R2Dtool: Positional interpretation of RNA-centric information in the context of transcriptomic and genomic features

Aditya J. Sethi[1,2,3], Pablo Acera Mateos[1,2,3], Rippei Hayashi[1], Nikolay Shirokikh[1], Eduardo Eyras[1,2,3]

[1] Shine-Dalgarno Centre for RNA Innovation, John Curtin School of Medical Research, Australian National University, Canberra, Australia
[2] Centre for Computational Biomedical Sciences, John Curtin School of Medical Research, Australian National University, Canberra, Australia
[3] EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, Australia

Abstract

Long-read sequencing enables isoform-resolved detection of functionally important RNA elements, such as RNA chemical modifications, RNA secondary structure, or RNA-protein interaction sites. Importantly, the functional impact of these elements can relate to their positions relative to isoform-specific transcript features, such as start and stop codons, open reading frames, exon-exon junctions and transcript termini. Relating transcriptomic and genomic features during sequencing data analysis is challenged by the flexibility of RNA biogenesis and the diversity of alternative mRNA transcripts. To address these challenges and streamline the mapping between transcriptome-mapped data and genome-mapped data, we developed R2Dtool. We illustrate R2Dtool's capability to process long-read transcriptomic information and potential for interpretation in the context of transcript and genomic annotation features using epitranscriptomics data. R2Dtool embraces and expedites analysis of RNA complexity, and we anticipate it will empower multiple transcriptomic studies for interpretation and discovery. R2Dtool is freely available under the MIT license from https://github.com/comprna/R2Dtool.

Introduction

Long-read sequencing provides a new capacity to map RNA molecular elements, such as RNA modifications (Acera Mateos et al., 2022; Liu et al., 2019; Stephenson et al., 2020), RNA secondary structure (Aw et al., 2021; Stephenson et al., 2022), and RNA-protein interaction sites (Lin et al., 2022) to specific transcript isoforms. The roles of these functional RNA elements are strongly determined by their positioning and distribution around both gene-specific and transcript-specific landmarks, including transcription starts sites, exon-exon junctions, start and stop codons, and polyadenylation sites. For example, the positional enrichment of N6-methyladenosine (m6A) downstream of stop codons is associated with the regulation of transcript deadenylation at the 3' end of transcripts (Dominissini et al., 2012; Lee et al., 2020). Long-read sequencing reveals that exon secondary structures are typically isoform-specific, and can affect transcript-specific translational efficiency by modulating the availability of the start codon (Aw et al., 2021). Protein-RNA interaction sites often distribute around RNA landmarks such as start and stop codons (Van Nostrand et al., 2020), and can occur both competitively and cooperatively on single transcripts as recently revealed by long-read sequencing (Lin et al., 2022).

Determining the positioning of RNA elements relative to local transcriptomic landmarks can be confounded by the complexity of the various transcript isoforms generated by alternative transcription and RNA processing mechanisms, which are highly coordinated (Neugebauer, 2019; Schor et al., 2013). On average, more than ten alternative transcripts per protein-coding gene usually are found (Pertea et al., 2018), many of which encode transcript-specific open reading frames (Reixachs-Solé and Eyras, 2022). A combined analysis of transcriptomic and genomic features is necessary to accurately position RNA elements in the context of transcriptomic landmarks, therefore requiring precise knowledge of transcript isoforms. Long-read sequencing enables isoform-specific attribution

of candidate functional elements, and direct RNA sequencing immediately includes a possibility of modification detection in individual molecules, opening an opportunity to map isoform-specific features to the genome coordinates and back. Creating a unified representation of the mapping to genomic and transcriptomic references can facilitate analysis and interpretation of transcript-specific data. Only mapping long RNA reads to a genome reference may lead to uncertainties in how individual reads correspond to specific transcripts and complicate understanding of the positional biases of RNA features from transcript-specific landmarks. Further, studying the distributions of RNA features in genomic space, e.g., using a genome browser, is difficult when using transcript-centric analyses. On the other hand, long RNA reads can be mapped to transcriptomic references (Dong et al., 2022) to enable precise localization of RNA elements, but this complicates the comparison of common positions between alternative transcripts, such as constitutive exons. To aid the functional interpretation of long-read transcriptomics studies, there is thus a need for tools that enable simple and reproducible mapping between RNA and DNA coordinate systems and can systematically annotate transcript-centric positions in the context of transcript and genomic landmarks. To address these challenges, we developed R2Dtool for mapping and interpreting long-read transcriptomics information in the context of transcript and genome references (Fig. 1, central panel).

Implementation

R2Dtool is a utility for analyzing and interpreting transcriptomic positions corresponding to RNA features of interest, such as modified RNA nucleotides, in the context of the relevant transcript-specific and genome-specific landmarks (Fig. 1, central panel). R2Dtool takes as input the transcript-centric sites obtained from long-read based experiments, formatted as 6+ column browser-extensible data (BED) file. Additionally, it allows for any number of additional data fields to be included in columns 7+. By design, R2Dtool remains compatible with a range of upstream methods. Further, gene structure information is provided to R2Dtool in the *de facto* standards general transfer format (GTF/GFF2). We outline below R2Dtool's major capabilities and simultaneously demonstrate its potential for interpretation and discovery using isoform-resolved epitranscriptomic data for m6A calculated from nanopore long-read sequencing (Acera Mateos et al., 2022).

Genomic transposition of isoform-specific sites

To enable the characterization of transcriptomic-centric modification calls in the context of the feature of the corresponding genic locus, R2Dtool (R2_liftover.R command) maps transcript sites to their genomic locus while simultaneously annotating them with relevant transcript and genomic features (Fig. 1, left panel). The mapping of bed-like transcript information to genomic space uses the genomicFeatures R package (Lawrence et al., 2013). The output is an extension of the standard BED format, where the first 6 columns can be reused to visualize the sites using genomic browsers and compare across conditions. To illustrate this command, we mapped m6A methylation calls to genomic coordinates (Fig. 1, left). In this case, liftover enables the comparison of methylation between alternative transcript isoforms (Fig. 1, left).

Isoform-specific positional annotation

Using the transcriptomic positions of relevant sites provided in transcript-centric BED format and the corresponding gene structures in GTF/GFF format, the R2Dtool command **R2_annotate.R** calculates their relative position along transcript features. Relative positions in the context of a fixed-length virtual transcript, i.e., metatranscript coordinates, are calculated using the approach described previously (Olarerin-George and Jaffrey, 2017). **R2_annotate.R** also calculates for each site of interest the distances to the available annotation features, such as the start and end of the ORF, the start and end

of the transcript, and upstream and downstream nearest splice junctions, all in a transcript-specific manner, enabling a range of downstream positional analyses (Fig. 1, right panel).

Metatranscript plots

R2Dtool incorporates the command **R2_plotMetaTranscript.R** to produce density plots from the output from the R2_annotate.R command. To illustrate the usage of R2Dtool metaplots, we produced the metatranscript plot for the density of m6A along mRNA transcripts from annotated transcript-centric methylation calls (Fig. 1, right panel).

Metajunction plots

Plots can be tailored to landmark positions of the annotation, such as the open reading frame start and end, or the exon-exon junctions, using the command **R2_metaJunction.R**. We illustrate this command using the relative distances of m6A modifications to their nearest upstream and downstream junctions obtained from R2_annotate.R (Fig. 1, right panel). The metaJunction plot shows a clear exclusion of m6A sites around 200nt away from the junctions, which is consistent with previous findings reporting a depletion of m6A from the vicinity of exon-exon junctions (Uzonyi et al., 2022) (Ke et al., 2017).

Conclusions

R2Dtool provides a simple, robust, and reproducible framework to analyze and interpret functional information from isoform-resolved RNA sequencing data, particularly long-read transcriptomics. The novelty of R2Dtool is that it provides a comprehensive annotation of identified transcript-centric sites in a simple-to-use, convenient framework. The isoform specificity of long-read transcriptomics results in the faithful identification of candidate relevant transcript positions, and therefore an accurate connection of the metatranscript-mapped coordinates and distances to the transcript features such as exon-exon junctions. R2Dtool is particularly useful for epitranscriptomics, where there have been recent significant advances in the isoform-specific identification of RNA modifications using direct nanopore sequencing reads. We anticipate that R2Dtool could empower multiple long-read transcriptomic studies and will be applied to interpretation and discovery of RNA features.

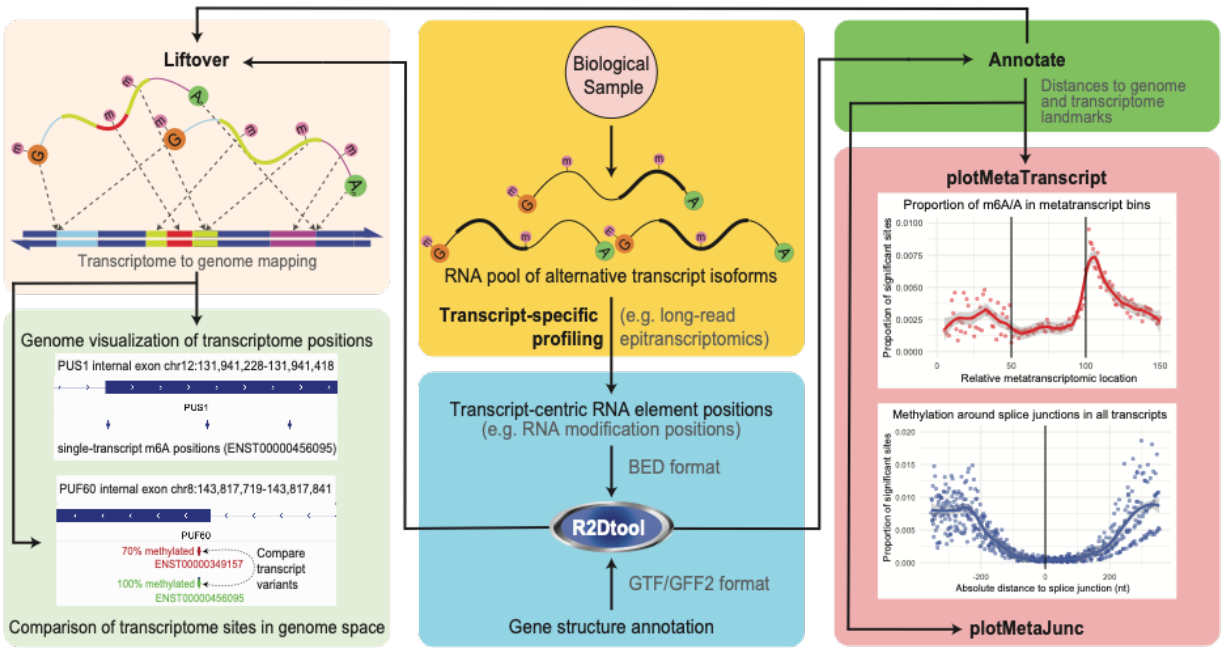

Figure 1. R2Dtool facilitates the analysis and interpretation of RNA-centric information in the context of relevant transcriptomic and genomic features. Center panel: Diagram describing the inputs and operations in R2Dtool. Left panels: IGV browser snapshots created with R2Dtool output to visualize transcriptome-specific m6A modifications inferred from direct nanopore sequencing reads in HEK293 cells. The upper panel shows three m6A sites in the genomic context of the gene PUS1 together with their transcript of origin. The lower panel shows two modified sites from two distinct isoforms with different stoichiometry that nonetheless share identical genomic positions. Right panels: R2Dtool meta-plots for isoform-resolved m6A modifications. The upper panel shows the output of R2Dtool *plotMetaTranscript* for the density of m6A along relative positions in the 5' untranslated region (UTR), coding sequence (CDS), and 3'UTR of mRNA transcripts in human. The lower panel shows the output of *plotMetaJunction* for the density of m6A modifications relative to their distances to the closest downstream (left) and upstream (right) exon-exon junction.

Funding

This research was supported by the Australian Research Council (ARC) Discovery Project grants DP210102385 (to RH and EE), DP220101352 (to EE), by the National Health and Medical Research Council (NHMRC) Investigator Grant GNT1175388 (to NS). This work was also supported by a Bootes grant (2021-2022).

Conflicts of interest: none declared

References

Acera Mateos, P., Sethi, A.J., Guarnacci, M., Ravindran, A., Srivastava, A., Xu, J., Woodward, K., Hamilton, W., Gao, J., Starrs, L.M., Burgio, G., Hayashi, R., Wickramasinghe, V., Dehorter, N., Preiss, T., Shirokikh, N., Eyras, E., 2022. Identification of m6A and m5C RNA modifications at single-molecule resolution from Nanopore sequencing (preprint). Bioinformatics. https://doi.org/10.1101/2022.03.14.484124

Alfonzo, J.D., Brown, J.A., Byers, P.H., Cheung, V.G., Maraia, R.J., Ross, R.L., 2021. A call for direct sequencing of full-length RNAs to identify all modifications. Nat. Genet. 53, 1113–1116. https://doi.org/10.1038/s41588-021-00903-1

Aw, J.G.A., Lim, S.W., Wang, J.X., Lambert, F.R.P., Tan, W.T., Shen, Y., Zhang, Y., Kaewsapsak, P., Li, C., Ng, S.B., Vardy, L.A., Tan, M.H., Nagarajan, N., Wan, Y., 2021. Determination of isoform-specific RNA structure with nanopore long reads. Nat. Biotechnol. 39, 336–346. https://doi.org/10.1038/s41587-020-0712-z

Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avşar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S.G., Aguilo, F., Bujnicki, J.M., 2022. MODOMICS: a database of RNA modification pathways. 2021 update. Nucleic Acids Res. 50, D231–D235. https://doi.org/10.1093/nar/gkab1083

Boo, S.H., Kim, Y.K., 2020. The emerging role of RNA modifications in the regulation of mRNA stability. Exp. Mol. Med. 52, 400–408. https://doi.org/10.1038/s12276-020-0407-z

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R., Rechavi, G., 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485, 201–206. https://doi.org/10.1038/nature11112

Dong, X., Du, M.R.M., Gouil, Q., Tian, L., Baldoni, P.L., Smyth, G.K., Amarasinghe, S.L., Law, C.W., Ritchie, M.E., 2022. Benchmarking long-read RNA-sequencing analysis tools using *in silico* mixtures (preprint). Bioinformatics. https://doi.org/10.1101/2022.07.22.501076

Furlan, M., Delgado-Tejedor, A., Mulroney, L., Pelizzola, M., Novoa, E.M., Leonardi, T., 2021. Computational methods for RNA modification detection from nanopore direct RNA sequencing data. RNA Biol. 18, 31–40. https://doi.org/10.1080/15476286.2021.1978215

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J.J., Vågbø, C.B., Geula, S., Hanna, J.H., Black, D.L., Darnell, J.E., Darnell, R.B., 2017. m ^{6} A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. Genes Dev. 31, 990–1006. https://doi.org/10.1101/gad.301036.117

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. 9, e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lee, Y., Choe, J., Park, O.H., Kim, Y.K., 2020. Molecular Mechanisms Driving mRNA Degradation by m6A Modification. Trends Genet. 36, 177–188. https://doi.org/10.1016/j.tig.2019.12.007

Lin, Y., Kwok, S., Thai, B.Q., Alabi, Y., Ostrowski, M., Wu, K., Floor, S.N., 2022. Single molecule co-occupancy of RNA-binding proteins with an evolved RNA deaminase. bioRxiv 2022.09.06.506853. https://doi.org/10.1101/2022.09.06.506853

Liu, H., Begik, O., Lucas, M.C., Ramirez, J.M., Mason, C.E., Wiener, D., Schwartz, S., Mattick, J.S., Smith, M.A., Novoa, E.M., 2019. Accurate detection of m6A RNA modifications in native RNA sequences. Nat. Commun. 10, 4079. https://doi.org/10.1038/s41467-019-11713-9

Mendel, M., Delaney, K., Pandey, R.R., Chen, K.-M., Wenda, J.M., Vågbø, C.B., Steiner, F.A., Homolka, D., Pillai, R.S., 2021. Splice site m6A methylation prevents binding of U2AF35 to inhibit RNA splicing. Cell 184, 3125-3142.e25. https://doi.org/10.1016/j.cell.2021.03.062

Neugebauer, K.M., 2019. Nascent RNA and the Coordination of Splicing with Transcription. Cold Spring Harb. Perspect. Biol. 11, a032227. https://doi.org/10.1101/cshperspect.a032227

Olarerin-George, A.O., Jaffrey, S.R., 2017. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. Bioinformatics 33, 1563–1564. https://doi.org/10.1093/bioinformatics/btx002

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A., Salzberg, S.L., 2018. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 19, 208. https://doi.org/10.1186/s13059-018-1590-2

Reixachs-Solé, M., Eyras, E., 2022. Uncovering the impacts of alternative splicing on the proteome with current omics techniques. WIREs RNA 13. https://doi.org/10.1002/wrna.1707

Schor, I.E., Gómez Acuña, L.I., Kornblihtt, A.R., 2013. Coupling Between Transcription and Alternative Splicing, in: Wu, J.Y. (Ed.), RNA and Cancer, Cancer Treatment and Research. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–24. https://doi.org/10.1007/978-3-642-31659-3_1

Schumann, U., Zhang, H.-N., Sibbritt, T., Pan, A., Horvath, A., Gross, S., Clark, S.J., Yang, L., Preiss, T., 2020. Multiple links between 5-methylcytosine content of mRNA and translation. BMC Biol. 18, 40. https://doi.org/10.1186/s12915-020-00769-5

Stephenson, W., Razaghi, R., Busan, S., Weeks, K.M., Timp, W., Smibert, P., 2022. Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. Cell Genomics 2, 100097. https://doi.org/10.1016/j.xgen.2022.100097

Stephenson, W., Razaghi, R., Busan, S., Weeks, K.M., Timp, W., Smibert, P., 2020. Direct detection of RNA modifications and structure using single molecule nanopore sequencing (preprint). Genomics. https://doi.org/10.1101/2020.05.31.126763

Uzonyi, A., Slobodin, B., Schwartz, S., 2022. Exon-intron architecture determines mRNA stability by dictating m6A deposition (preprint). Molecular Biology. https://doi.org/10.1101/2022.06.29.498130

Van Nostrand, E.L., Pratt, G.A., Yee, B.A., Wheeler, E.C., Blue, S.M., Mueller, J., Park, S.S., Garcia, K.E., Gelboin-Burkhart, C., Nguyen, T.B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X.-D., Graveley, B.R., Yeo, G.W., 2020. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol. 21, 90. https://doi.org/10.1186/s13059-020-01982-9

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., He, C., 2014. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature 505, 117–120. https://doi.org/10.1038/nature12730