Genome analysis

Metadensity: a background-aware python pipeline for summarizing CLIP signals on various transcriptomic sites

Hsuan-Lin Her\textsuperscript{1,2}, Evan Boyle\textsuperscript{2} and Gene W. Yeo\textsuperscript{1,2,3,4,}*

\textsuperscript{1}Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA 92093, USA, \textsuperscript{2}Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92093, USA, \textsuperscript{3}Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093, USA and \textsuperscript{4}Stem Cell Program, University of California San Diego, La Jolla, CA 92093, USA

*To whom correspondence should be addressed.

Associate Editor: Marieke Kuijjer

Received on June 20, 2022; revised on September 15, 2022; editorial decision on October 8, 2022; accepted on November 7, 2022

Abstract

\textbf{Motivation:} Cross-linking and immunoprecipitation (CLIP) is a technology to map the binding sites of RNA-binding proteins (RBPs). The region where an RBP binds within RNA is often indicative of its molecular function in RNA processing. As an example, the binding sites of splicing factors are found within or proximal to alternatively spliced exons. To better reveal the function of RBPs, we developed a tool to visualize the distribution of CLIP signals around various transcript features.

\textbf{Results:} Here, we present Metadensity (https://github.com/YeoLab/Metadensity), a software that allows users to generate metagene plots. Metadensity allows users to input features such as branchpoints and preserves the near-nucleotide resolution of CLIP technologies by not scaling the features by length. Metadensity normalizes immunoprecipitated libraries with background controls, such as size-matched inputs, then windowing in various user-defined features. Finally, the signals are averaged across a provided set of transcripts.

\textbf{Availability and implementation:} Metadensity is available at https://github.com/YeoLab/Metadensity, with example notebooks at https://metadensity.readthedocs.io/en/latest/tutorial.html.

\textbf{Contact:} geneyeo@ucsd.edu

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics Advances online.

1 Introduction

RNA-binding proteins (RBPs) are key modulators of RNA metabolism (Hentze \textit{et al.}, 2018). Cross-linking and immunoprecipitation (IP) followed by sequencing (HITS-CLIP/CLIP-seq) (Darnell, 2010; Licatalosi \textit{et al.}, 2008; Yeo \textit{et al.}, 2009) and derivatives such as PAR-CLIP (Hafner \textit{et al.}, 2010), iCLIP (Briese \textit{et al.}, 2019; König \textit{et al.}, 2010), enhanced CLIP (eCLIP) (Van Nostrand \textit{et al.}, 2016) and irCLIP (Zarnegar \textit{et al.}, 2016) are technologies to discover transcriptome-wide RNA interaction sites of RBPs. Briefly, after crosslinking of the RBP to RNA and limited digestion of unprotected RNA, the protected RBP–RNA fragment is isolated by IP, converted into cDNA and then sequenced. During library preparation, depending on the reverse transcription conditions, the crosslinking of the nucleotide causes reverse transcription stoppage or mutation (Chakrabarti \textit{et al.}, 2018; Hauer \textit{et al.}, 2015; Van Nostrand \textit{et al.}, 2017). As a result, crosslink-induced read truncations (CITs) or mutations (CIMs) can enable near-nucleotide resolution recovery of a fraction of the binding sites.

Enriched RBP binding at specific transcript features provide important clues to the function of the RBP. To illustrate, splicosomal proteins are enriched at the 5’- and 3’-splice sites (ss) (Moore and Sharp, 1993), and RNA decay factors often interact within the 3’-untranslated regions (UTRs) of protein-coding genes (Muers, 2013). By examining the distribution of RBP-binding sites around canonical features in genes, one can infer the functions of RBPs.

The distribution of transcriptome-wide signals is often summarized in metagene plots. However, existing metagene packages (Olarerin-George and Jaffrey, 2017) emphasize the 5’-UTR-CDS-3’-UTR model on mature messenger RNAs (mRNAs). Such a model is useful in studying RNA stability and/or translational regulators. However, many RBPs bind premature mRNAs to regulate splicing, polyadenylation and export (Hentze \textit{et al.}, 2018). To thoroughly comprehend an RBP’s role in RNA processing, a software tool that includes multiple models of metagene density is needed. In addition, CLIP-seq data contain various background signals (Van Nostrand \textit{et al.}, 2016) and existing metagene packages do not support background normalization. The coverage at each position is strongly
Here, we present Metadensity, a python package that supports multiple types of metagene plots and allows user-customized feature creation. In addition, it has a built-in normalization procedure to account for background in the SMInput library. Finally, it allows the user to not only utilize the read coverage as an approximation of binding, but also support the extraction of various diagnostic signals such as CITs and CIMs.

2 Overview

Metadensity starts by extracting CLIP diagnostic signals from BAM/BIGWIG files for each transcript, using either the read coverage or summation of CITs and CIMs. Alternatively, to speed up computation, a WIG track can be pre-computed (Fig. 1A), which allows us to accommodate other sequencing technologies that have signals such as CITs and CIMs. The package allows users to input customized, non-Genencode features. For example, in Figure 1B–D, the metagene is supplied with branchpoints detected by CaptureSeq (Mercer et al., 2015; Signal et al., 2018). With this feature, we clearly recapitulate SF3B4’s role in branchpoint recognition (Broser et al., 1993; Krammer et al., 1987; Moore and Sharp, 1993). Similarly, proteins part of the U2 complex has strongest enrichment at the 3’-ss. In addition, it can compute the regular 5’-UTR-CDS-3’-UTR model (Supplementary Fig. S1), and densities around polyadenylation sites (Supplementary Figs S2 and S3).

3 Conclusion

Here, we provide a user-friendly package to generate various metagene plots for visualizing CLIP-seq data, including pre-mRNA features such as branchpoints and polyadenylation sites. The package performs background normalization and outputs RBP maps for transcriptome-wide eCLIP visualization. Users can utilize these visualizations to interrogate RBP functions. We showcase how the U2 and SF3B complex’s density align with current knowledge and their role in the spliceosome. Similarly, U2 proteins have strongest binding at the 3’-ss. The various metagene models will allow us to propose testable hypotheses for RBPs on their impact in various steps of RNA-processing.

Funding

This work was supported by US National Institutes of Health research grants HG004659 and HG009889.

Conflict of Interest: G.W.Y. is a co-founder, member of the Board of Directors, on the SAB, equity holder, and paid consultant for Locanabio and Eclipse BioInnovations. G.W.Y. is a visiting professor at the National University of Singapore. G.W.Y.’s interests have been reviewed and approved
by the University of California, San Diego in accordance with its conflict-of-interest policies. The authors declare no other competing financial interests.

**Data availability**

All ENCODE eCLIP datasets are available through the ENCODE website (encodeproject.org). Annotations of transcriptomic features are available at GENCODE (https://www.gencodegenes.org/).

**References**

Briese, M. et al. (2019) A systems view of spliceosomal assembly and branch-points with iCLIP. Nat. Struct. Mol. Biol., 26, 930–940.

Brosi, R. et al. (1993) Separation of splicing factor SF3 into two components and purification of SF3a activity. J. Biol. Chem., 268, 17640–17646.

Chakrabarti, A.M. et al. (2018) Data science issues in studying protein–RNA interactions with CLIP technologies. Annu. Rev. Biomed. Data Sci., 1, 235–261.

Darnell, R.B. (2010) HITS-CLIP: panoramic views of protein–RNA regulation in living cells. Wiley Interdiscip. Rev. RNA, 1, 266–286.

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

Hauer, C. et al. (2015) Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP. Nat. Commun., 6, 7921.

Hentze, M.W. et al. (2018) A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol., 19, 327–341.

König, J. et al. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol., 17, 909–915.

Krämer, A. et al. (1987) Separation of multiple components of HeLa cell nuclear extracts required for pre-messenger RNA splicing. J. Biol. Chem., 262, 17630–17640.

Licatalosi, D.D. et al. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 456, 464–469.

Mercer, T.R. et al. (2015) Genome-wide discovery of human splicing branchpoints. Genome Res., 25, 290–303.

Moore, M.J. and Sharp, P.A. (1993) Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. Nature, 365, 364–368.

Muers, M. (2013) mRNA decay factors regulate transcription. Nat. Rev. Genet., 14, 444.

Olarerin-George, A.O. and Jaffrey, S.R. (2017) MetaPlotR: a perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. Bioinformatics, 33, 1563–1564.

Signal, B. et al. (2018) Machine learning annotation of human branchpoints. Bioinformatics, 34, 920–927.

Van Nostrand, E.L. et al. (2020) Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol., 21, 90.

Van Nostrand, E.L. et al. (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods, 13, 508–514.

Van Nostrand, E.L. et al. (2017) Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions. Methods, 126, 29–37.

Yeo, G.W. et al. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA–protein interactions in stem cells. Nat. Struct. Mol. Biol., 16, 130–137.

Zarnegar, B.J. et al. (2016) rCLIP platform for efficient characterization of protein–RNA interactions. Nat. Methods, 13, 489–492.