1105–1111.

Uren,P. *et al.* (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, **28**, 301–320.

Wang,T. *et al.* (2014) A model-based approach to identify binding sites in clip-seq data. *PLoS One*, **9**, e93248.