Methods 72 (2015) 76–85

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Transcriptome-wide ribonuclease-mediated protein footprinting to identify RNA–protein interaction sites

Ian M. Silverman, Brian D. Gregory *

Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
Cell and Molecular Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form 17 September 2014
Accepted 20 October 2014
Available online 28 October 2014

Keywords:
RNA-binding proteins
RNA–protein interactions
RBPome
Ribonuclease footprinting
Genomics
Post-transcriptional regulation

ABSTRACT

RNA-binding proteins (RBPs) are intimately involved in all aspects of RNA processing and regulation and are linked to neurodegenerative diseases and cancer. Therefore, investigating the relationship between RBPs and their RNA targets is critical for a broader understanding of post-transcriptional regulation in normal and disease processes. The majority of approaches to study RNA–protein interactions interrogate only individual RBPs. However, there are hundreds of these proteins encoded in the human genome, and each cell type expresses a different repertoire, greatly limiting the ability of current methods to capture the global landscape of RNA–protein interactions. To address this gap, we and others have recently developed methods to globally identify regions of RNAs that are bound by proteins in an unbiased manner. Here, we describe a detailed protocol for performing our ribonuclease-mediated protein footprint sequencing approach, termed protein interaction profile sequencing (PIP-seq). In this protocol, RNA–protein interactions are stabilized by cross-linking, and unbound regions are digested with ribonucleases (RNases), leaving only the protein-bound regions intact. To control for RNase insensitive regions, proteins are first denatured and degraded, then protein-depleted RNAs are subjected to RNase treatment. After high-throughput sequencing of the remaining fragments, peak calling is performed to identify protein-protected sites (PPSs). We describe the application of this protocol to a human embryonic kidney cell line (HEK293T) and perform basic quality control, reproducibility, and benchmarking analyses. Finally, we delineate the landscape of protein-interactions in HEK293T cells, underscoring the value of this approach. Future applications of this method to study the dynamics of RNA–protein interactions in developmental and disease processes will help to further uncover the role of RBPs in post-transcriptional regulation.

1. Introduction

Gene expression is regulated at both the transcriptional and post-transcriptional level by *cis*-elements and trans-acting factors. Transcriptional regulation controls only the rate of RNA production, whereas numerous and diverse processes, including RNA splicing, localization, translation and degradation, regulate RNAs at the post-transcriptional level (reviewed in [1–4]). DNA–protein interactions are largely governed by sequence, since DNA is a mostly long double-helical molecule. Conversely, RNA–protein interactions are much more complex, due to the ability of RNA to form a variety of secondary and tertiary structures [5]. This complexity is further compounded by the fact that hundreds of canonical RBPs are encoded in metazoan genomes; with more than 1000 proteins shown to directly interact with messenger RNA (mRNAs) in human cells [6,7]. Therefore, understanding the rules by which RBPs interact with target RNAs is critical to our understanding of the enigmatic and multilayered process of post-transcriptional gene regulation.

Due to the fundamental role of RBPs in basic cell function, inappropriate expression or mutations of their encoding genes or target sites often leads to misregulation of target RNAs and ultimately disease [8]. Such mutations often manifest in neurological phenotypes, the etiology of which remains unclear. These observations have led to the hypothesis that specific disruption of neurological function is a result of the higher complexity of post-transcriptional regulation in the nervous system (i.e. localized translation in axons) [8]. Two specific examples of RBPs involved in neurological disorders are TAR DNA-binding protein (TDP43) and fused in

Abbreviations: PIP-seq, protein interaction profile sequencing; PPSs, protein-protected sites; DSN, duplex-specific nuclease.

* Corresponding author at: Department of Biology, University of Pennsylvania, 433 S. University Ave., Philadelphia, PA 19104, USA. Fax: +1 (215) 898 8780.
E-mail address: bdgregor@sas.upenn.edu (B.D. Gregory).

http://dx.doi.org/10.1016/j.ymeth.2014.10.021
1046-2023/

sarcoma (FUS) that are commonly mutated in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) [9]. These examples highlight the functional significance of RBPs in human biology.

Due to this fundamental importance, numerous approaches have been developed to characterize RNA–protein interactions, with the hopes of obtaining a better understanding of the biochemical mechanisms that dictate these associations. This has resulted in the development of two basic experimental paradigms for understanding RNA–protein interactions at the molecular level. Methods such as SELEX [10] and RNAcompete [11], utilize *in vitro* screening of selected RBPs for affinity with random oligonucleotides. These approaches have generated catalogs of RNA-binding motifs for numerous RBPs in many organisms [12] and are undoubtedly a valuable resource for the community. However, due to the *in vitro* nature of these assays, they may not reflect true biological interactions. The other general framework involves *in vivo* purification of an RBP of interest and screening of bound target RNAs by quantitative real-time PCR (qRT-PCR), microarray, or deep sequencing. For instance, RNA immunoprecipitation (RIP) enables identification of RBP target RNAs, and has been applied to study numerous proteins in a variety of cell types and organisms [13]. However, RIP is limited to the resolution of full-length transcripts and unable to identify the precise binding location of an RBP. The development of RNA:Protein immunoprecipitation in tandem (RIPiT) and cross-linking and immunoprecipitation (CLIP) increased the power of classical approaches by chemically stabilizing RNA–protein interactions, and digesting unbound fragments, allowing for higher precision mapping of the binding sites of individual RBPs as well as RBP complexes [14,15]. In fact, three recent approaches, photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP), individual-nucleotide resolution CLIP (iCLIP) and cross-linking induced mutation site (CIMS) analysis, have allowed for identification of RBP binding sites at single nucleotide resolution [16–18]. Together, these methods have rapidly expanded our understanding of individual RBP–RNA interactions and their roles in post-transcriptional gene regulation.

Although studying the binding sites of individual RBPs is essential, the landscape of RNA–protein interactions in the cell is much more complex. Numerous RBPs bind and regulate a single RNA at different points in its lifecycle, and the combinatorial effects of these proteins can dictate the overall regulatory program for RNA molecules. Moreover, groups of RBPs have been shown to bind and post-transcriptionally regulate functionally related transcripts in an operon-like fashion, underscoring the importance of these multiple interactions [19]. Therefore, methods that take a more global approach for defining RBP–RNA interaction sites are necessary to monitor the multiple interactions on RNAs by the numerous RBPs by which they are bound during the entirety of their life cycle.

To address this, we and others have developed approaches to identify RNA–protein interaction sites by interrogating the global landscape of RNA–protein interactions, without restriction to a specific RBP. These RNA-centric methodologies include PIP-seq, global PAR-CLIP, and protein occupancy profiling [6,20,21]. Development of these approaches is a significant advance over previous single protein approaches, due to the large number and cell-type specific expression patterns of RBPs. Furthermore, the use of these RNA-centric approaches allows the study of global RBP–RNA dynamics through biological processes and treatments or between different cell types [20,22]. Thus, global methods such as PIP-seq allow large-scale studies of RBP co-binding and dynamics that are not possible by single RBP approaches.

The PIP-seq methodology is analogous to DNase footprinting, which has been used for many years to identify histone-free regions of the eukaryotic genome. Specifically, RNA–protein interactions are stabilized by treatment with formaldehyde and subjected to RNase digestion (Fig. 1). In this protocol, we use two different RNase treatments, one that's specific for single-stranded RNA (ssRNA) and one for double-stranded RNA (dsRNA). This increases the number and type of RBP binding sites that are detectable by our methodology. The cross-links are then reversed and the remaining RNA fragments are ligated between sequencing adapters and PCR amplified to generate strand-specific sequencing libraries. One important consideration in this protocol is that RNase insensitive regions would also appear to be protein-binding sites. Therefore, we also produce paired control libraries, in which proteins are first denatured and partially digested prior to treatment with each RNase. By combining these two types of libraries, we are able to reveal the regions of the transcriptome that are specifically protected from RNase digestion by proteins.

PIP-seq has several major advantages over alternative approaches. Unlike, similar approaches, PIP-seq can be applied to any biological sample because it is not dependent on using non-natural nucleotides or ultraviolet light, which is limited to culture cells and simple organisms. Additionally, PIP-seq does not involve a polyA$^+$ selection step, which expands the repertoire of RNAs that can be analyzed using this method. Instead, we use thermostable duplex-specific nuclease (DSN) treatment, which selects against high levels of ribosomal RNA reads in the final libraries without any need for polyA$^+$ selection prior to RNase digestion. This is one of the first approaches beyond gene expression profiling to utilize this technique and provides evidence that it is a powerful means to perform reliable high-throughput RNA sequencing experiments on digested or highly degraded RNA samples. Finally, the use of two structure-specific RNases (an ss- and dsRNase) allows for empirical structure determination as we previously described [23,24].

Using PIP-seq in HeLa cells, we have previously demonstrated that protein interaction sites are functionally conserved regions in human transcripts [21]. PIP-seq was used to uncover uncharacterized protein-binding motifs, some of which may represent multi-RBP or multi-RNA-binding domain interaction sequences. Additionally, the motif data generated in our study of HeLa cells provided additional support for the post-transcriptional operon hypothesis, whereby mRNAs encoding functionally related proteins are co-regulated through the coordinated binding of a specific collection of RBPs [25]. Finally, using this approach we uncovered enrichment for disease-associated SNPs (specifically synonymous SNPs) within protein-protected regions, suggesting that affecting RNA–protein interactions might be a common mechanism of these mutations in human disease. Taken together, these findings support the utility of this assay in studying RNA–protein interaction sites in eukaryotic transcriptomes.

Here, we present a detailed protocol for performing PIP-seq in human culture cells (Fig. 1). We focused this study on HEK293T cells, which have been commonly used in previous single RBP–RNA interaction studies. From this analysis, we find that PIP-seq is highly reproducible and use it to identify ~300,000 PPSs in the HEK293T transcriptome in a single experiment. We find that these RNA–protein interaction sites are highly conserved, enriched for known binding sites of multiple RBPs, and are localized throughout the coding and non-coding portions of mRNA transcripts. In total, our results suggest that this will be the preferable approach for global analysis of RNA–protein interaction sites in eukaryotic transcriptomes.