COMPUTATIONAL EPITRANSCRIPTOMICS: BIOINFORMATIC APPROACHES FOR THE ANALYSIS OF RNA MODIFICATIONS

EDITED BY: Mattia Pelizzola, Pavel V. Baranov and Erik Dassi

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-440-5
DOI 10.3389/978-2-88966-440-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL EPITRANSCRIPTOMICS: BIOINFORMATIC APPROACHES FOR THE ANALYSIS OF RNA MODIFICATIONS

Topic Editors:
Mattia Pelizzola, Italian Institute of Technology (IIT), Italy
Pavel V. Baranov, University College Cork, Ireland
Erik Dassi, University of Trento, Italy

Citation: Pelizzola, M., Baranov, P. V., Dassi, E., eds. (2021). Computational Epitranscriptomics: Bioinformatic Approaches for the Analysis of RNA Modifications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-440-5

Table of Contents

05 ***Editorial: Computational Epitranscriptomics: Bioinformatic Approaches for the Analysis of RNA Modifications***
Erik Dassi, Pavel V. Baranov and Mattia Pelizzola

07 ***Graphical Workflow System for Modification Calling by Machine Learning of Reverse Transcription Signatures***
Lukas Schmidt, Stephan Werner, Thomas Kemmer, Stefan Niebler, Marco Kristen, Lilia Ayadi, Patrick Johe, Virginie Marchand, Tanja Schirmeister, Yuri Motorin, Andreas Hildebrandt, Bertil Schmidt and Mark Helm

18 ***The Identification of Differentially Expressed Genes Showing Aberrant Methylation Patterns in Pheochromocytoma by Integrated Bioinformatics Analysis***
Dengqiang Lin, Jinglai Lin, Xiaoxia Li, Jianping Zhang, Peng Lai, Zhifeng Mao, Li Zhang, Yu Zhu and Yujun Liu

32 ***Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2′-O-Methylations in RNA by RiboMethSeq***
Florian Pichot, Virginie Marchand, Lilia Ayadi, Valérie Bourguignon-Igel, Mark Helm and Yuri Motorin

41 ***Quantifying RNA Editing in Deep Transcriptome Datasets***
Claudio Lo Giudice, Domenico Alessandro Silvestris, Shalom Hillel Roth, Eli Eisenberg, Graziano Pesole, Angela Gallo and Ernesto Picardi

56 ***PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation***
Bowen Song, Yujiao Tang, Zhen Wei, Gang Liu, Jionglong Su, Jia Meng and Kunqi Chen

64 ***MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets***
Luca Cozzuto, Huanle Liu, Leszek P. Pryszcz, Toni Hermoso Pulido, Anna Delgado-Tejedor, Julia Ponomarenko and Eva Maria Novoa

75 ***RNA N6-Methyladenosine-Related Gene Contribute to Clinical Prognostic Impact on Patients With Liver Cancer***
Wei Wang, Bo Sun, Yang Xia, Shenghong Sun and Chiyi He

86 ***Direct RNA Sequencing for the Study of Synthesis, Processing, and Degradation of Modified Transcripts***
Mattia Furlan, Iris Tanaka, Tommaso Leonardi, Stefano de Pretis and Mattia Pelizzola

98 ***How Do You Identify m^6 A Methylation in Transcriptomes at High Resolution? A Comparison of Recent Datasets***
Charlotte Capitanchik, Patrick Toolan-Kerr, Nicholas M. Luscombe and Jernej Ule

110 ***LITHOPHONE: Improving lncRNA Methylation Site Prediction Using an Ensemble Predictor***
Lian Liu, Xiujuan Lei, Zengqiang Fang, Yujiao Tang, Jia Meng and Zhen Wei

120 N6-Methyladenosine RNA Methylation Regulators Have Clinical Prognostic Values in Hepatocellular Carcinoma

Wei Liu, Cuiqing Zhong, Deguan Lv, Mengjie Tang and Feng Xie

134 tRNA Fragments Populations Analysis in Mutants Affecting tRNAs Processing and tRNA Methylation

Anahi Molla-Herman, Margarita T. Angelova, Maud Ginestet, Clément Carré, Christophe Antoniewski and Jean-René Huynh

EDITORIAL
published: 11 December 2020
doi: 10.3389/fgene.2020.630360

Editorial: Computational Epitranscriptomics: Bioinformatic Approaches for the Analysis of RNA Modifications

Erik Dassi[1], *Pavel V. Baranov*[2,3] *and Mattia Pelizzola*[4]*

[1] Laboratory of RNA Regulatory Networks, Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy, [2] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland, [3] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia, [4] Center for Genomic Science, Fondazione Istituto Italiano di Tecnologia, Milan, Italy

Keywords: RNA modification, bioinformatics and computational biology, RNA methylation, RNA editing, pseudouridine (Ψ), nanopore

Editorial on the Research Topic

Computational Epitranscriptomics: Bioinformatic Approaches for the Analysis of RNA Modifications

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

****Correspondence:***
Mattia Pelizzola
mattia.pelizzola@iit.it

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *17 November 2020*
Accepted: *24 November 2020*
Published: *11 December 2020*

Citation:
Dassi E, Baranov PV and Pelizzola M (2020) Editorial: Computational Epitranscriptomics: Bioinformatic Approaches for the Analysis of RNA Modifications.
Front. Genet. 11:630360.
doi: 10.3389/fgene.2020.630360

RNA modifications were discovered decades ago, and more than 150 different marks have been found decorating various RNA species, including coding and non-coding transcripts (Boccaletto et al., 2018). Yet, only in the last decade this research field rapidly expanded, due to the development of simple and effective methods for the genome-wide identification of some of these marks, such as MeRIP-seq for the profiling of N6-methyladenosine (m6A) (Dominissini et al., 2012; Meyer et al., 2012). The renowned interest in the field led to the identification of key effectors—writers, erasers, and readers—that establish and decode the patterning of specific marks. This suggested that RNA modifications have the potential to be dynamically controlled, similarly to their genomic counterparts, the modifications of DNA and chromatin. Indeed, in analogy to the epigenome, the collective set of RNA modifications was named epitranscriptome. Altogether, the epitranscriptome is considered an important determinant of RNA fate, and specific marks were found to be involved in various steps of the RNA life cycle including, while not limited to, transcription, processing, decay, and translation (Roundtree et al., 2017).

As often occurs, following the birth of a new omics, the development of computational methods that are tailored to the analysis of those high-throughput datasets started to flourish. In analogy to the development of computational epigenomics (Bock and Lengauer, 2008), this research field could be referred to as computational epitranscriptomics.

This Research Topic collects a number of contributions in this field. Few manuscripts focused on the development of novel methods for the prediction of RNA modifications. A Galaxy-based user friendly graphical workflow was developed that cover the preprocessing of omics data, the quantification of mismatch and arrest rates with single-nucleotide resolution, and the subsequent machine learning, modification calling and visualization (Schmidt et al.). A computational workflow dedicated to 2′-O-methylation marks was optimized, allowing a more accurate detection of these marks and a more precise quantification of their level variations (Pichot et al.). A novel tool (LITOPHONE) was developed that adopts an ensemble predictor relying on sequence features to predict m6A sites in long non-coding RNAs (Liu L. et al.). A web server (PIANO) was implemented that relies on various genomic features, including sequence information, for the prediction and functional annotation of pseudouridine sites (Song et al.). Finally, a bioinformatic pipeline (tRFs-Galaxy) was developed for the study of small non-coding RNAs derived from tRNAs (tRFs), allowing the study of tRFs biogenesis in *Drosophila melanogaster* (Molla-Herman et al.).

Two additional contributions discussed pitfalls in the analysis of specific marks. A first study discussed the impact of different bioinformatics steps on the detection of RNA editing events, describing key metrics for the quantification of their level of activity (Giudice et al.). A second contribution compared m6A genome-wide maps generated in various studies based on eight different methods, discussing the agreement of the data and the challenges in their comparative analysis, revealing an expression bias in the detected genes (Capitanchik et al.).

Two contributions were focused on the use of direct RNA sequencing through the Nanopore platform that enables long-reads sequencing of native transcripts. A perspective discussed how these data could allow quantifying the dynamics of modified RNAs at the level of individual isoforms (Furlan et al.). A second study introduced MasterOfPores, a NextFlow workflow that facilitates the analysis of these data, allowing the prediction of RNA modifications and the estimation of polyA tail lengths (Cozzuto et al.).

Finally, three different studies introduced bioinformatics workflows for studying the impact of RNA modifications in various tumor types. In the first study, a workflow based on consensus clustering and gene set enrichment analysis was presented that allowed the subsequent construction of a prognostic risk model suggesting the involvement of three m6A-related genes in liver cancer (Wang et al.). In the second study, bioinformatics analyses revealed a risk signature based on three m6A regulators, proposing candidate prognostic markers predictor of the clinicopathological features in hepatocellular carcinoma (Liu W. et al.). In the third study, integrated bioinformatics analyses led to the identification of differentially expressed transcripts with aberrant methylation patterns in malignant pheochromocytoma (Lin et al.).

Despite the rapid advance of the field, which allowed expanding the set of known marks, profiling their pattern, and disclosing their functional roles, a number of open questions remain (Frye et al., 2016). Most modifications remain poorly characterized, it is unclear whether different marks crosstalk and whether an epitranscriptional code exists. We are only starting to understand how, where and when these modifications are altered and whether they represent potential therapeutic targets in diseases. Key for answering these and other questions will be the continuous development of methods to map and analyze these marks. This research would benefit from the establishment of large scale collaborative and networking efforts such as the European Epitranscriptomics Network (www.epitran.eu) (Jantsch et al., 2018).

AUTHOR CONTRIBUTIONS

All authors contributed writing this Editorial and managing the corresponding Research Topic.

FUNDING

This work was supported by the EPITRAN COST Action (CA16120).

REFERENCES

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Bock, C., and Lengauer, T. (2008). Computational epigenetics. *Bioinformatics* 24, 1–10. doi: 10.1093/bioinformatics/btm546

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G., and Suzuki, T. (2016). RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.* 17, 365–372. doi: 10.1038/nrg.2016.47

Jantsch, M. F., Quattrone, A., O'Connell, M., Helm, M., Frye, M., Macias-Gonzales, M., et al. (2018). Positioning Europe for the EPITRANSCRIPTOMICS challenge. *RNA Biol.* 15, 829–831. doi: 10.1080/15476286.2018.1460996

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

TECHNOLOGY AND CODE
published: 25 September 2019
doi: 10.3389/fgene.2019.00876

Graphical Workflow System for Modification Calling by Machine Learning of Reverse Transcription Signatures

Lukas Schmidt[1†], Stephan Werner[1†], Thomas Kemmer[2], Stefan Niebler[3], Marco Kristen[1], Lilia Ayadi[4,5], Patrick Johe[1], Virginie Marchand[4], Tanja Schirmeister[1], Yuri Motorin[4,5], Andreas Hildebrandt[2]*, Bertil Schmidt[3]* and Mark Helm[1]*

[1] Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University, Mainz, Germany, [2] Institute of Computer Science, Scientific Computing and Bioinformatics, Johannes Gutenberg-University, Mainz, Germany, [3] Institute of Computer Science, High Performance Computing, Johannes Gutenberg-University, Mainz, Germany, [4] Next-Generation Sequencing Core Facility UMS2008 IBSLor CNRS-UL-INSERM, Biopôle, University of Lorraine, Vandœuvre-lès-Nancy, France, [5] IMoPA UMR7365 CNRS-UL, Biopôle, University of Lorraine, Vandœuvre-lès-Nancy, France

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Istituto Italiano di Tecnologia, Italy

Reviewed by:
Yang Zhao,
Vanderbilt University,
United States
Tommaso Leonardi,
University of Cambridge,
United Kingdom

****Correspondence:***
Andreas Hildebrandt
andreas.hildebrandt@uni-mainz.de
Bertil Schmidt
bertil.schmidt@uni-mainz.de
Mark Helm
mhelm@uni-mainz.de

† These authors share first authorship

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 June 2019*
Accepted: *21 August 2019*
Published: *25 September 2019*

Citation:
Schmidt L, Werner S, Kemmer T, Niebler S, Kristen M, Ayadi L, Johe P, Marchand V, Schirmeister T, Motorin Y, Hildebrandt A, Schmidt B and Helm M (2019) Graphical Workflow System for Modification Calling by Machine Learning of Reverse Transcription Signatures. Front. Genet. 10:876. doi: 10.3389/fgene.2019.00876

Modification mapping from cDNA data has become a tremendously important approach in epitranscriptomics. So-called reverse transcription signatures in cDNA contain information on the position and nature of their causative RNA modifications. Data mining of, e.g. Illumina-based high-throughput sequencing data, is therefore fast growing in importance, and the field is still lacking effective tools. Here we present a versatile user-friendly graphical workflow system for modification calling based on machine learning. The workflow commences with a principal module for trimming, mapping, and postprocessing. The latter includes a quantification of mismatch and arrest rates with single-nucleotide resolution across the mapped transcriptome. Further downstream modules include tools for visualization, machine learning, and modification calling. From the machine-learning module, quality assessment parameters are provided to gauge the suitability of the initial dataset for effective machine learning and modification calling. This output is useful to improve the experimental parameters for library preparation and sequencing. In summary, the automation of the bioinformatics workflow allows a faster turnaround of the optimization cycles in modification calling.

Keywords: RT signature, Watson–Crick face, Galaxy platform, RNA modifications, machine learning, m^{1}A

INTRODUCTION

In the rapidly growing field of epitranscriptomics (Saletore et al., 2012), the detection of RNA modifications is typically based on a combination of reagents and tools for wet work on the one hand, and bioinformatics processing of massive amounts of RNA-Seq data, on the other hand. Because of a sequence space that may include up to 10^7 nucleotides and more, transcriptomes must be scrutinized by computer-assisted detection schemes, resulting in what is called modification calling (Helm and Motorin, 2017).

With the exception of the up-and-coming nanopore direct RNA sequencing technology (Byrne et al., 2017; Garalde et al., 2018; Smith et al., 2019), RNA-Seq data are obtained after reverse transcription of the modified RNA template into DNA, a process during which information about modification type and position may get erased, partially or completely, since the newly synthesized cDNA is composed only of the four canonical deoxynucleotides. Attempts to circumvent this problem included, for example, the use of various chemical reagents, which specifically react with a given modification, to alter cDNA synthesis at sites of RNA modifications. One such reagent is CMCT, a carbodiimide leading to stalling of cDNA synthesis at sites of pseudouridine modification in the RNA template (Ofengand and Bakin, 1997; Carlile et al., 2014; Schwartz et al., 2014). Other modifications do not require chemical derivatization to alter cDNA synthesis. In particular, modifications with chemical alterations on their Watson–Crick face are liable to cause cDNA synthesis differing from that expected of an unmodified RNA template. A case in point is m^1A, a modification featuring a methyl group on the Watson–Crick face of adenosine, which interferes with proper base pairing, in RNA structure (Helm et al., 1998) (Helm et al., 1999) (Lempereur et al., 1985; Zhou et al., 2016), as well as during cDNA synthesis by reverse transcription (Motorin et al., 2007). In the particular case of m^1A, the resulting cDNA was shown to contain products of transcription arrest, i.e. abortive cDNA fragments, as well as misincorporation, most frequently of dATP being incorporated instead of dTTP at the position corresponding to the modification site. The ensemble of erroneous events in cDNA synthesis has been termed *reverse transcription signature* and was shown to depend on a number of factors including e.g. the nature of the penultimate base encountered by the RT enzyme before engaging the modified RNA residue (Hauenschild et al., 2015). The RT signature of m^1A can be experimentally altered e.g. by enzymatic demethylation with the AlkB enzyme (Zheng et al., 2015; Liu et al., 2016; Li et al., 2017) or at alkaline pH, which induces a Dimroth rearrangement to m^6A (Dominissini et al., 2016; Safra et al., 2017). Since these processes are relatively specific to m^1A, they can be exploited to increase confidence in modification calling, therein being used as a validation (Helm and Motorin, 2017).

All of the above processes require significant computing efforts to extract information on RNA modifications from RNA-Seq data. Given that the composition of RT signature of a given modification in terms of RT arrest, misincorporation, and even template nucleotide skipping ("jumps") (Ebhardt et al., 2009; Findeiss et al., 2011; Ryvkin et al., 2013; Hauenschild et al., 2015) is subject to variations caused by factors that are not fully characterized and thus cannot be entirely controlled, an innovative approach to account for a maximum of these features and exploit them for computer-based prediction ("modification calling") involves machine learning. A particular brand of machine learning, the random forest, was used for the purpose of modification by several groups, including us (Hauenschild et al., 2015).

Optimizing the performance of a modification calling protocol requires multiple rounds, beginning with a wet work part of library preparation and subsequent Illumina sequencing, as illustrated in **Figure 1A**. Here, a pretreatment (A1) of the samples by using auxiliary reagents such as the demethylase AlkB or changes in the library preparation part (A2), e.g. by employing different reverse transcriptase enzymes or variegated reaction conditions, are implemented experimentally. After sequencing (A3), a fast evaluation of their influence on the RT signature and consequently on RF performance (A4) is necessary to proceed with the next round of optimized library preparation in the wet lab. The associated computational data mining thus represents a bottleneck on the path to optimal modification calling.

To address this shortcoming, we here present an automated workflow implementation based on Galaxy (Afgan et al., 2018), whose components are depicted in **Figure 1B**. The Galaxy implementation provides a first module (B1) for the automation of typical and recurrent RNA-Seq–associated operations such as trimming and mapping. While these operations can be customized to accommodate a range of data formats, it allows procedurally stable and reproducible treatment of data package of comparable content, such as RNA-Seq data obtained under variegated conditions for library preparation. This, in turn, allows a comparative evaluation of those experimental conditions, as outlined above. The same holds true for subsequent modules (B2), designed and implemented following the requirement for fast comparison of data packages. The implemented tools allow to quantify mismatch, jump, and arrest rates in the relevant transcriptome, thus compiling RT signatures at single-nucleotide resolution. Still automatized, RT signatures of modified RNA nucleotides can be transferred as positive instances for machine learning, along with negative instances, i.e. signatures of unmodified nucleotides. Positive and negative instances are then used to train a Python-based random forest implementation of machine learning, and the performance of the trained machine in modification calling is evaluated and reported as a feedback in a further round of experimental optimization. Finally, with the implementation of a visualization module, graphics can be displayed and extracted for visual examination and comparison of individual sequence segments as well as the entire RNA fragments in a publishable manner.

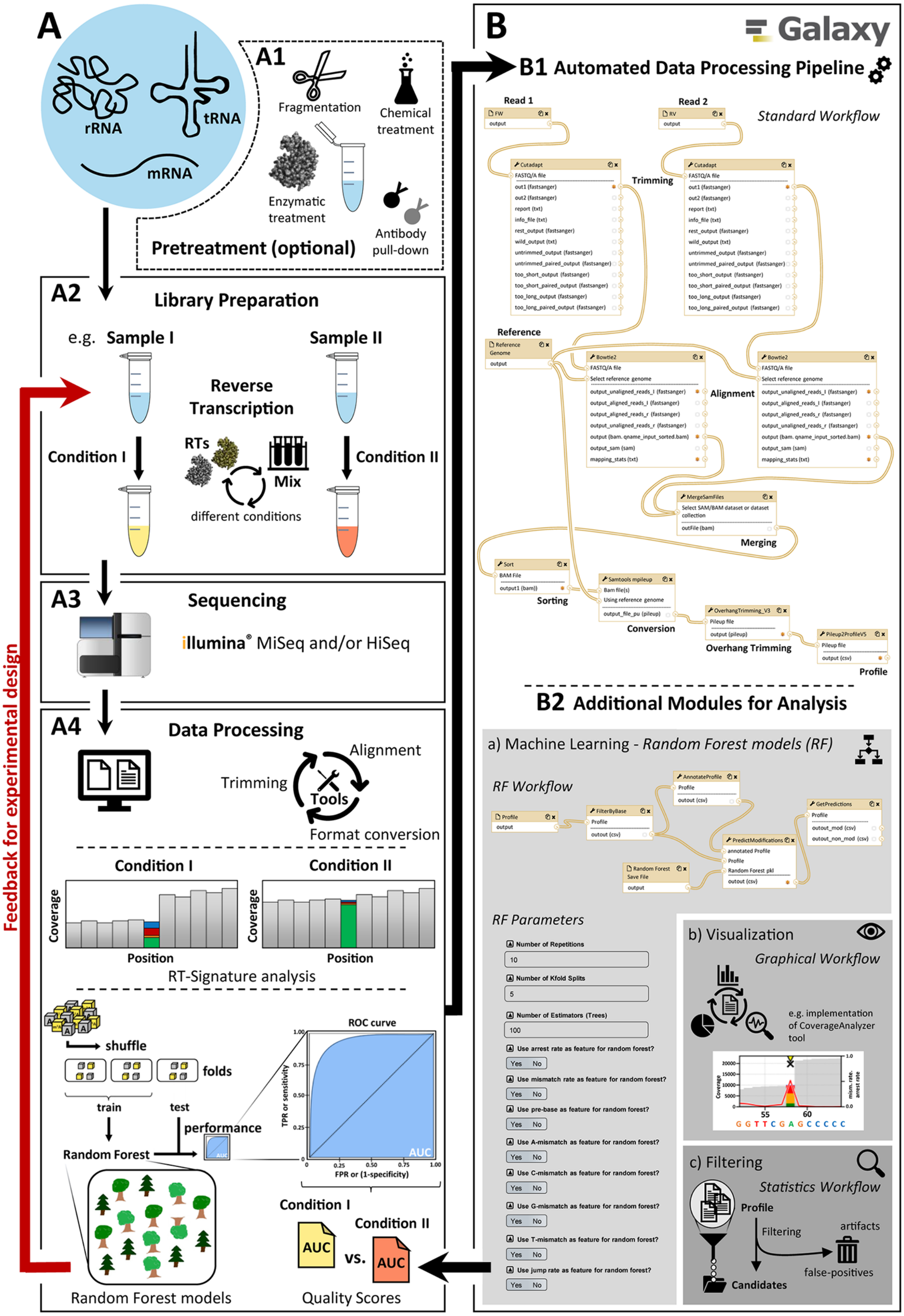
MATERIALS AND METHODS

RNA Sequencing Analysis

The present workflow serves as the main process for the analysis of RNA sequencing data in respect to the detection of several modifications. Its Galaxy distribution comes with a number of adjustable elements for variegated workflows, in which the particular element (Workflow *RNA_Seq_Standard_Workflow)* serves as basis for the remaining workflows and functionalities. Therefore, it is referred to as "standard workflow." The overall scheme of the workflow is illustrated in **Figure 1** (B1) and consists of the following steps:

Preprocessing of Raw Reads (Trimming)

The raw reads from the sequencing data (stored in fastq-format) are first subjected to removal of auxiliary sequences such as adapters, barcodes, and unique molecular identifiers (UMIs). For this task, the workflow uses the Cutadapt trimming software (Martin, 2011). Due to the necessity to remove multiple sequences from the raw reads, their respective arrangement, and the configuration of Cutadapt, the trimming is separated into



RF Parameters

Number of Repetitions: 10

Number of Kfold Splits: 5

Number of Estimators (Trees): 100

Use arrest rate as feature for random forest? Yes No

Use mismatch rate as feature for random forest? Yes No

Use pre-base as feature for random forest? Yes No

Use A-mismatch as feature for random forest? Yes No

Use C-mismatch as feature for random forest? Yes No

Use G-mismatch as feature for random forest? Yes No

Use T-mismatch as feature for random forest? Yes No

Use jump rate as feature for random forest? Yes No

FIGURE 1 | Main overview of the modification calling pipeline. A diagram showing the different steps for creating and analyzing RNA-Seq data. The pipeline has two parts: **(A)** general workflow for the processing of RNA samples and **(B)** the implemented automated graphical workflow system with the available modules for bioinformatics data analysis. **(A)** consists of (A1) possible and partly necessary pretreatments for different RNA species, (A2) library preparation with the possibility of adaptations (e.g. conditions for reverse transcription), (A3) sequencing with Illumina sequencing platforms (e.g. MiSeq/NextSeq and HiSeq), and (A4) data processing including basic data treatment like adapter trimming, alignment, and format conversion, as well as data analysis (e.g. machine learning and RT-signature analysis). The elaborate data processing (A4) was fully automated in **(B)** by using the open-source Galaxy platform to create and provide a quick and user-friendly feedback mechanism to optimize the experimental design, sample preparation, and data processing. The standard workflow (B1) is supplemented by various additional modules (B2) including workflows for (a) machine learning, (b) visualization, and (c) filtering.

multiple steps. In a typical Illumina paired-end sequencing run, the forward and reverse reads are stored in individual fastq files; the reads show slightly different characteristics concerning the auxiliary elements; hence, the trimming for forward and reverse reads is performed separately. The first substep in the trimming process consists of the removal of Illumina adapter sequences. In a second step, terminal barcode sequences and UMIs (Miner et al., 2004; McCloskey et al., 2007; Casbon et al., 2011) are cut from the raw reads.

Alignment

Mapping to a given sequence reference file is performed with Bowtie 2 (Langmead and Salzberg, 2012). Again, this process is performed separately for forward and reverse reads (–nofw/–norc option) and therefore in single-end mode. For the detection of RT-impairing modifications like m^1A, it is necessary to allow for mismatches (One mismatch ["N1"] allowed in seed length of 6 ["L6"]). Values are tailored toward tRNAs (e.g. high amounts of RT-impairing modifications). Additionally, if the evaluation is performed on samples containing a large number of modifications (affecting the RT), the amount of allowed mismatch occurrences has to be increased by adjusting the seed-length option (Bowtie standard parameters allow for one mismatch within a given seed; hence, seed length has to be decreased for highly modified samples). The alignment is stored in BAM format.

The two BAM files, one for the forward and one for the reverse reads correspondingly, are merged using the SAMtools (Li et al., 2009) "merge" function, and the aligned reads are sorted according to chromosomal coordinates.

File Conversion and Overhang Trimming

Further analysis steps require information of mapped reads at single base resolution for each position in the reference sequence, as every position is evaluated for mismatch and arrest properties. Accordingly, the BAM-file is converted into Pileup-format using the SAMtools (Li et al., 2009) "mpileup" function. As described in Tserovski et al. (2016), the library preparation includes a step in which C-tailing at the 3′ end of the cDNA strand was performed. Due to this tailing step in the library preparation protocol, despite the previous trimming steps, some tailing bases (overhangs) can remain and were then aligned with the reads. As these overhangs can impede the detection of modified sites, they have to be removed from the alignment. Therefore, a Python-based algorithm for postalignment manipulation was developed. This algorithm finds read-ending bases and compares them to reference base and removes them in case of a mismatch. After the overhang trimming, the data are still stored in Pileup format.

Feature Extraction

Information on each position of the reference is then extracted from the Pileup format and subsequently stored in a format termed "Profile" (example shown in **Table 1**). The information consists of the following features:

Arrest rate: Drop in coverage in relation to the preceding (N+1) position (arrest).

Mismatch rate: Relative amount of mapped nucleobases not matching the respective base in the reference (mismatch).

Jump rate: Relative amount of deletions (bases left out during reverse transcription) occurring at the given position in the reference (jump). A distinction is made between deletions at the given position in the reference (single jumps direct), deletions at the neighboring position (−1 position) (single jump delayed), and deletions at the given position, as well as the neighboring position (double jump).

In addition, the reference name (ref seg), reference base (refbase), reference position (pos), and coverage at the respective position (cov) are stored in the Profile. Also included is detailed information on the alignment numbers for each type of base (A, C, G, T) and unknown read bases (N), as well as the type of base preceding the position (prebase) in question.

In many cases, modified positions heavily differ from nonmodified positions in these key characteristics. Nonmodified bases are not expected to cause arrest and mismatch signals (at least not at high levels), making these features a main target for differentiation between modified and unmodified sites.

Downstream Analysis

The generation of the Profile file concludes the standard workflow. From this point on, the proceedings heavily vary depending on the question being investigated, with the Profile file serving as the starting point. Options for downstream analysis are shown in **Figure 1** (B2) and include the following:

Filtering

An option for further evaluation is a simple filtering process. Here, adenosine instances can be separated into two categories, namely, "likely m^1A" and "likely non-m^1A." The selectable filter criteria include threshold values for mismatch and arrest rates, minimum coverage, and the nucleobase of interest. In most cases, the arrest and mismatch rates should be sufficient to separate m^1As from non-m^1As.

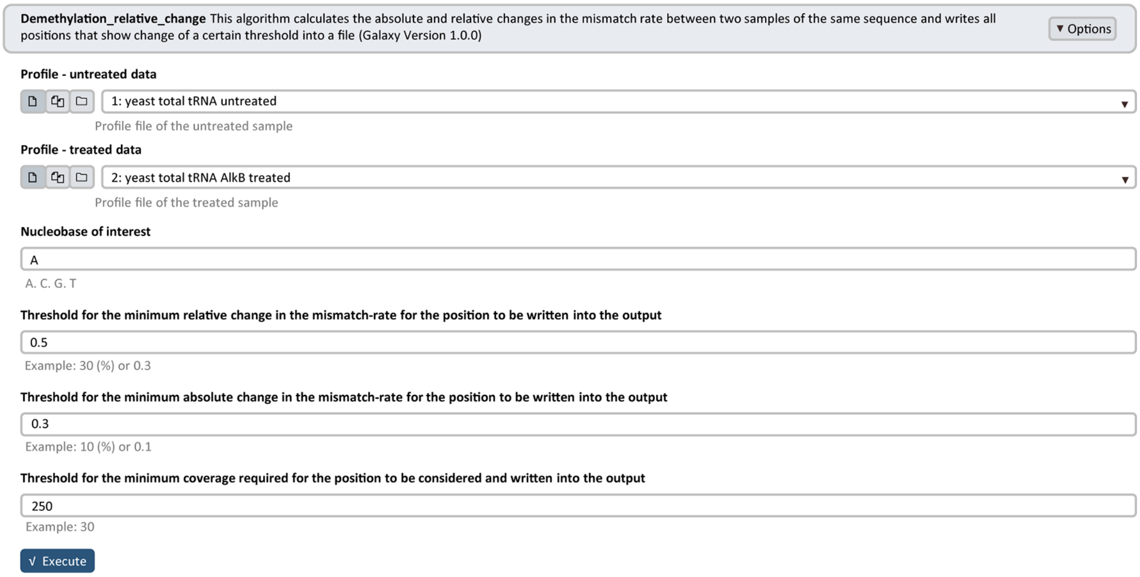
Another filtering option includes the comparison of two samples after different treatment. In our Galaxy pipeline, the sample comparison after enzymatic or chemical treatment is implemented wherein one sample serves as a reference (**Figure 2**). The algorithm calculates the absolute and relative changes in the mismatch rate between 2 samples for each position and filters by means of adjustable thresholds for changes and coverage. The resulting Profile file contains candidates filtered according to the selected thresholds. This module can be used for verification of modification candidates by e.g. applying enzymatic or chemical treatment to remove the alterations at the Watson–Crick face impeding reverse transcription and therefore decreasing the mismatch rate (exemplary analysis shown in Results section).

Machine Learning

For the prediction of m^1A and other modifications, a machine learning model for binary classification is included in the Galaxy distribution (Workflow *Workflow_Prediction*). The associated

TABLE 1 | Extracted Profile file after filtering with *Demethylation_relative_change* module with all m^1A candidate positions.

ref_seg	pos	refbase	cov	prebase	mismatch	A	G	T	C	N	a	g	t	c	n	single_ jump_ direct	single_ jump_ delayed	double_ jump	arrest
tdbR00000370|Saccharomyces_ cerevisiae|4932|Arg|TCT	57	A	699	C	0.29471	493	8	2	94	0	0	5	5	92	0	0.00000	0.02710	0.00285	0.10941
tdbR00000300|Saccharomyces_ cerevisiae|4932|Asn|GTT	59	A	961	C	0.37045	605	7	6	125	0	0	7	69	142	0	0.00000	0.00407	0.02238	0.15544
tdbR00000021|Saccharomyces_ cerevisiae|4932|Cys|GCA	57	A	405	T	0.21728	317	13	39	0	0	0	7	28	1	0	0.00000	0.00000	0.00000	0.43399
tdbM00000003|Saccharomyces_ cerevisiae|4932|Gln|TTG	57	A	475	A	0.15789	400	11	18	1	0	0	12	29	4	0	0.00000	0.00000	0.00000	0.26810
tdbR00000170|Saccharomyces_ cerevisiae|4932|Ile|AAT	59	A	919	T	0.38085	569	55	88	6	0	0	67	127	7	0	0.00429	0.00000	0.01072	0.15350
tdbM00000006|Saccharomyces_ cerevisiae|4932|Ile|TAT	58	A	373	T	0.25469	278	13	28	4	0	0	7	34	9	0	0.00000	0.00000	0.00000	0.31934
tdbR00000192|Saccharomyces_ cerevisiae|4932|Lys|CTT	58	A	2715	G	0.16317	2272	102	103	9	0	0	108	112	9	0	0.00037	0.00000	0.00293	0.07658
tdbR00000193|Saccharomyces_ cerevisiae|4932|Lys|TTT	58	A	619	G	0.43942	347	49	75	10	0	0	62	68	8	0	0.00478	0.00000	0.00955	0.16511
tdbR00000323|Saccharomyces_ cerevisiae|4932|Pro|TGG	57	A	459	T	0.43573	259	3	69	0	0	0	12	112	4	0	0.00000	0.00000	0.00000	0.18905
tdbR00000324|Saccharomyces_ cerevisiae|4932|Pro|TGG	57	A	439	T	0.43508	248	4	56	1	0	0	9	121	0	0	0.00000	0.00000	0.00000	0.20364
tdbR00000443|Saccharomyces_ cerevisiae|4932|Thr|AGT	58	A	396	A	0.28283	284	23	23	3	0	0	28	30	5	0	0.00000	0.00222	0.12195	0.38608
tdbR00000444|Saccharomyces_ cerevisiae|4932|Thr|AGT	58	A	616	A	0.31656	421	39	47	5	0	0	41	54	9	0	0.00145	0.00000	0.10320	0.30152
tdbR00000464|Saccharomyces_ cerevisiae|4932|Val|AAC	59	A	1066	T	0.18386	870	33	55	22	0	0	18	61	7	0	0.00187	0.00000	0.00094	0.69026



Demethylation_relative_change This algorithm calculates the absolute and relative changes in the mismatch rate between two samples of the same sequence and writes all positions that show change of a certain threshold into a file (Galaxy Version 1.0.0)

Profile - untreated data

1: yeast total tRNA untreated ▾

Profile file of the untreated sample

Profile - treated data

2: yeast total tRNA AlkB treated ▾

Profile file of the treated sample

Nucleobase of interest

A

A, C, G, T

Threshold for the minimum relative change in the mismatch-rate for the position to be written into the output

0.5

Example: 30 (%) or 0.3

Threshold for the minimum absolute change in the mismatch-rate for the position to be written into the output

0.3

Example: 10 (%) or 0.1

Threshold for the minimum coverage required for the position to be considered and written into the output

250

Example: 30

FIGURE 2 | Galaxy Filtering module *Demethylation_relative_change* interface. As input, two Profile files, yeast total tRNA untreated and yeast total tRNA AlkB treated, are used with the following selected parameters for filtering: adenosine (A) as nucleobase of interest, 0.5 or 50 (%) and 0.3 or 30 (%) as thresholds for the minimum relative and absolute changes in the mismatch rate and 250 as threshold for the minimum coverage required.

workflows for training and prediction are based on a random forest model from the "scikit-learn" Python package (Pedregosa et al., 2011). For the training process, the positive class (modified bases) and negative class (nonmodified bases) are given as input in a 1:1 ratio. This ratio is used in order to counter the tendency of RF models to bias toward the majority class. This RF property frequently leads to false negatives for the positive class (the modifications) when making predictions. Importantly, this bias is not necessarily reflected by the evaluation scores. The random forest performs e.g. 10 repetitions of a 5-fold cross-validation. These parameters can be adjusted as required for different models. The model's performance is measured by the area under the receiver operating characteristic curve. A detailed description of the concept of the random forest model used for this workflow can be found in Hauenschild et al. (2015). The prediction workflow requires a trained random forest model and a Profile file as input and performs a binary classification.

Visualization

A graphical representation of the position of interest within sequence context can be created using a Python-based script (Workflow *Visualize_V3*), extracted from the CoverageAnalyzer tool (Hauenschild et al., 2016). The user can plot a sequence containing up to 1000 bases where the leftmost and rightmost bases can be selected by position. In addition, various sizes can be adjusted, including the width and height of the plot, the font size, and the size of markers within the graphic (exemplary plot shown in **Figure 3**).

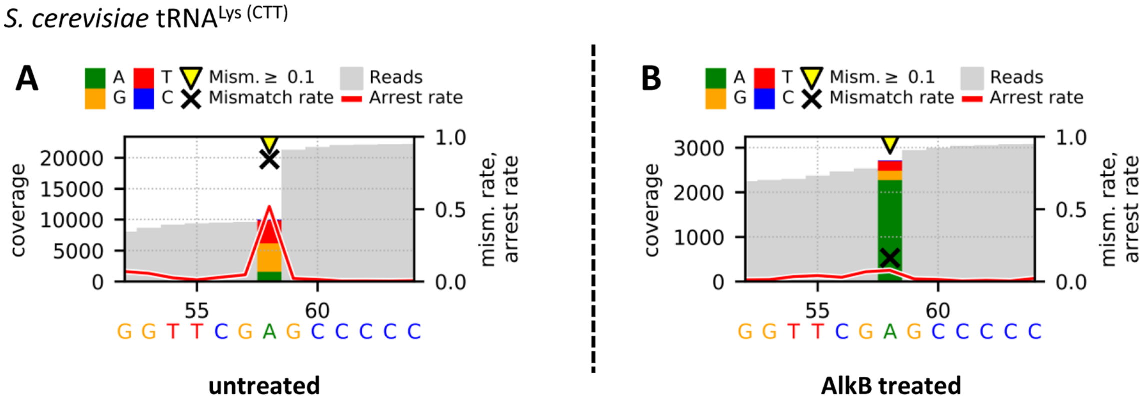


FIGURE 3 | Graphical plots of untreated **(A)** and AlkB-treated **(B)** yeast tRNA$^{Lys\ (CTT)}$ using the additional module *Visualize_V3* for visualization. Sites with error rates of more than 10% are highlighted with yellow arrows, with colored bars indicating the nature of the reads. Mismatch rates are depicted as black crosses, and arrest rates as red lines. The m^{1}A site is located in the middle of the shown sequence segment at position 58.

RNA Sequencing—Sample Preparation

Library Preparation and Sequencing

Sample preparation and sequencing are performed according to a previously published protocol (Hauenschild et al., 2015; Tserovski et al., 2016). This library preparation protocol includes the possibility to catch abortive products during the reverse transcription step, important for the detection of modifications impeding reverse transcription and generating a certain amount of RT stop products. The protocol also allows the adaptation of almost all necessary steps for preparation of RNA-Seq libraries, including adapter ligations, reverse transcription, and polymerase chain reaction. This allows fast screening of different conditions during sample preparation. Special experimental changes (e.g. buffer changes or pretreatment of the RNA) during library preparation for the preparation of our shown exemplary data are mentioned in the Results section.

RESULTS

Enzymatic Demethylation of m^1a Sites in Yeast tRNA With AlkB

In an exemplary sample processing, two samples of total tRNA from *Saccharomyces cerevisiae* were used for sample preparation, sequencing, data processing, and analysis. One of the samples had been subjected to pretreatment (**Figure 1** [A1]) with α-ketoglutarate–dependent dioxygenase AlkB that "repairs" alkylated DNA and RNA containing 3-methylcytosine (m^3C) or 1-methyladenine (m^1A) by oxidative demethylation. Protein preparation and sample treatment were performed according to a previously published protocol (Zheng et al., 2015). The second sample was used as reference. Both samples were then used as starting material for library preparation and subsequent sequencing (**Figure 1** [A2, A3]). Library preparation and sequencing were performed as described in our published workflow by Hauenschild et al. (2015) and Tserovski et al. (2016). The sequencing output data packages in FASTQ format were then processed with the standard automated Galaxy workflow *RNA_Seq_Standard_Workflow* (**Figure 1** [B1]) to create Profile files for downstream analysis.

Filtering for Demethylation Candidates

The Profile files were used for statistical analysis. **Figure 2** illustrates the Galaxy Filtering module *Demethylation_relative_change*, which was used to filter and extract all positions that show an absolute and relative change in the mismatch rate of a certain threshold between the untreated and AlkB-treated sample. **Table 1** shows the extracted Profile file with all candidate positions after filtering. From our sample comparison, with our selected thresholds, 13 candidate positions fulfilling the requirements were filtered out, with high probability to be m^1A sites.

Visualization of Demethylation Candidates

In addition, the Profile files were used in the visualization workflow *Visualize_V3* to obtain graphical plots for each sample. The visual comparison of the untreated (A) and AlkB-treated (B) yeast $tRNA^{Lys\ (CTT)}$, which includes an m^1A at position 58, is shown in **Figure 3**. The strong decreases of the mismatch and arrest rate from 0.845 and 0.518 to 0.163 and 0.077 after AlkB treatment at position 58 of the shown sequence segment indicate a successful removal of the methylation and therefore enabled valid reverse transcription. Such changes in the reverse transcription signature are considered as effective validation of the actual presence of m^1A at the considered position.

Influence of Mn^{2+} on the RT Signature at m^1A Sites in Yeast tRNA

In a second exemplary sample processing, four samples of total tRNA from *S. cerevisiae* were used for sample preparation, sequencing, data processing, and analysis. The samples were used for library preparation and differed in the reverse transcription step (**Figure 1** [A2]). For reverse transcription, we used SuperScript® III Reverse Transcriptase (Thermo Fisher Scientific, Germany) in four different buffer mixtures to investigate the influence of Mn^{2+} during reverse transcription (Zhou et al., 2018). Sample A served as a reference and was prepared according to the supplier's manual, using the standard RT buffer with Mg^{2+}. For the other three test samples, custom-made RT buffers, including the standard buffer components, and Mn^{2+} in different concentrations (0.5 mM [B], 1.0 mM [C] or 3.0 mM [D]) instead of Mg^{2+}, were used. Library preparation and sequencing were performed as described in our published workflow by Hauenschild et al. (2015) and Tserovski et al. (2016). The sequencing output data packages in FASTQ format were then processed with the standard automated Galaxy workflow *RNA_Seq_Standard_Workflow* (**Figure 1** [B1]) to create Profile files for downstream analysis.

Visualization of $tRNA^{Asn\ (GTT)}$ Using Mg^{2+} or Mn^{2+} as Buffer Components for Reverse Transcription During Library Preparation

The Profile files were used in the visualization workflow *Visualize_V3* to obtain graphical plots for each sample. The visual comparison of the reference (**Figure 4A**) and the Mn^{2+} (0.5 mM [**Figure 4B**], 1.0 mM [**Figure 4C**], or 3.0 mM [**Figure 4D**]) yeast $tRNA^{Asn\ (GTT)}$ samples, including an m^1A at position 59, is shown in **Figure 4**. The high mismatch rates (≥90%) throughout all samples are driven by the prebase influence (Hauenschild et al., 2015), leading to a consistently high C mismatch. Considering the m^1A at position 59, the strong decrease in the arrest rate at position 59 from 0.846 (A) over 0.869 (B) and 0.704 (C) down to 0.070 (D) indicates an increasing read-through capability of the reverse transcriptase due to a stabilizing effect by increased Mn^{2+} concentrations. In addition, by exchanging Mg^{2+} through Mn^{2+}, the number of jumps (single_jump_direct, single_jump_delayed, double_jump) increases with higher Mn^{2+} concentrations, visible in **Table 2**, as well as in the graphical plots by coverage drops (through deletions/jumps), especially visible in **Figure 4D**.

DISCUSSION

We here present a versatile, user-friendly graphical workflow system for modification calling to analyze RNA-Seq data. It can also be used to analyze any high-throughput data as long as they follow the formats listed in this technology report. Although this package allows creation and implementation of various workflows for processing and analysis, the application of this

TABLE 2 | Extracted Profile data for yeast tRNA$^{Asn (GTT)}$ after library preparation with 4 different buffer mixtures for the reverse transcription step. Shown are data for positions 58, 59 (m^{1}A), and 60.

ref_seg	pos	refbase	cov	prebase	mismatch	A	G	T	C	N	a	g	t	c	n	single_jump_direct	single_jump_delayed	double_jump	arrest
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT Reference	58	A	3238	A	0.02471	3158	4	4	33	2	0	5	7	25	0	0.01927	0.00056	0.00000	0.4574
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 0.5 mM Mn	58	A	1380	A	0.04855	1313	4	1	47	3	0	2	0	10	0	0.02404	0.00000	0.00060	0.32355
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 1.0 mM Mn	58	A	3546	A	0.04061	3402	15	9	79	0	0	13	6	22	0	0.02913	0.00000	0.00067	0.14965
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 3.0 mM Mn	58	A	2239	A	0.04332	2142	9	6	37	7	0	12	5	21	0	0.05623	0.00172	0.00138	0.0565
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT Reference	59	A (m^{1}A)	6311	C	0.90160	621	79	36	3431	6	0	119	25	1994	0	0.00000	0.01048	0.04161	0.84647
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 0.5 mM Mn	59	A (m^{1}A)	2210	C	0.93167	151	37	59	1238	8	0	37	15	665	0	0.00041	0.01630	0.09902	0.86879
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 1.0 mM Mn	59	A (m^{1}A)	4454	C	0.95757	189	65	95	2208	1	0	75	35	1786	0	0.00038	0.02481	0.14907	0.70422
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 3.0 mM Mn	59	A (m^{1}A)	2568	C	0.96145	99	9	9	1149	14	0	7	5	1276	0	0.00000	0.05323	0.16101	0.06965
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT Reference	60	C	42890	C	0.00445	87	30	22	42699	21	20	10	1	0	0	0.00000	0.00000	0.00000	0.36943
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 0.5 mM Mn	60	C	18703	C	0.00733	51	12	10	18566	50	11	3	0	0	0	0.00000	0.00005	0.00000	0.43528
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 1.0 mM Mn	60	C	17706	C	0.00345	17	7	10	17645	10	9	6	2	0	0	0.00006	0.00011	0.00011	0.35852
tdbR00000300|Saccharomyces_cerevisiae|4932|Asn|GTT 3.0 mM Mn	60	C	3287	C	0.01156	2	1	9	3249	14	5	5	2	0	0	0.00000	0.00000	0.00030	0.03294

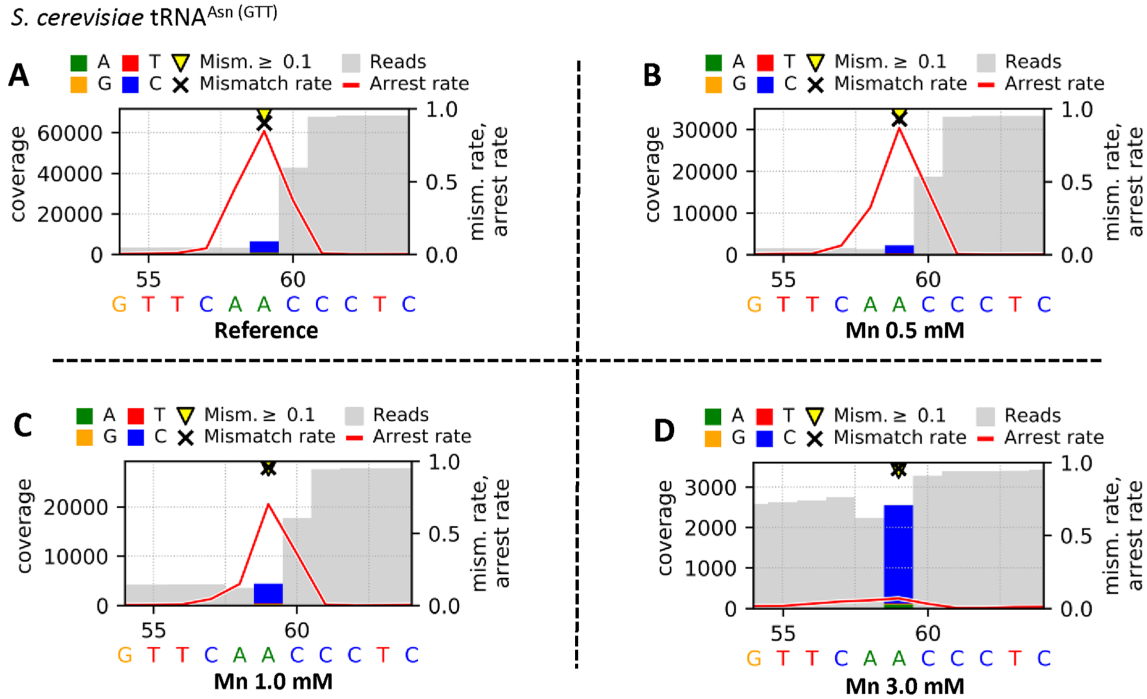


FIGURE 4 | Graphical plots of yeast tRNA$^{Asn (GTT)}$, which was used for library preparation, visualized by using the additional module *Visualize_V3*. The reverse transcription step was performed by using SuperScript® III Reverse Transcriptase in different reaction buffers. The supplier's standard reaction buffer (First Strand Synthesis buffer) with Mg^{2+} serves as reference **(A)**, and the tested buffer mixtures differ by increased concentrations of Mn^{2+} [0.5 mM **(B)**, 1.0 mM **(C)**, 3.0 mM **(D)**] as Mg^{2+} substitute. Sites with error rates of more than 10% are highlighted with yellow arrows, with colored bars indicating the nature of the reads. Mismatch rates are depicted as black crosses, and arrest rates as red lines. The m1A site is located in the middle of the shown sequence segment at position 59.

pipeline has limitations, which we would like to indicate hereafter and to point out possible solutions for adjustment.

Limitations and Adjustability

The limitations of the workflow pertain mostly to the specific characteristics of the library preparation protocol. The workflow is tailored to the analysis of short RNA sequences, mostly tRNAs, and uses a "splice unaware" alignment because in the examples given, splicing is irrelevant. Accordingly, analysis of transcriptomic data should use an alignment tool that is specifically tailored to mapping of splice variants ("splice aware").

Furthermore, algorithms such as the overhang trimming are not optimized for parallelization, which can lead to very long runtimes for the analysis, a problem potentially exacerbated by the large size of transcriptomic input data. Of course, as this Galaxy distribution makes use of the local computer's processing power, large-scale analysis should not be performed on a device with weak computing capabilities. This Galaxy distribution, developed in a Unix environment, has not been tested on Windows platforms.

Detection efficiency of modified ribonucleotides is highly dependent on the dataset. tRNA samples show a high number of RT-impairing modifications, which can negatively affect the RT signals for surrounding positions, making it more difficult to detect modified positions of interest through filtering or machine learning. We also observed that detectability is highly dependent on read coverage. In some cases, modified low-coverage sites could not be detected as the RT signatures were noisy and thus not very pronounced. Moreover, the machine learning and prediction processes require an adequate number of training instances for a given modification. Modifications that are present only in low amounts are not compatible with the available machine learning process. Lastly, the workflow here presented was created and optimized to detect modifications, which naturally impair reverse transcription. However, this does not preclude modifications, which are made accessible for analysis through changes in the structural or chemical characteristics in a pretreatment by generating RT events like increased mismatch and arrest rates. Examples include the generation of RT signatures for N^6-methyladenosine (m^6A) with an engineered polymerase with reverse transcriptase activity to induce mutations at m^6A sites (Aschenbrenner et al., 2018), the enzymatic introduction of a bio-orthogonal propargyl group to trigger RT termination for m^6A detection (Hartstock et al., 2018), and the site-specific installation of an allyl group to the N^6-position of adenosines, spontaneously inducing the formation of N^1,N^6-cyclized adenosine by iodination to create mutations to differentiate m^6A, which is inert to allyl labeling, from adenosines at individual RNA sites (Shu et al., 2017).

While the available workflows were tailored toward our specific library preparation protocol and were created with the goal of detecting m^1A, the workflows are easily adjustable for analysis of other modifications and other protocols. For example, the standard workflow also works without the overhang-trimming step, which allows the user to remove this step when using other library preparation protocols. In addition, the Galaxy interface allows for user-friendly customization of many input parameters. The customization is not limited to the software packages such as Cutadapt (Martin, 2011) and Bowtie (Langmead and Salzberg, 2012), but also includes individual Python scripts for the multiple workflows. Accordingly, adapter and barcode sequences can be replaced to fit the library preparation protocol, and other tasks like quality trimming can be performed. For the Python scripts, the range of adjustable parameters allows the user to change the modification of interest, filter criteria, features, and parameters for the machine learning model as well as several options for the visualization.

Furthermore, existing workflows can be easily rearranged to suit the desired analysis. The associated Galaxy toolshed allows for the installation of additional bioinformatics programs and enables the user to create entirely new workflows. For example, other alignment tools can be implemented that may improve or accelerate data processing or allow transcriptome-wide analysis for other data packages. In the provided tutorial, the installation of new software is described. As an example, we have incorporated the CUSHAW2 tool (Liu et al., 2012), which allows significant acceleration of the alignment speed, as a substitute for Bowtie 2. Our performance assessment showed that the alignment process could be sped up by a factor of up to six of the same datasets and on the same hardware platform. By reducing the time of the rather costly alignment step of the pipeline, it is possible to increase overall throughput. In return, the analysis of larger datasets is feasible within the same time in order to further increase the accuracy of the obtained results.

CONCLUSION/SUMMARY

Machine learning as an efficient tool for data mining is currently receiving enormous attention, which also extends to high-throughput sequencing data. Based on previous progress in machine learning for modification calling (Hauenschild et al., 2015), we here present a workflow that not only automatizes all steps, but which also, in principle, allows adaptation to "nonnatural" modifications, i.e. bioconjugate derivatives of RNA nucleotides after treatment with a chemical reagent or enzymes (Ofengand and Bakin, 1997; Carlile et al., 2014; Schwartz et al., 2014; Shu et al., 2017; Hartstock et al., 2018). In the course of development of reagent- and enzyme-based mapping procedures, repeated cycles of optimization, e.g. of reaction conditions, are necessary, but an assessment of modification calling performance for a given set of reaction conditions is extremely time consuming. The workflow here presents a solution to this bottleneck; while developed using the naturally occurring modification m1A as an example, it is conceived as such to be easily adaptable to the development of chemical reagents for modification mapping.

DATA AVAILABILITY

The graphical workflow system, an instruction manual, and a tutorial are available at: https://github.com/HelmGroup, Repository: Galaxy_modification_calling.

Operating system(s): Linux, Programming language for custom scripts: Python, Other requirements: Docker (software) needs to be installed.

The AlkB test datasets analyzed and generated for this study can be found in the repository: Galaxy_modification_calling (https://github.com/HelmGroup/Galaxy_modification_calling/tree/master/TestData/AlkB).

Compressed files are provided in PKZIP and ZIP format and were compressed with 7-Zip.

Files: total_tRNA_yeast_untreated_R1.fastq (untreated yeast total tRNA – Read 1)

total_tRNA_yeast_untreated_R2.fastq (untreated yeast total tRNA – Read 2)

total_tRNA_yeast_AlkB_treated_R1.fastq (AlkB-treated yeast total tRNA – Read 1)

total_tRNA_yeast_AlkB_treated_R2.fastq (AlkB-treated yeast total tRNA – Read 2)

total_tRNA_yeast_untreated.profile (untreated yeast total tRNA – Profile)

total_tRNA_yeast_AlkB_treated.profile (AlkB-treated yeast total tRNA – Profile)

total_tRNA_yeast_reference.fasta (Reference total tRNA yeast)

Files for testing of the machine learning workflow can be found in the repository: Galaxy_modification_calling (https://github.com/HelmGroup/Galaxy_modification_calling/tree/master/TestData/Prediction).

Files: Known_m1A_sites_yeast (list of known m^{1}A sites)

total_tRNA_yeast_untreated.profile (untreated yeast total tRNA – Profile)

All other data are available from the corresponding authors upon reasonable request.

AUTHOR CONTRIBUTIONS

Conception and design: LS, SW, and MH; biomolecular experiments: SW, MK, and PJ; sequencing service: LA, VM, and YM; analysis and interpretation of the data: LS, SW, MK and MH; development and testing of the Galaxy modules: LS, TK, SN, BS, MK and AH; writing of the paper: LS, SW, and MH; proofreading and discussion: TS, BS, and AH.

FUNDING

This work was supported by DFG grants HE3397/13-2, by DIP Grant RE 4193/1-1/RO 4681/6-1, and by JPND "RNA NEURO"/Bmbf grant FKZ: 01ED1804 and by the EPITRAN COST initiative (CA16120).

ACKNOWLEDGMENTS

We are grateful to Valérie Igel-Bourguignon (NGS Core Facility, Biopôle, University of Lorraine, Nancy, France) for excellent technical assistance. We further thank the help of Hannah Back, BSc (Biomedical Chemistry, Johannes Gutenberg-University, Mainz, Germany), for experimental assistance, as well as Dr. Lyudmil Tserovski and Dr. Ralf Hauenschild for their contributions during the early stages of this project. The pET30a-AlkB plasmid was a gift from Tao Pan (Addgene plasmid #79050; http://n2t.net/addgene:79050; RRID: Addgene_79050).

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Aschenbrenner, J., Werner, S., Marchand, V., Adam, M., Motorin, Y., Helm, M., et al. (2018). Engineering of a DNA polymerase for direct m^{6}A sequencing. *Angew. Chem. Int. Edit.* 57, 417–421. doi: 10.1002/anie.201710209

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi: 10.1038/ncomms16027

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 515, 143–146. doi: 10.1038/nature13802

Casbon, J. A., Osborne, R. J., Brenner, S., and Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* 39, e81. doi: 10.1093/nar/gkr217

Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., et al. (2016). The dynamic N^1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* 530, 441–446. doi: 10.1038/nature16998

Ebhardt, H. A., Tsang, H. H., Dai, D. C., Liu, Y., Bostan, B., and Fahlman, R. P. (2009). Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* 37, 2461–2470. doi: 10.1093/nar/gkp093

Findeiss, S., Langenberger, D., Stadler, P. F., and Hoffmann, S. (2011). Traces of post-transcriptional RNA modifications in deep sequencing data. *Biol. Chem.* 392, 305–313. doi: 10.1515/bc.2011.043

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi: 10.1038/nmeth.4577

Hartstock, K., Nilges, B. S., Ovcharenko, A., Cornelissen, N. V., Püllen, N., Lawrence-Dörner, A.-M., et al. (2018). Enzymatic or *in vivo* installation of propargyl groups in combination with click chemistry for the enrichment and detection of methyltransferase target sites in RNA. *Angew. Chem. Int. Edit.* 57, 6342–6346. doi: 10.1002/anie.201800188

Hauenschild, R., Tserovski, L., Schmid, K., Thuring, K., Winz, M. L., Sharma, S., et al. (2015). The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* 43, 9950–9964. doi: 10.1093/nar/gkv895

Hauenschild, R., Werner, S., Tserovski, L., Hildebrandt, A., Motorin, Y., and Helm, M. (2016). CoverageAnalyzer (CAn): a tool for inspection of modification signatures in RNA sequencing profiles. *Biomolecules* 6, 42. doi: 10.3390/biom6040042

Helm, M., and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.* 18, 275–291. doi: 10.1038/nrg.2016.169

Helm, M., Brulé, H., Degoul, F., Cepanec, C., Leroux, J. P., Giegé, R., et al. (1998). The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial tRNA. *Nucleic Acids Res.* 26, 1636–1643. doi: 10.1093/nar/26.7.1636

Helm, M., Giege, R., and Florentz, C. (1999). A Watson–Crick Base-Pair-Disrupting Methyl Group (m′A9′) is sufficient for cloverleaf folding of human mitochondrial tRNA(lys). *Biochemistry* 38, 13338–133346. doi: 10.1021/bi991061g

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B., and Bachellerie, J. P. (1985). Conformation of yeast 18S rRNA. Direct chemical probing of the 5′ domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate–accessible. *Nucleic Acids Res.* 13, 8339–8357. doi: 10.1093/nar/13.23.8339

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, X., Xiong, X., Zhang, M., Wang, K., Chen, Y., Zhou, J., et al. (2017). Base-resolution mapping reveals distinct m(1)A methylome in nuclear- and mitochondrial-encoded transcripts. *Mol. Cell* 68, 993–1005, e1009. doi: 10.1016/j.molcel.2017.10.019

Liu, F., Clark, W., Luo, G., Wang, X., Fu, Y., Wei, J., et al. (2016). ALKBH1-mediated tRNA demethylation regulates translation. *Cell* 167, 1897. doi: 10.1016/j.cell.2016.11.045

Liu, Y., Schmidt, B., and Maskell, D. L. (2012). CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows–Wheeler transform. *Bioinformatics* 28, 1830–1837. doi: 10.1093/bioinformatics/bts276

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10. doi: 10.14806/ej.17.1.200

McCloskey, M. L., Stoger, R., Hansen, R. S., and Laird, C. D. (2007). Encoding PCR products with batch-stamps and barcodes. *Biochem. Genet.* 45, 761–767. doi: 10.1007/s10528-007-9114-x

Miner, B. E., Stoger, R. J., Burden, A. F., Laird, C. D., and Hansen, R. S. (2004). Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res.* 32, e135. doi: 10.1093/nar/gnh132

Motorin, Y., Muller, S., Behm-Ansmant, I., and Branlant, C. (2007). Identification of modified residues in RNAs by reverse transcription–based methods. *Methods Enzymol.* 425, 21–53. doi: 10.1016/S0076-6879(07)25002-5

Ofengand, J., and Bakin, A. (1997). Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.* 266, 246–268. doi: 10.1006/jmbi.1996.0737

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Ryvkin, P., Leung, Y. Y., Silverman, I. M., Childress, M., Valladares, O., Dragomir, I., et al. (2013). HAMR: high-throughput annotation of modified ribonucleotides. *RNA (New York, NY)* 19, 1684–1692. doi: 10.1261/rna.036806.112

Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., et al. (2017). The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* 551, 251–255. doi: 10.1038/nature24456

Saletore, Y., Meyer, K., Korlach, J., Vilfan, I. D., Jaffrey, S., and Mason, C. E. (2012). The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* 13, 175. doi: 10.1186/gb-2012-13-10-175

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., Leon-Ricardo, B. X., et al. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* 159, 148–162. doi: 10.1016/j.cell.2014.08.028

Shu, X., Dai, Q., Wu, T., Bothwell, I. R., Yue, Y., Zhang, Z., et al. (2017). N6-allyladenosine: a new small molecule for RNA labeling identified by mutation assay. *J. Am. Chem. Soc.* 139, 17213–17216. doi: 10.1021/jacs.7b06837

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* 14, e0216709. doi: 10.1371/journal.pone.0216709

Tserovski, L., Marchand, V., Hauenschild, R., Blanloeil-Oillo, F., Helm, M., and Motorin, Y. (2016). High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA. *Methods (San Diego, Calif.)* 107, 110–121. doi: 10.1016/j.ymeth.2016.02.012

Zheng, G., Qin, Y., Clark, W. C., Dai, Q., Yi, C., He, C., et al. (2015). Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* 12, 835–837. doi: 10.1038/nmeth.3478

Zhou, H., Kimsey, I. J., Nikolova, E. N., Sathyamoorthy, B., Grazioli, G., McSally, J., et al. (2016). m(1)A and m(1)G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* 23, 803–810. doi: 10.1038/nsmb.3270

Zhou, K. I., Clark, W. C., Pan, D. W., Eckwahl, M. J., Dai, Q., and Pan, T. (2018). Pseudouridines have context-dependent mutation and stop rates in high-throughput sequencing. *RNA Biol.* 15, 892–900. doi: 10.1080/15476286.2018.1462654

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 November 2019
doi: 10.3389/fgene.2019.01181

The Identification of Differentially Expressed Genes Showing Aberrant Methylation Patterns in Pheochromocytoma by Integrated Bioinformatics Analysis

Dengqiang Lin[1†], Jinglai Lin[1†], Xiaoxia Li[2†], Jianping Zhang[1], Peng Lai[1], Zhifeng Mao[1], Li Zhang[1], Yu Zhu[3] and Yujun Liu[1*]*

[1] Department of Urology, Xiamen Hospital of Zhongshan Hospital, Fudan University, Xiamen, China, [2] Department of Radiology, Xiamen Hospital of Zhongshan Hospital, Fudan University, Xiamen, China, [3] Department of Urology, Ruijin Hospital, Medical School of Shanghai Jiaotong University, Shanghai, China

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Italian Institute of Technology (IIT),
Italy

Reviewed by:
Nitish Kumar Mishra,
University of Nebraska Medical Center, United States
Andrew Dellinger,
Elon University, United States

***Correspondence:**
Yu Zhu
zhuyuruijin@163.com
Yujun Liu
liuyj12018@163.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 July 2019*
Accepted: *24 October 2019*
Published: *15 November 2019*

Citation:
Lin D, Lin J, Li X, Zhang J, Lai P, Mao Z, Zhang L, Zhu Y and Liu Y (2019) The Identification of Differentially Expressed Genes Showing Aberrant Methylation Patterns in Pheochromocytoma by Integrated Bioinformatics Analysis. Front. Genet. 10:1181. doi: 10.3389/fgene.2019.01181

Malignant pheochromocytoma (PHEO) can only be fully diagnosed when metastatic foci develop. However, at this point in time, patients gain little benefit from traditional therapeutic methods. Methylation plays an important role in the pathogenesis of PHEO. The aim of this research was to use integrated bioinformatics analysis to identify differentially expressed genes (DEGs) showing aberrant methylation patterns in PHEO and therefore provide further understanding of the molecular mechanisms underlying this condition. Aberrantly methylated DEGs were first identified by using R software (version 3.6) to combine gene expression microarray data (GSE19422) with gene methylation microarray data (GSE43293). An online bioinformatics database (DAVID) was then used to identify all overlapping DEGs showing aberrant methylation; these were annotated and then functional enrichment was ascertained by gene ontology (GO) analysis. The online STRING tool was then used to analyze interactions between all overlapping DEGs showing aberrant methylation; these results were then visualized by Cytoscape (version 3.61). Next, using the cytoHubba plugin within Cytoscape, we identified the top 10 hub genes and found that these were predominantly enriched in pathways related to cancer. Reference to The Cancer Genome Atlas (TCGA) further confirmed our results and further identified an upregulated hypomethylated gene (*SCN2A*) and a downregulated hypermethylated gene (*KCNQ1*). Logistic regression analysis and receiver operating characteristic (ROC) curve analysis indicated that *KCNQ1* and *SCN2A* represent promising differential diagnostic biomarkers between benign and malignant PHEO. Finally, clinical data showed that there were significant differences in the concentrations of potassium and sodium when compared between pre-surgery and post-surgery day 1. These suggest that *KCNQ1* and *SCN2A*, genes that encode potassium and sodium channels, respectively, may serve as putative diagnostic targets for the diagnosis and prognosis of PHEO and therefore facilitate the clinical management of PHEO.

Keywords: pheochromocytoma, bioinformatics, expression, methylation, KCNQ1, SCN2A

INTRODUCTION

Pheochromocytoma (PHEO) arises from the extra-adrenal sympathetic and parasympathetic ganglia (also referred to as the paraganglioma), as well as the intra-adrenal medulla. This tumor is rare, with a reported incidence of 1 in 300,000 (Else et al., 1993; Lefebvre and Foulkes 2014; Lenders et al., 2014). However, PHEO is a frequent cause of secondary hypertension, a potentially life-threatening cardiovascular complication (Zelinka et al., 2012; Prejbisz et al., 2013). Clinical reports show that up to 36% of patients develop malignancy (Pacak et al., 2007). On the other hand, reports from autopsy research estimate that 0.05–0.1% of cases remain undiagnosed (Jain et al., 2019). Current guidelines for the early treatment of PHEO recommend radical surgical resection. The 5-year survival rate post-surgery in benign cases of PHEO ranges from 84% to 96%, but is less than 50% in malignant cases; the recurrence rate can be as high as 65.45 within 5 years (Schurmeyer et al., 1988; Walther et al., 1999; Kopf et al., 2001; Edstrom Elder et al., 2003). Once PHEO enters an advanced stage, effective treatment modalities are limited, but include radionuclide therapy (^{131}I-MIBG) (van Hulsteijn et al., 2014), chemotherapy (a combination of cyclophosphamide, vincristine, and dacarbazine) (Vogel et al., 2014), and external beam radiation therapy (Vogel et al., 2014). However, patients suffering from the advanced stages of PHEO gain little benefit from such treatment modalities. Therefore, there is an urgent need to investigate the key genes involved in the progression of this disease. The identification of new biomarkers could help us to improve the prognosis of patients and facilitate clinical management.

Research studies have identified germline mutations in around one third of patients with PHEO (Lenders and Eisenhofer, 2017) and have identified a range of susceptibility genes, including *RET*, *HIF2A*, *VHL*, *NF1*, *SDHx* (*SDHA*, *SDHB*, *SDHC*, *SDHD*, *SDHAF2*), *FH*, *TMEM127*, and *MAX* (Wallace et al., 1990; Latif et al., 1993; Mulligan et al., 1993; Baysal et al., 2000; Niemann and Muller, 2000; Astuti et al., 2001; Hao et al., 2009; Burnichon et al., 2010; Qin et al., 2010; Comino-Mendez et al., 2011; Castro-Vega et al., 2014). Although genomic variation appears to occur more commonly in PHEO than in any other human tumors (Karagiannis et al., 2007; Fishbein and Nathanson, 2012), research has failed to identify specific genes related to carcinogenesis. Over recent years, the use of microarrays and sequencing has become a promising and effective technique with which to screen hub disease-causing genes and identify biomarkers of diagnostic, prognostic, and therapeutic value. To our knowledge, a complete bioinformatic analysis of PHEO, using the Gene Expression Omnibus (GEO) database and The Cancer Genome Atlas (TCGA), has yet to be carried out, particularly with regards to gene expression and methylation.

In this study, we first identified and screened differentially expressed genes (DEGs) showing aberrant methylation in PHEO by combining gene expression microarray data (GSE19422) and gene methylation microarray data (GSE43293). We then identified 10 core genes showing differential expression and aberrant methylation to act as suitable candidates for further interaction network analysis. TCGA was then used to verify the expression of these core genes and investigate their prognostic value. Our overall goal was to explore new genetic targets that may help us to improve patient outcomes.

MATERIALS AND METHODS

Microarray Data

Two gene expression profiles were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/): platform GPL6480—Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (GSE19422, including 84 PHEO tissues and six normal adrenal tissues); and the gene methylation dataset—Illumina HumanMethylation450 arrays (GSE43293, including 22 PHEO tissues and two normal adrenal tissues).

Data Processing

All aberrantly methylated DEGs were analyzed with R software (version 3.6) (https://www.r-project.org/). For DEGS, we used a |log(fold change [FC])| value >1 and an adjusted *P* value <0.05 as cutoff criteria following normalization and background correction with the affyPLM package in R. Data relating to aberrantly methylated genes were first normalized using the beta-mixture quantile dilation (BMIQ) method in the R wateRmelon package. We then used a β value >0.2 and an adjusted *P* value <0.05 as cutoff standards.

Gene Ontology Functional Enrichment Analysis

An online bioinformatics database (DAVID, Database for Annotation, Visualization, and Integrated Discovery, https://david.ncifcrf.gov/) was used to identify all overlapping DEGs showing aberrant methylation. These were annotated and then functional enrichment was ascertained by gene ontology (GO) analysis, including biological processes (BP), molecular function (MF), and cellular component (CC) (Consortium, 2006; Huang da et al., 2009). The GO functional enrichment results were visualized using the ggplot2 package in R.

Protein–Protein Interaction Network and Module Analysis

The online STRING tool (http://string-db.org) (Park et al., 2009) was used to search for potential correlations among the overlapping DEGs showing aberrant methylation. Cytoscape software (version 3.61; https://cytoscape.org) (Haffner et al., 2017) was then used to build a protein–protein interaction (PPI) network and analyze potential interactions. The cytoHubba plugin and the maximal clique centrality (MCC) method were then used to identify the top 10 hub genes. We then used the MCODE plugin to screen core modules of the PPI network with a standard degree cutoff of 2, a node score cutoff of 0.2, a k-core of 2, and a maximum depth of 100.

Expression Analysis of Candidate Genes in TCGA

The cBioPortal (https://www.cbioportal.org/) and UCSC Xena (http://xena.ucsc.edu/welcome-to-ucsc-xena/) platforms, in combination with the TCGA database (TCGA-PCPG), were used to analyze genetic alterations, gene expression levels, and the relationship between expression and methylation. In total, TCGA featured 184 datasets that were available for methylation and expression analysis. We also used the Human Protein Atlas (HPA) database to investigate the expression levels of candidate genes in normal adrenal tissues.

Kaplan–Meier Survival Analysis for Candidate Genes in TCGA

The Kaplan–Meier plotter (http://www.kmplot.com/) was used to determine the prognostic value of candidate genes in TCGA. *P* values <0.05 were considered to be statistically significant.

Clinical Information

With the approval of our institutional ethics review board, we collected clinical information, including tumor size and biochemical data, from 136 patients who underwent adrenalectomy and were subsequently diagnosed with PHEO following surgery. The clinical data (**Supplementary Table 1**) were collected between January 2016 and May 2019 from the Department of Urology in Ruijin Hospital affiliated to the Medical School of Shanghai Jiaotong University in China.

Statistical Analysis

All data are presented as means ± standard deviation. Statistical analyses were performed with SPSS software (version 23.0;IBM). Bar graphs and scatter diagrams were created by GraphPad Prism 7 software. Data analysis and correlation were carried out using paired *t* tests and either Pearson's or Spearman's correlation analysis, as well as line regression analysis. Outliers were analyzed using Spearman's correlation analysis. We then created a logistic model featuring two selected variables, the expression levels of KCNQ1 and SCN2A, to act as a test for differential diagnosis. Finally, a receiver operating characteristic (ROC) curve was drawn to evaluate the effect of this differential diagnostic test. *P* values <0.05 were considered to be statistically significant.

RESULTS

The Identification of Aberrantly Methylated DEGs in PHEO

In order to identify genes that were differentially expressed in PHEO and normal tissues, we first downloaded the gene expression profile dataset GSE19422 (84 PHEO tissues and six normal tissues) from the NCBI GEO database. Analysis of GSE19422 led to the identification of 1,935 significant DEGs (948 upregulated and 987 downregulated) for further study (**Figures 1A, B**). Methylated data were then standardized in the GSE43293 dataset to further identify 3,444 hypermethylated and 5,660 hypomethylated genes (**Figure 1C**). To identify DEGs showing aberrant methylation, all 948 upregulated genes and 5,660 hypomethylated genes were imported collectively into a Venn diagram. This led to the identification of 412 hypomethylated and highly expressed genes for further analysis (**Figure 2A**). Analysis also identified 148 hypermethylated genes with low expression levels (**Figure 2B**).

GO Enrichment Analysis of Aberrantly Methylated DEGs by DAVID 6.8

Next, we attempted to identify the biological function of the 560 aberrantly methylated DEGs. To do this, we used the DAVID 6.8 online tool to carry out GO functional enrichment analysis. As shown in **Figure 3**, the top 5 functions for BP were as follows: development of the nervous system, the positive regulation of GTPase activity, homophilic cell adhesion *via* plasma membrane adhesion molecules, axonal guidance, and signal transduction. The top 5 functions for MF were as follows: enriched in hydrolase activity, acting on carbon–nitrogen (but not peptide) bonds, Ras guanyl-nucleotide exchange factor activity, microtubule binding, transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding, and structural constituent of cytoskeleton. The top 5 locations for CC were plasma membrane, cell junction, postsynaptic membrane, postsynaptic density, and axon.

The Identification of Hub Genes by Protein–Protein Interaction Analysis Using STRING and Cytoscape 3.61

Next, we attempted to identify hub genes among the 560 aberrantly methylated DEGs. To do this, we used PPI analysis and the online STRING platform to examine protein interaction effects among aberrantly methylated DEGs. As illustrated in **Figure 4A**, the PPI network included a total of 550 nodes and 1,463 edges (PPI enrichment $P < 1.0 \times 10^{-16}$); these results were imported into Cytoscape 3.61 software for visual analysis. Using the cytoHubba plugin and the MCC method, we identified the top 10 hub genes: *CALM1*, *CACNA1C*, *KCNH2*, *KCNQ2*, *KCNMA1*, *KCNN2*, *GRIA2*, *KCNQ1*, *KCNN3*, and *SCN2A* (**Figure 2B**). The MCODE plugin of Cytoscape 3.61 was then used to analyze the whole network; this identified 13 sub-networks (**Figure 2C**). Of these, module 1 achieved the highest score (score: 6.667), while module 2 featured the most hub genes (five in total: *KCNH2*, *KCNMA1*, *KCNN2*, *GRIA2*, *KCNQ1*, and *KCNN3*). Core module analysis indicated that hub genes may play roles in pathways related to cancer, such as the phospholipase D signaling, cAMP signaling, IL-17 signaling, Toll-like receptor signaling, TNF signaling, and MAPK signaling (**Figure 5**). Consequently, these 10 candidate hub genes were selected for further analysis.

Expression Levels of Candidate Hub Genes in TCGA

The TCGA database was used to further verify our selection of key hub genes. Analysis showed that the 10 hub genes in PHEO

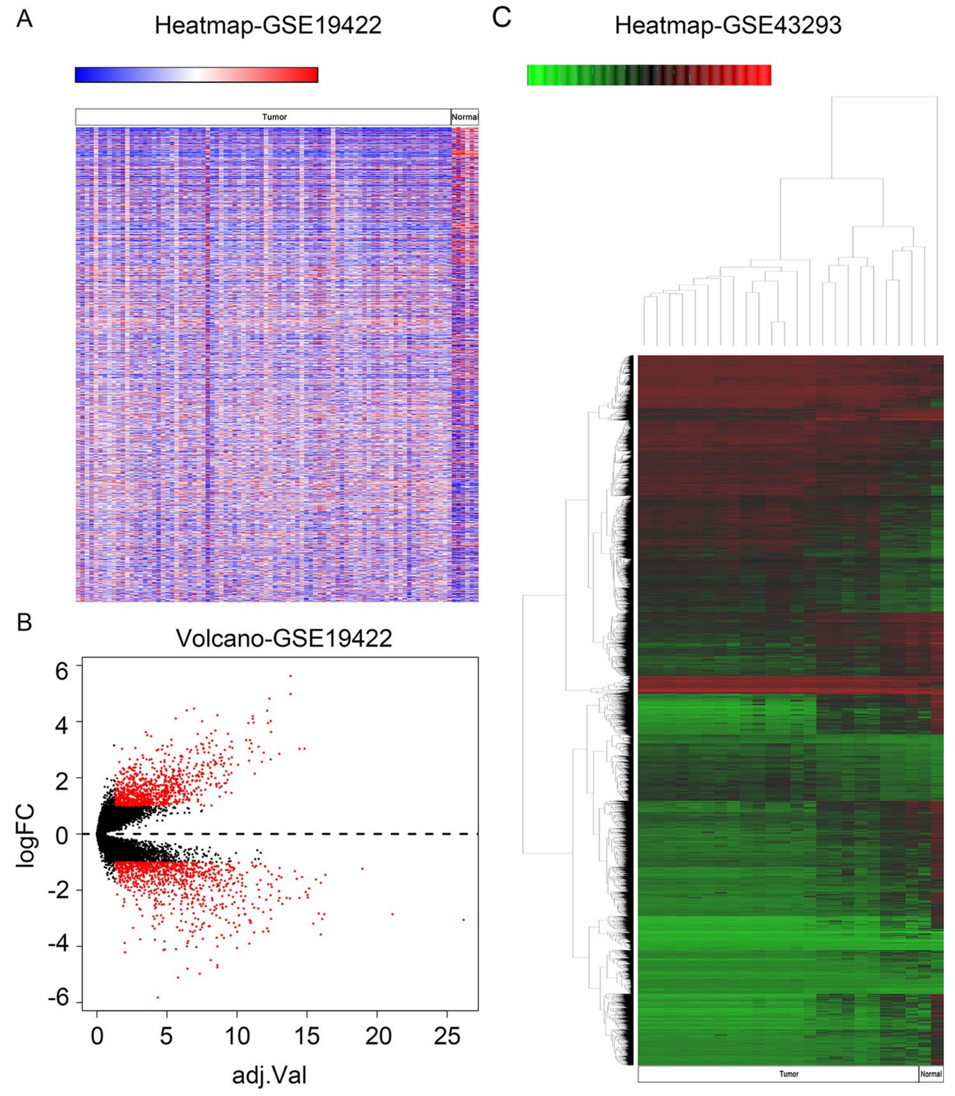


FIGURE 1 | Identification of differentially expressed genes (DEGs) and differentially methylated genes. **(A)** Heat map of DEGs in GSE19422. *Red*, upregulated genes; *blue*, downregulated genes. **(B)** Volcano plot of DEGs in GSE19422 (*red dots*). **(C)** Heat map of DEGs in GSE43293. *Red*, hypermethylated genes; *green*, hypomethylated genes.

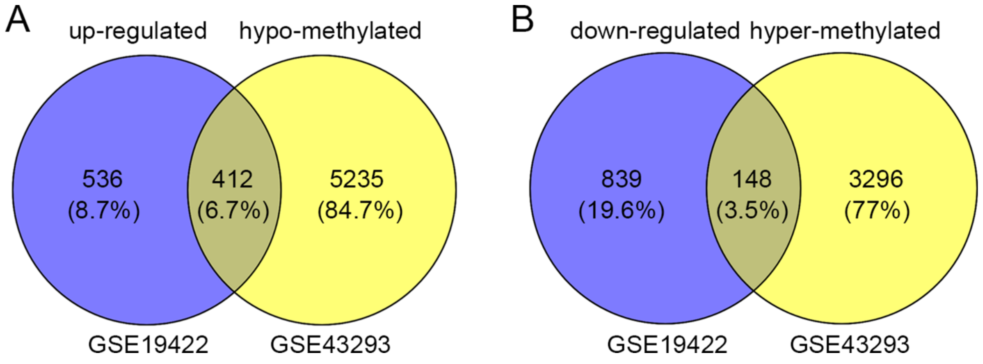


FIGURE 2 | Identification of aberrantly methylated differentially expressed genes (DEGs). **(A)** 412 upregulated and hypomethylated genes were identified. **(B)** 148 downregulated and hypermethylated genes were identified.

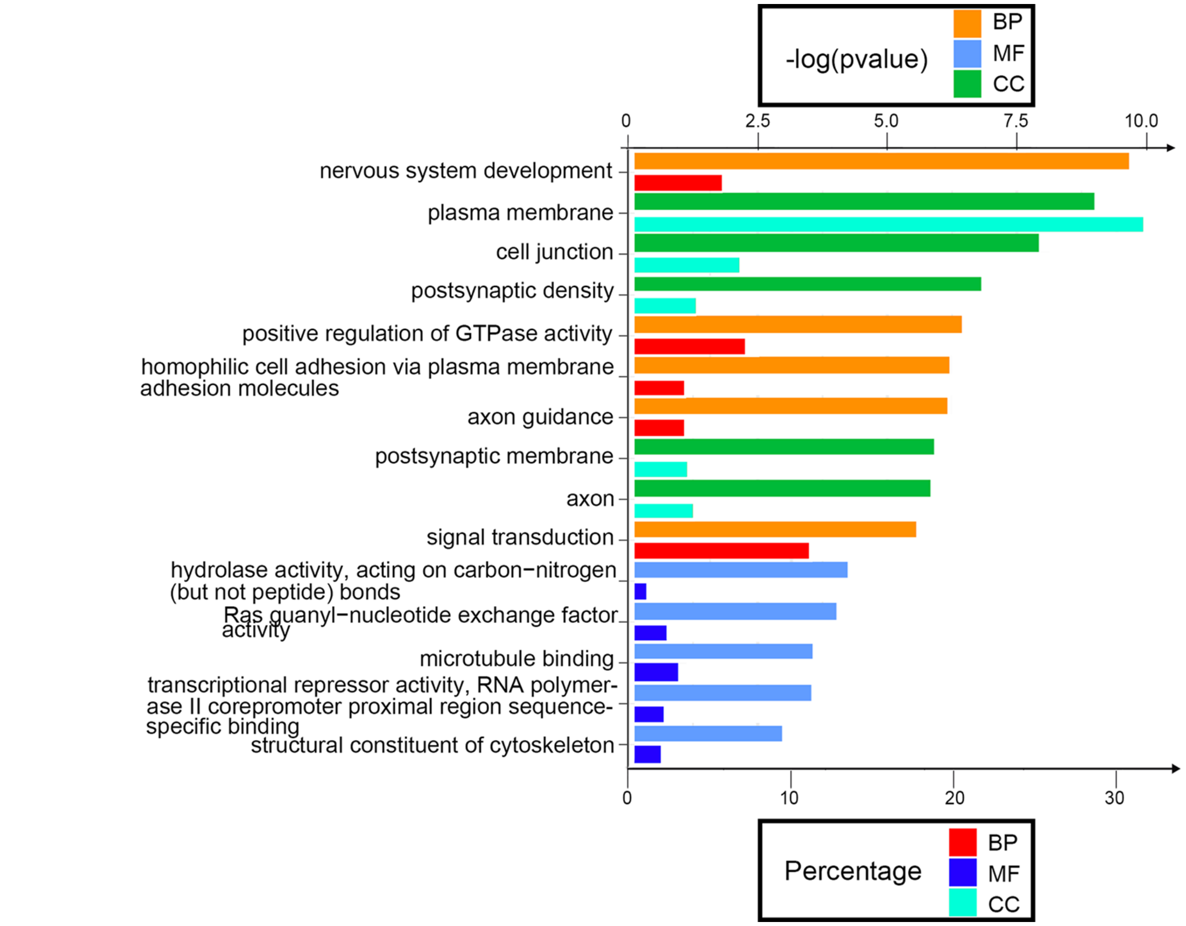


FIGURE 3 | Enrichment analysis of aberrantly methylated differentially expressed genes (DEGs). *BP,* biological processes; *MF,* molecular function; *CC,* cellular component.

tissues showed similar expression levels when compared between the TCGA and the GSE19422 dataset (**Figure 6A**) and similar methylation patterns (**Figure 6B**). As shown in **Figures 7A, B**, these hub genes showed alterations in 44.57% of the 184 cases, including mutation (3.26%) and amplification (4.89%). In addition, we found that the mRNA expression levels of the 10 hub genes showed a significant and negative relationship to the levels of DNA methylation (**Figure 7C**). Collectively, these findings indicated that DNA methylation exerts a significant effect on the progression of PHEO progression by influencing the expression of hub genes.

The Clinical Value of Candidate Hub Genes in PHEO

To evaluate the prognostic value of the candidate hub genes, we performed survival analysis using the online Kaplan–Meier plotter. **Figure 8A** shows that the overexpression of *KCNH2*, *KCNQ2*, and *KCNQ1* was significantly associated with a good prognosis; in contrast, the overexpression of *SCN2A* was significantly associated with a poor prognosis. Because of the overexpression of *KCNH2* and *KCNQ2* in PHEO when compared with normal tissues, we eliminated these genes in our subsequent analysis (**Figure 8A**). Immunohistochemical staining results from the Human Protein Atlas database indicated that *KCNQ1* showed strong expression levels in normal adrenal tissue (**Figure 8B**); in contrast, *SCN2A* was only expressed in very small levels in normal adrenal tissue (**Figure 8C**). Therefore, we investigated *KCNQ1* and *SCN2A* further to identify their potential therapeutic value. As depicted in **Figure 9A**, the expression levels of *KCNQ1* in PHEO tissues were negatively associated (Spearman's $r = -0.46$, $P < 0.0001$, and line regression coefficient = −0.4018, $P < 0.0001$) with the expression levels of *SCN2A*, suggesting that patients with PHEO may benefit from interventions targeting one of them. To this end, we established a relationship network (**Figure 9B**), including *KCNQ1* and *SCN2A*, as well as the 50 most frequently altered neighboring genes. Furthermore, some cancer drugs targeted to *KCNQ1* and *SCN2A* were included in the network, some of which are known to exhibit anti-PHEO effects, such as Propofol (Wang et al., 2018) and lidocaine (Tan et al., 2016).

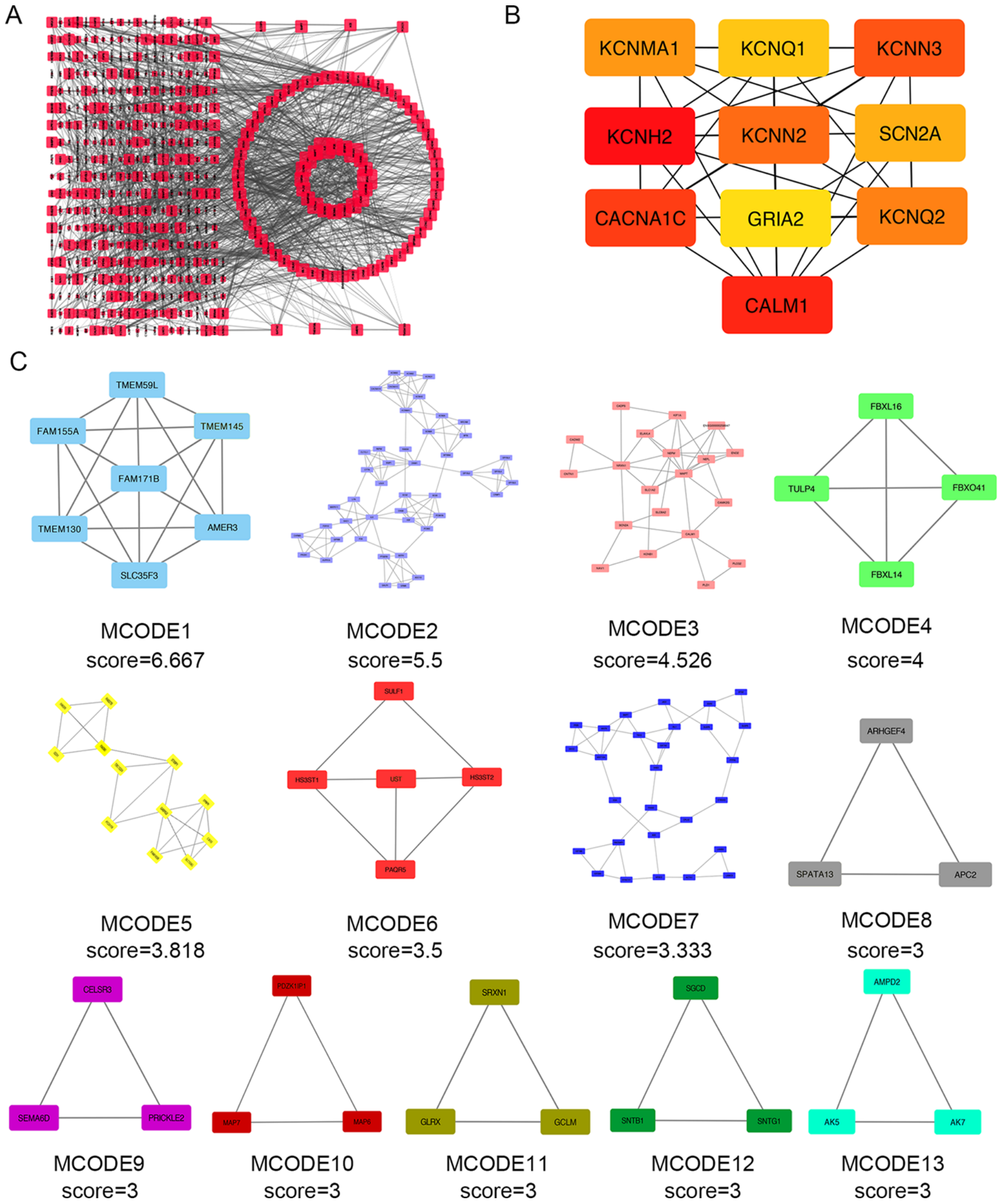


FIGURE 4 | Protein–protein interaction (PPI) network analysis and the identification of hub genes for the aberrantly methylated genes. **(A)** The PPI network included a total of 550 nodes and 1,463 edges. **(B)** The top 10 hub genes were evaluated using the maximal clique centrality method. **(C)** Module analysis for aberrantly methylated DEGs.

Due to the dilemma posed by the differential diagnosis of benign and malignant PHEOs, we performed logistic regression analysis. We attempted to improve efficiency of differential diagnosis by analyzing two gene expression datasets: GSE39716 (36 benign and nine malignant profiles) and GSE67066 (40 benign and 11 malignant profiles). Two variables, the expression levels of *KCNQ1* and *SCN2A*, were entered into backwards stepwise logistic regression analysis (**Table 1**). *KCNQ1* from the GSE39716 dataset showed the largest relative risk (RR) (50.562, *P* = 0.028), followed by SCN2A from the GSE67066 dataset (4.424, *P* = 0.009).

Then, we created a ROC curve to evaluate the value of this procedure for differential diagnosis. The area under the ROC curves for GSE39716 (**Figure 9C**) and GSE67066 (**Figure 9D**) were 0.756 [*P* = 0.019, 95% confidence interval (CI) = 0.606–0.906] and 0.786 (*P* = 0.004, 95% CI = 0.619–0.954), respectively. Corresponding sensitivity and specificity were 0.667 and 0.778, and 0.818 and 0.7, respectively, indicating

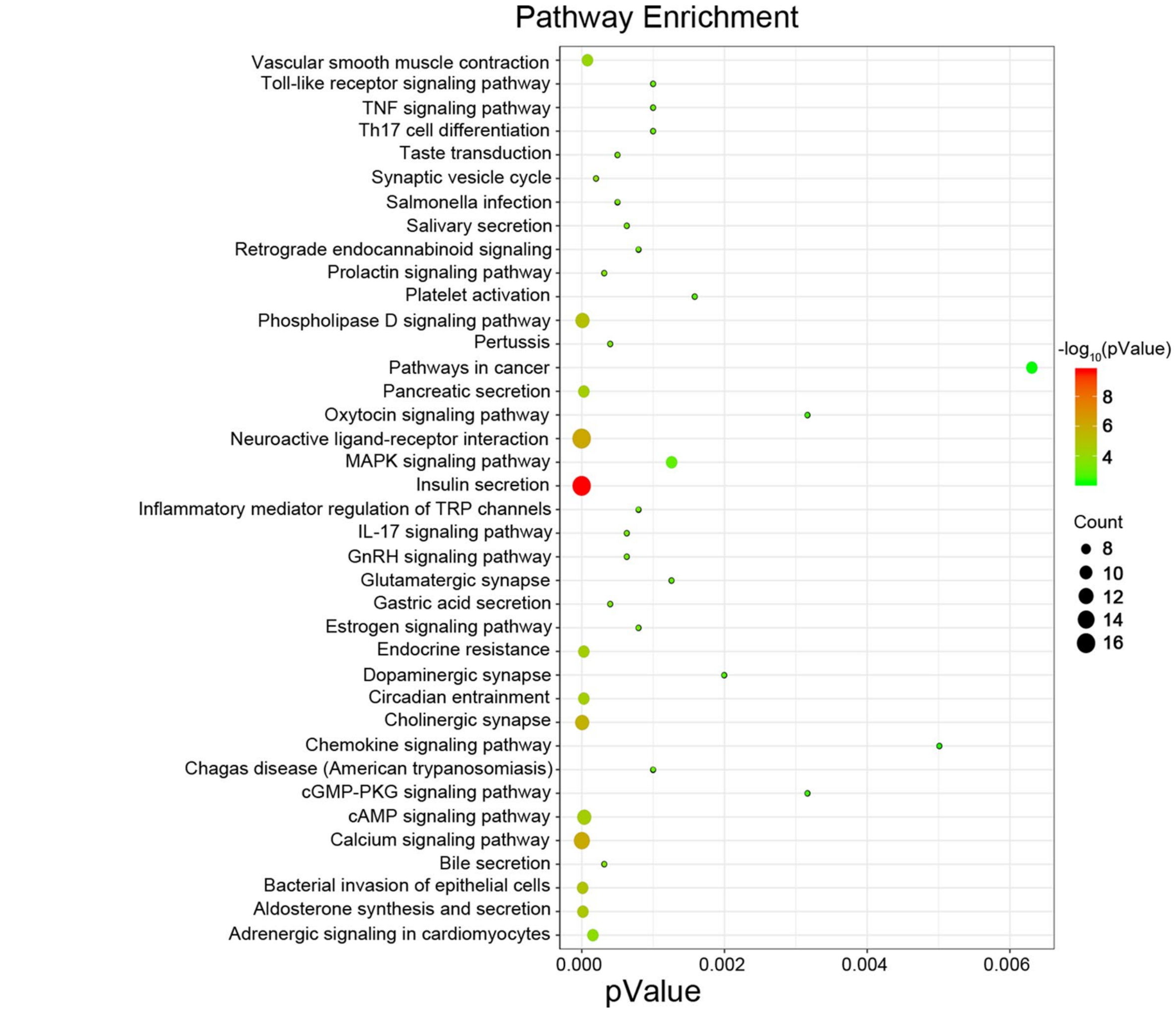


FIGURE 5 | Pathway enrichment analysis for the genes from a core module MCODE 2.

that *KCNQ1* and *SCN2A* may represent promising differential diagnostic biomarkers.

Finally, we found that there was significant difference between potassium concentration (3.98 ± 0.29 mmol/l vs. 3.63 ± 0.33 mmol/l, $P < 0.0001$) and sodium concentration (140.36 ± 2.26 mmol/l vs. 137.90 ± 3.66 mmol/l, $P < 0.0001$) when compared between pre-surgery and post-surgery day 1 (**Figures 10A**, **B**). However, the concentrations of potassium and sodium prior to surgery were not associated with tumor size (**Figures 10C**, **D**).

DISCUSSION

Despite significant effort, there is still little we can do to improve the prognosis of patients with PHEO, particularly in malignant cases. Consequently, there is a clear need to explore the specific pathogenesis of this disease and identify core genes or proteins that may facilitate clinical diagnosis and treatment. As a free and commonly used resource, the NCBI GEO database features a significant body of microarray profiling and next-generation sequencing for a variety of human tumors. Using this database, we downloaded gene expression microarray data (GSE19422) and gene methylation microarray data (GSE43293) for further analysis. In particular, we screened two of the most important hub genes, a downregulated hypermethylated gene (*KCNQ1*) and an upregulated hypomethylated gene (*SCN2A*) in PHEO tissues, both of which were further validated by the TCGA database. Functional enrichment results indicated that these hub genes played a role in the pathogenesis and progression of PHEO through certain pathways. We aimed to provide a new perspective for the pathogenesis, diagnosis, and treatment of PHEO, thus leading to improved patient outcomes.

Using R software, we identified a total of 560 aberrantly methylated DEGs. GO enrichment analysis further indicated that aberrantly methylated DEGs were predominantly involved in cancer-related biological processes, such as the positive regulation of GTPase activity, homophilic cell adhesion *via*

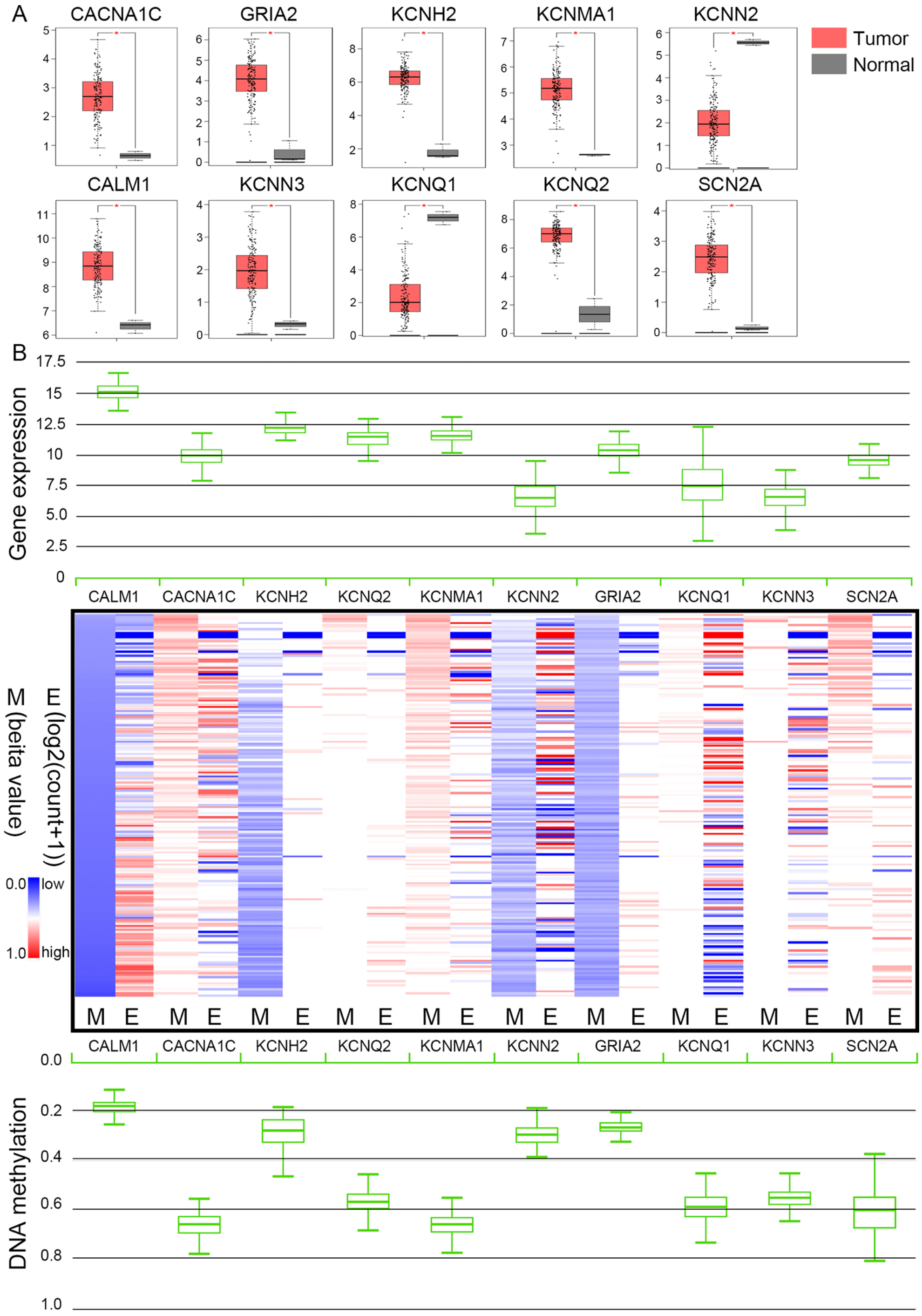


FIGURE 6 | Expression and methylation of 10 hub genes in The Cancer Genome Atlas (TCGA) database. **(A)** Box plots showing 10 hub genes at the mRNA expression level using data from the TCGA database and the GEPIA tool. **(B)** Heat map showing the correlation between the mRNA expression and DNA methylation of the 10 hub genes with the UCSC Xena platform. *M* DNA methylation, *E* mRNA expression. *Red*, upregulated genes in E or hypermethylated genes in M; *blue*, downregulated genes in E or hypomethylated genes in M.

plasma membrane adhesion molecules, axonal guidance, and signal transduction. By cycling between an inactive GDP-bound and an active GTP-bound state, the family of GTPases can act as molecular switches and are involved in a range of cellular processes, including cell proliferation, apoptosis, and migration (Olson et al., 1995; Vega and Ridley, 2008; Cherfils and Zeghouf, 2013; Croise et al., 2014; Mack and Georgiou, 2014). The relative effects of GTPases depend on whether they

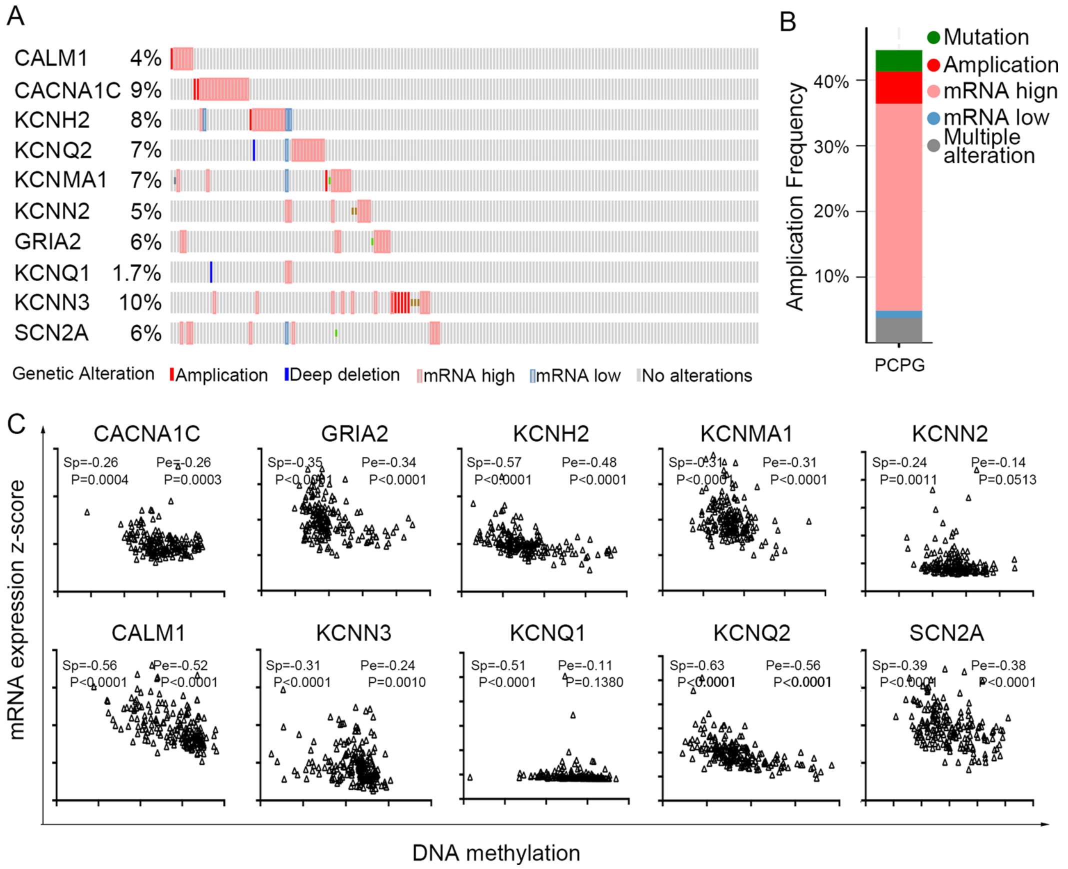


FIGURE 7 | Use of the TCGA database to validate the 10 hub genes. **(A)** Genetic alterations in the 10 hub genes. **(B)** An overview of genetic alteration in the 10 hub genes. **(C)** Correlation between mRNA expression and DNA methylation for the 10 hub genes. *Sp* Spearman correlation analysis, *Pe*: Pearson's correlation analysis.

exert action during the initiation or progression of tumors (Ellenbroek and Collard, 2007; Orgaz et al., 2014). The most commonly investigated members of the family of GTPases are RhoA, Cdc42, and Rac1. In a previous study, Croise et al. (2016) reported that PHEO was associated with the reduced activity of Cdc42 and Rac1, and the reduced expression of two Rho-GEFs, FARP1 and ARHGEF1. Our own previous research demonstrated that it inhibited PHEO progression that promotes adhesion molecules, E-cadherin and β-catenin, translocation from cytoplasm to membrane (Lin et al., 2019).

Visualization of the PPI network using Cytoscape software identified a total of 550 nodes and 1,463 edges, thus indicating that almost all of the aberrantly methylated DEGs interacted with each other, either directly or indirectly. These data imply that by manipulating the expression of core genes, it may be possible to interfere with the initiation and progression of PHEO. To this end, we used the cytoHubba plugin to identify the top 10 key genes: *CALM1*, *CACNA1C*, *KCNH2*, *KCNQ2*, *KCNMA1*, *KCNN2*, *GRIA2*, *KCNQ1*, *KCNN3*, and *SCN2A*. Similar to the GEO database, there were similar patterns of expression and methylation for these 10 core genes in PHEO when compared with normal tissues in the TCGA database, such as the downregulated and hypermethylated genes *KCNN2* and *KCNQ1*. In total, 44.57% of the 184 PHEO tissues showed genetic alterations in *KCNN2* and *KCNQ1*. These results demonstrated that these 10 core genes may play important roles in the initiation and progression of PHEO. However, only two core genes, *KCNQ1* and *SCN2A*, showed any potential prognostic value when we considered their expression patterns in PHEO.

The *KCNQ1* gene is located on chromosome 11 and has 16 exons and 15 introns. This gene encodes for the pore-forming α-subunit of a voltage-gated potassium channel that allows potassium to flow out of the cell membrane following depolarization. Under physiological conditions, this process maintains homeostasis with regards to ion concentration, cell volume, and pH (Felipe et al., 2006; Huang and Jan, 2014). An increasing body of evidence now supports the essential role of potassium channels in the initiation and progression of tumors, particularly in colorectal cancer (den Uil et al., 2016; Rapetti-Mauss et al., 2017), hepatocellular carcinoma (Fan et al., 2018), and gastric cancer (Liu et al., 2015). Research carried out by Rapetti-Mauss et al. indicated that *KCNQ1* is a target gene for the Wnt/β-catenin pathway and that the loss of *KCNQ1* promoted the disruption of cell–cell contact, thus contributing to EMT (epithelial–mesenchymal transition), cell proliferation, and invasion in colorectal cancer (Rapetti-Mauss et al., 2017). In a previous study, we demonstrated that ApoG2,

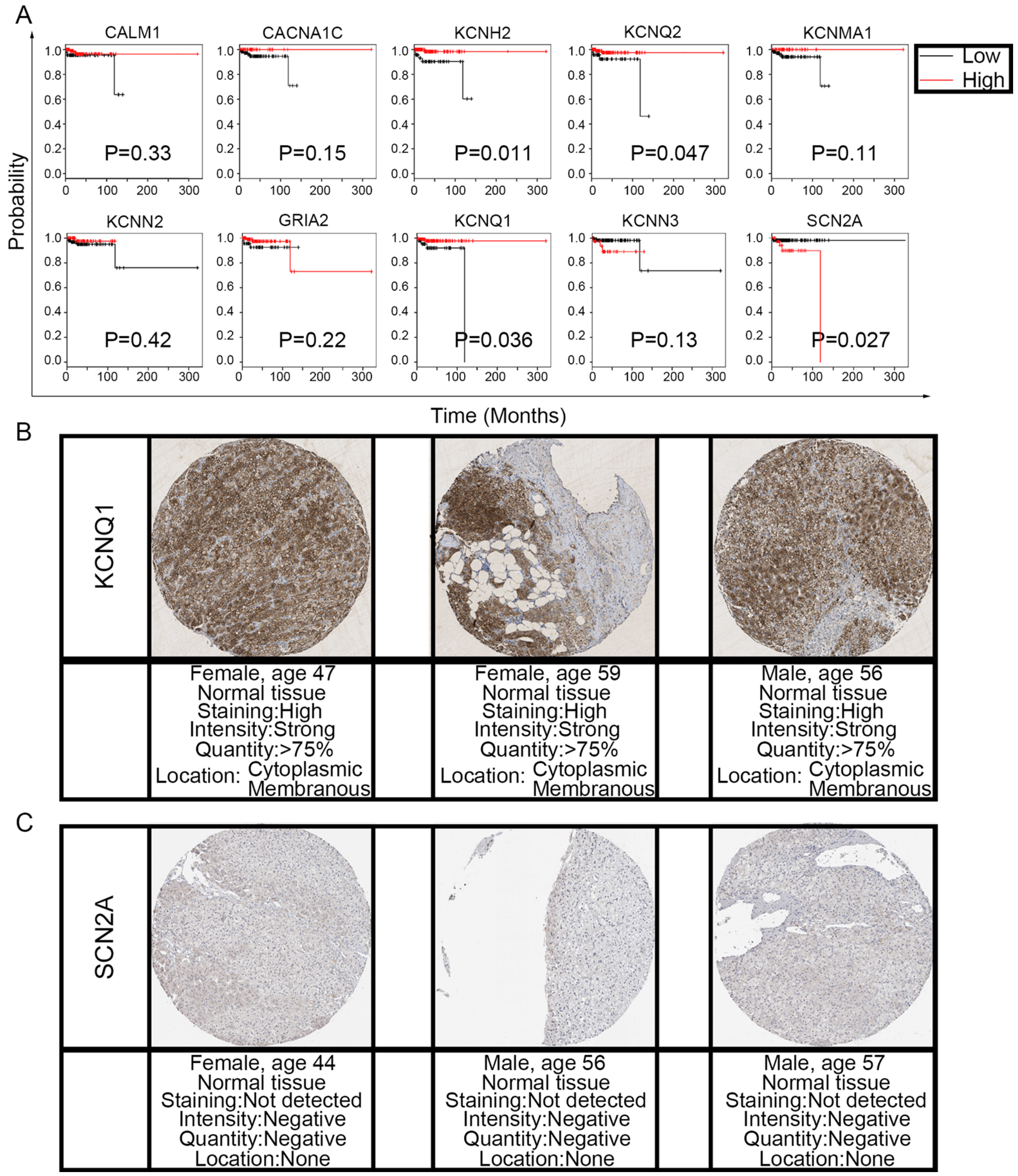


FIGURE 8 | Prognostic value of the 10 hub genes and the expression levels of *KCNQ1* and *SCN2A* in normal tissues. **(A)** Correlation between gene expression and prognosis. **(B)** *KCNQ1* showed strong expression in normal adrenal tissue. **(C)** Normal adrenal tissue was negative for *SCN2A* immunostaining.

a small molecular inhibitor, could inhibit PHEO cell migration and invasion by promoting the translocation of E-cadherin and β-catenin from the cytoplasm to the membrane dependent and that this process depended on downregulation of the PI3K/AKT pathway. This suggested that the regulation of β-catenin by *KCNQ1* may play a similar role in the metastasis of PHEO (Lin et al., 2019). Although the rate of *KCNQ1* mutation was only 1.7% (3/179), the expression level of *KCNQ1* was closely associated with the prognosis of patients with PHEO. Based on our current findings, we speculate that the methylation rate of *KCNQ1* might be more relevant than the rate of DNA mutation; this requires verification by further research.

In addition, we hypothesize that *KCNQ1*, as a potassium channel gene, could also influence the levels of potassium. As expected, analysis of our clinical data showed a significant difference for potassium concentration when compared between the pre-surgical state and post-surgery day 1. However, it remains unknown as to whether the concentration of potassium could serve as a prognostic biomarker or not. This is because the levels of potassium can be influenced by a range of factors, including, but not limited to, the progression of PHEO. Furthermore, it is not clear whether cutoff points for potassium concentration would be instructive in clinical practice. These points need to be addressed in future research.

Voltage-gated sodium channels are transmembrane glycoprotein complexes composed of a large α-subunit with 24 transmembrane domains and one or more regulatory β-subunits. The *SCN2A* gene is located on chromosome 2 and has 31 exons

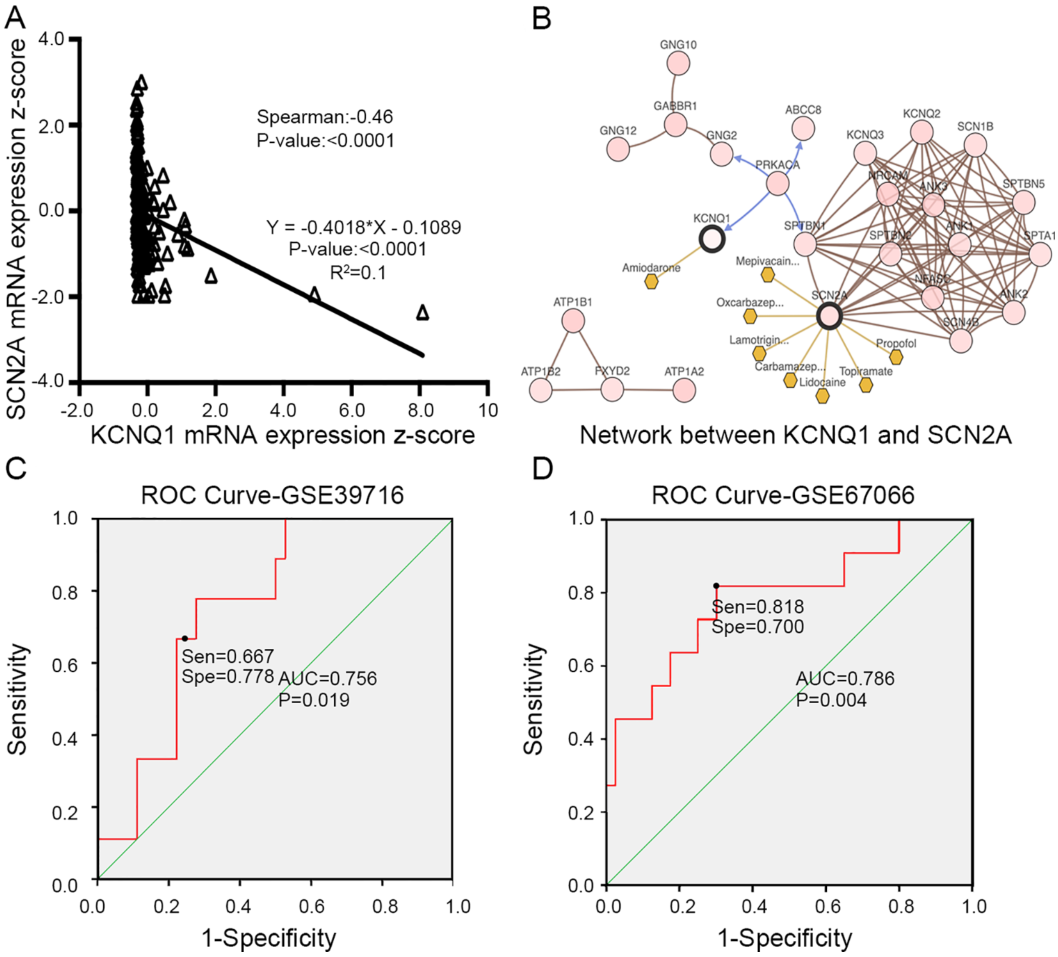


FIGURE 9 | (A) Correlation between *KCNQ1* and *SCN2A* expression levels by Spearman's correlation and line regression analysis. **(B)** Interaction network between *KCNQ1* and *SCN2A*, along with other cancer drugs targeted to *KCNQ1* and *SCN2A*. **(C)** ROC analysis of GSE39716 to discriminate between benign and malignant PHEOs, showing the sensitivity and specificity of this test. **(D)** ROC analysis of GSE67066 to discriminate between benign and malignant PHEOs, showing the sensitivity and specificity of this test. *Sen* sensitivity, *Spe* specificity, *AUC* area under the curve, *ROC* receiver operating characteristic.

TABLE 1 | Logistic regression analysis.

Dataset	Genes	Level (Ben vs. Mal)	B	P value	RR (95%CI)
GSE39716	*KCNQ1*	7.12 ± 0.46 vs. 6.82 ± 0.22	3.923	0.028	50.562 (1.542–1658.4)
	SCN2A	8.18 ± 0.82 vs. 8.33 ± 0.48	0.434	0.556	1.544 (0.364–6.554)
GSE67066	*KCNQ1*	7.39 ± 0.83 vs. 7.26 ± 0.68	0.95	0.097	2.586 (0.841–7.957)
	SCN2A	7.49 ± 0.84 vs. 6.50 ± 1.13	1.487	0.009	4.424 (1.444–13.556)

Ben, benign PHEO; Mal, malignant PHEO; B, regression coefficient; RR, relative risk; CI, confidence interval.

that encode one member of the sodium channel α-subunit gene family. Several previous publications have reported an association between *SCN2A* gene mutation and a variety of seizure types (Liu et al., 2015). However, mutation of the gene has not been associated with pathogenesis of tumors, including PHEO. However, analysis of our clinical data revealed a significant difference for sodium concentration when compared between the pre-surgical state and post-surgery day 1, thus suggesting that the pre-surgical concentration was influenced by the tumor. Furthermore, the rate of mutation in the *SCN2A* gene was as high as 6% (10/179) and its expression levels were closely associated with the prognosis of patients with PHEO. Consequently, the biological role and clinical value of *SCN2A* in PHEO clearly warrant further investigation.

CONCLUSION

In summary, we used two microarray datasets (GSE19422 and GSE43293) to identify a number of important DEGs showing aberrant methylation in PHEO-related pathways. These findings may help us to develop a better understanding of how genetic alterations are involved in the initiation and progression of PHEO and identify which genes and pathways we should investigate

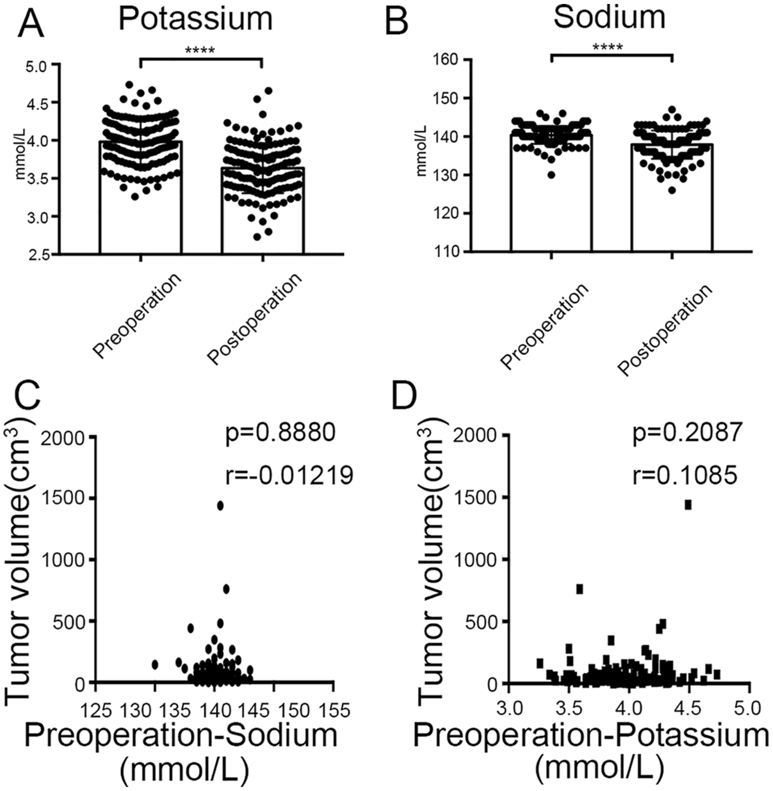


FIGURE 10 | Clinical values of *KCNQ1* and *SCN2A*. **(A)** Histogram showing a significant difference between the pre-surgical status and post-surgery day 1 for potassium concentration. ****$P < 0.0001$. **(B)** Histogram showing a significant difference between the pre-surgical status and post-surgery day 1 for sodium concentration. ****$P < 0.0001$. **(C)** Correlation between tumor volume and pre-surgical sodium concentration. **(D)** Correlation between tumor volume and pre-surgical potassium concentration.

further. Most importantly, we showed that two of the DEGs showing aberrant methylation (*KCNQ1* and *SCN2A*) represent potential biomarkers for the prognosis of patients with PHEO and may help in differential diagnosis between benign and malignant tissues. Consequently, *KCNQ1* and *SCN2A* represent valuable targets for the diagnosis and treatment of PHEO.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE19422, GSE43293, GSE39716, GSE67066.

AUTHOR CONTRIBUTIONS

DL, JL, and XL contributed equally to this work and should be considered as co-first authors. DL conceived and designed the study. DL and JL analyzed the data. DL and XL prepared the figures and wrote the text for the main manuscript. JZ and PL provided technical guidance. ZM and LZ revised the manuscript. YZ and YL provided funding support. All authors reviewed the manuscript and approved the final version for publication.

FUNDING

This study was supported by the National Natural Science Foundation of China (reference number 81272936) and the Shanghai Nature Science Foundation (no. 17ZR1417300).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01181/full#supplementary-material

SUPPLEMENTARY TABLE 1 | Biochemical level of patients.

REFERENCES

Astuti, D., Latif, F., Dallol, A., Dahia, P. L., Douglas, F., and George, E. (2001). Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am. J. Hum. Genet.* 69, 49–54. doi: 10.1086/321282

Baysal, B. E., Ferrell, R. E., Willett-Brozick, J. E., Lawrence, E. C., Myssiorek, D., and Bosch, A. (2000). Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science* 287, 848–851. doi: 10.1126/science.287.5454.848

Burnichon, N., Briere, J. J., Libe, R., Vescovo, L., Riviere, J., and Tissier, F. (2010). SDHA is a tumor suppressor gene causing paraganglioma. *Hum. Mol. Genet.* 19, 3011–3020. doi: 10.1093/hmg/ddq206

Castro-Vega, L. J., Buffet, A., De Cubas, A. A., Cascon, A., Menara, M., and Khalifa, E. (2014). Germline mutations in FH confer predisposition to malignant pheochromocytomas and paragangliomas. *Hum. Mol. Genet.* 23, 2440–2446. doi: 10.1093/hmg/ddt639

Cherfils, J., and Zeghouf, M. (2013). Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiol. Rev.* 93, 269–309. doi: 10.1152/physrev.00003.2012

Comino-Mendez, I., Gracia-Aznarez, F. J., Schiavi, F., Landa, I., Leandro-Garcia, L. J., and Leton, R. (2011). Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat. Genet.* 43, 663–667. doi: 10.1038/ng.861

Consortium, G. O. (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* 34, D322–D326. doi: 10.1093/nar/gkj021

Croise, P., Estay-Ahumada, C., Gasman, S., and Ory, S. (2014). Rho GTPases, phosphoinositides, and actin: a tripartite framework for efficient vesicular trafficking. *Small GTPases* 5, e29469. doi: 10.4161/sgtp.29469

Croise, P., Houy, S., Gand, M., Lanoix, J., Calco, V., and Toth, P. (2016). Cdc42 and Rac1 activity is reduced in human pheochromocytoma and correlates with FARP1 and ARHGEF1 expression. *Endocr. Relat. Cancer* 23, 281–293. doi: 10.1530/erc-15-0502

den Uil, S. H., Coupe, V. M., Linnekamp, J. F., van den Broek, E., Goos, J. A., and Delis-van Diemen, P. M. (2016). Loss of KCNQ1 expression in stage II and stage III colon cancer is a strong prognostic factor for disease recurrence. *Br. J. Cancer* 115, 1565–1574. doi: 10.1038/bjc.2016.376

Edstrom Elder, E., Hjelm Skog, A. L., Hoog, A., and Hamberger, B. (2003). The management of benign and malignant pheochromocytoma and abdominal paraganglioma. *Eur. J. Surg. Oncol.* 29, 278–283. doi: 10.1053/ejso.2002.1413

Ellenbroek, S. I., and Collard, J. G. (2007). Rho GTPases: functions and association with cancer. *Clin. Exp. Metastasis* 24, 657–672. doi: 10.1007/s10585-007-9119-1

Else, T., Greenberg, S., and Fishbein, L. (1993). *Hereditary Paraganglioma-Pheochromocytoma Syndromes. In: GeneReviews ((R)).* Seattle (WA): University of Washington, Seattle.

Fan, H., Zhang, M., and Liu, W. (2018). Hypermethylated KCNQ1 acts as a tumor suppressor in hepatocellular carcinoma. *Biochem. Biophys. Res. Commun.* 503, 3100–3107. doi: 10.1016/j.bbrc.2018.08.099

Felipe, A., Vicente, R., Villalonga, N., Roura-Ferrer, M., Martinez-Marmol, R., and Sole, L. (2006). Potassium channels: new targets in cancer therapy. *Cancer Detect Prev.* 30, 375–385. doi: 10.1016/j.cdp.2006.06.002

Fishbein, L., and Nathanson, K. L. (2012). Pheochromocytoma and paraganglioma: understanding the complexities of the genetic background. *Cancer Genet.* 205, 1–11. doi: 10.1016/j.cancergen.2012.01.009

Haffner, M. C., Esopi, D. M., Chaux, A., Gurel, M., Ghosh, S., and Vaghasia, A. M. (2017). AIM1 is an actin-binding protein that suppresses cell migration and micrometastatic dissemination. *Nat. Commun.* 8, 142. doi: 10.1038/s41467-017-00084-8

Hao, H. X., Khalimonchuk, O., Schraders, M., Dephoure, N., Bayley, J. P., and Kunst, H. (2009). SDH5, a gene required for flavination of succinate dehydrogenase, is mutated in paraganglioma. *Science* 325, 1139–1142. doi: 10.1126/science.1175689

Huang, X., and Jan, L. Y. (2014). Targeting potassium channels in cancer. *J. Cell Biol.* 206, 151–162. doi: 10.1083/jcb.201404136

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jain, A., Baracco, R., and Kapur, G. (2019). Pheochromocytoma and paraganglioma-an update on diagnosis, evaluation, and management. *Pediatr. Nephrol.* doi: 10.1007/s00467-018-4181-2

Karagiannis, A., Mikhailidis, D. P., Athyros, V. G., and Harsoulis, F. (2007). Pheochromocytoma: an update on genetics and management. *Endocr. Relat. Cancer* 14, 935–956. doi: 10.1677/erc-07-0142

Kopf, D., Goretzki, P. E., and Lehnert, H. (2001). Clinical management of malignant adrenal tumors. *J. Cancer Res. Clin. Oncol.* 127, 143–155. doi: 10.1007/s004320000170

Latif, F., Tory, K., Gnarra, J., Yao, M., Duh, F. M., and Orcutt, M. L. (1993). Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* 260, 1317–1320. doi: 10.1126/science.8493574

Lefebvre, M., and Foulkes, W. D. (2014). Pheochromocytoma and paraganglioma syndromes: genetics and management update. *Curr. Oncol.* 21, e8–e17. doi: 10.3747/co.21.1579

Lenders, J. W. M., and Eisenhofer, G. (2017). Update on modern management of pheochromocytoma and paraganglioma. *Endocrinol. Metab. (Seoul)* 32, 152–161. doi: 10.3803/EnM.2017.32.2.152

Lenders, J. W., Duh, Q. Y., Eisenhofer, G., Gimenez-Roqueplo, A. P., Grebe, S. K., and Murad, M. H. (2014). Pheochromocytoma and paraganglioma: an endocrine society clinical practice guideline. *J. Clin. Endocrinol. Metab.* 99, 1915–1942. doi: 10.1210/jc.2014-1498

Lin, D., Wang, X., Li, X., Meng, L., Xu, F., and Xu, Y. (2019). Apogossypolone acts as a metastasis inhibitor *via* up-regulation of E-cadherin dependent on the GSK-3/AKT complex. *Am. J. Transl. Res.* 11, 218–232.

Liu, X., Chen, Z., Zhao, X., Huang, M., Wang, C., and Peng, W. (2015). Effects of IGF2BP2, KCNQ1 and GCKR polymorphisms on clinical outcome in metastatic gastric cancer treated with EOF regimen. *Pharmacogenomics* 16, 959–970. doi: 10.2217/pgs.15.49

Mack, N. A., and Georgiou, M. (2014). The interdependence of the Rho GTPases and apicobasal cell polarity. *Small GTPases* 5, 10. doi: 10.4161/21541248.2014.973768

Mulligan, L. M., Kwok, J. B., Healey, C. S., Elsdon, M. J., Eng, C., and Gardner, E. (1993). Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. *Nature* 363, 458–460. doi: 10.1038/363458a0

Niemann, S., and Muller, U. (2000). Mutations in SDHC cause autosomal dominant paraganglioma, type 3. *Nat. Genet.* 26, 268–270. doi: 10.1038/81551

Olson, M. F., Ashworth, A., and Hall, A. (1995). An essential role for Rho, Rac, and Cdc42 GTPases in cell cycle progression through G1. *Science* 269, 1270–1272. doi: 10.1126/science.7652575

Orgaz, J. L., Herraiz, C., and Sanz-Moreno, V. (2014). Rho GTPases modulate malignant transformation of tumor cells. *Small GTPases* 5, e29019. doi: 10.4161/sgtp.29019

Pacak, K., Eisenhofer, G., Ahlman, H., Bornstein, S. R., Gimenez-Roqueplo, A. P., and Grossman, A. B. (2007). Pheochromocytoma: recommendations for clinical practice from the First International Symposium. *Nat. Clin. Pract. Endocrinol. Metab.* 3, 92–102. doi: 10.1038/ncpendmet0396 October 2005.

Park, H. U., Suy, S., Danner, M., Dailey, V., Zhang, Y., and Li, H. (2009). AMP-activated protein kinase promotes human prostate cancer cell growth and survival. *Mol. Cancer Ther.* 8, 733–741. doi: 10.1158/1535-7163.Mct-08-0631

Prejbisz, A., Lenders, J. W., Eisenhofer, G., and Januszewicz, A. (2013). Mortality associated with phaeochromocytoma. *Horm. Metab. Res.* 45, 154–158. doi: 10.1055/s-0032-1331217

Qin, Y., Yao, L., King, E. E., Buddavarapu, K., Lenci, R. E., and Chocron, E. S. (2010). Germline mutations in TMEM127 confer susceptibility to pheochromocytoma. *Nat. Genet.* 42, 229–233. doi: 10.1038/ng.533

Rapetti-Mauss, R., Bustos, V., Thomas, W., McBryan, J., Harvey, H., and Lajczak, N. (2017). Bidirectional KCNQ1:beta-catenin interaction drives colorectal cancer cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 114, 4159–4164. doi: 10.1073/pnas.1702913114

Schurmeyer, T., Dralle, H., Schuppert, F., and von zur Muhlen, A. (1988). [Preoperative diagnosis of suspected pheochromocytoma–retrospective assessment of diagnostic criteria]. *Acta Med. Austriaca* 15, 106–108.

Tan, Y., Wang, Q., Zhao, B., She, Y., and Bi, X. (2016). GNB2 is a mediator of lidocaine-induced apoptosis in rat pheochromocytoma PC12 cells. *Neurotoxicology* 54, 53–64. doi: 10.1016/j.neuro.2016.03.015

van Hulsteijn, L. T., Niemeijer, N. D., Dekkers, O. M., and Corssmit, E. P. (2014). (131)I-MIBG therapy for malignant paraganglioma and phaeochromocytoma: systematic review and meta-analysis. *Clin. Endocrinol. (Oxf)* 80, 487–501. doi: 10.1111/cen.12341

Vega, F. M., and Ridley, A. J. (2008). Rho GTPases in cancer cell biology. *FEBS Lett.* 582, 2093–2101. doi: 10.1016/j.febslet.2008.04.039

Vogel, J., Atanacio, A. S., Prodanov, T., Turkbey, B. I., Adams, K., and Martucci, V. (2014). External beam radiation therapy in treatment of malignant pheochromocytoma and paraganglioma. *Front. Oncol.* 4, 166. doi: 10.3389/fonc.2014.00166

Wallace, M. R., Marchuk, D. A., Andersen, L. B., Letcher, R., Odeh, H. M., and Saulino, A. M. (1990). Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science* 249, 181–186. doi: 10.1126/science.2134734

Walther, M. M., Keiser, H. R., and Linehan, W. M. (1999). Pheochromocytoma: evaluation, diagnosis, and treatment. *World J. Urol.* 17, 35–39.doi: 10.1007/s003450050102

Wang, H., Zhang, S., Zhang, A., and Yan, C. (2018). Propofol prevents the progression of malignant pheochromocytoma *in vitro* and *in vivo*. *DNA Cell Biol.* 37, 308–315. doi: 10.1089/dna.2017.3972

Zelinka, T., Petrak, O., Turkova, H., Holaj, R., Strauch, B., and Krsek, M. (2012). High incidence of cardiovascular complications in pheochromocytoma. *Horm. Metab. Res.* 44, 379–384. doi: 10.1055/s-0032-1306294

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 13 February 2020
doi: 10.3389/fgene.2020.00038

Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2′-O-Methylations in RNA by RiboMethSeq

Florian Pichot[1,2,3], *Virginie Marchand*[2], *Lilia Ayadi*[1,2], *Valérie Bourguignon-Igel*[1,2], *Mark Helm*[3] and *Yuri Motorin*[1,2]*

[1] IMoPA UMR7365 CNRS-UL, BioPole Université de Lorraine, Vandœuvre-lès-Nancy, France, [2] Epitranscriptomics and RNA Sequencing (EpiRNA-Seq) Core Facility, UMS2008 IBSLor (CNRS-UL)/US40 (INSERM), Université de Lorraine, Vandœuvre-lès-Nancy, France, [3] Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg-University Mainz, Mainz, Germany

OPEN ACCESS

Edited by:
Erik Dassi,
University of Trento, Italy

Reviewed by:
Ernesto Picardi,
University of Bari Aldo Moro,
Italy
Christopher L. Holley,
Duke University,
United States

***Correspondence:**
Yuri Motorin
iouri.motorine@univ-lorraine.fr

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 21 October 2019
Accepted: 13 January 2020
Published: 13 February 2020

Citation:
Pichot F, Marchand V, Ayadi L, Bourguignon-Igel V, Helm M and Motorin Y (2020) Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2′-O-Methylations in RNA by Ribomethseq. Front. Genet. 11:38. doi: 10.3389/fgene.2020.00038

A major trend in the epitranscriptomics field over the last 5 years has been the high-throughput analysis of RNA modifications by a combination of specific chemical treatment (s), followed by library preparation and deep sequencing. Multiple protocols have been described for several important RNA modifications, such as 5-methylcytosine (m^5C), pseudouridine (ψ), 1-methyladenosine (m^1A), and 2′-O-methylation (Nm). One commonly used method is the alkaline cleavage-based RiboMethSeq protocol, where positions of reads' 5'-ends are used to distinguish nucleotides protected by ribose methylation. This method was successfully applied to detect and quantify Nm residues in various RNA species such as rRNA, tRNA, and snRNA. Such applications require adaptation of the initially published protocol(s), both at the wet bench and in the bioinformatics analysis. In this manuscript, we describe the optimization of RiboMethSeq bioinformatics at the level of initial read treatment, alignment to the reference sequence, counting the 5′- and 3′-ends, and calculation of the RiboMethSeq scores, allowing precise detection and quantification of the Nm-related signal. These improvements introduced in the original pipeline permit a more accurate detection of Nm candidates and a more precise quantification of Nm level variations. Applications of the improved RiboMethSeq treatment pipeline for different cellular RNA types are discussed.

Keywords: 2′-O-methylation, RNA, ribose methylation, high-throughput sequencing, bioinformatic pipeline, receiver operating characteristic

INTRODUCTION

The precise and high-throughput mapping of modified nucleotides in RNA is a real challenge in the field of epitranscriptomics (RNA modifications). Several recent publications have demonstrated the presence of numerous RNA modifications, not only in rather well studied species such as tRNA/rRNA/sn(sno)RNA, but also in coding RNAs (mRNA), in all living organisms studied to date. Thus, several high-throughput methods for the identification of RNA modifications have been developed and successfully applied for mapping m^5C, m^6A, pseudouridines, and (more recently) 2′-O-methylations (2′-O-Me), along with m^1A and $m^7G/m^3C/D$ (Dominissini et al., 2012; Squires et al., 2012; Carlile et al., 2014; Schwartz et al., 2014; Hauenschild et al., 2015; Dai et al., 2017; Marchand et al., 2018; Schwartz, 2018).

The major impediment to applying high-throughput screening methods resides in various experimental and bioinformatics biases, which are only partially controlled and may affect the precision of the final result or even lead to a nonnegligible number of false-positive identifications. Taking into account the size of eukaryotic transcriptomes, thousands of false-positive signals may appear at the transcriptome-wide level, even with extremely strict criteria for candidate site selection (e.g., False Discovery Rate [FDR]< 1%). Thus, every step in bioinformatic data treatment, conversion and manipulation should be optimized in order to minimize the number of potential false-positive signals.

Recently, we published a high-throughput deep sequencing-based approach, named RiboMethSeq, for mapping of 2′-O-methylations in highly abundant RNAs, mostly in rRNA (Marchand et al., 2016; Erales et al., 2017), with possible extension to tRNA (Marchand et al., 2017a; Freund et al., 2019). This protocol is also suitable for low abundance RNAs (Krogh et al., 2017). The RiboMethSeq protocol is based on the enhanced protection of the phosphodiester bond in RNA from nucleolytic attack and cleavage due to the presence of 2′-O-methylation at the 5′-neighboring ribose moiety. This enhanced protection is evaluated as a normalized number of 5′- and 3′-ends of randomly cleaved fragments present in the sequencing library. If a residue is 2′-O-methylated, this reduces the cleavage efficiency and thus the relative number of fragments starting and ending at +1 nucleotide relative to the modification. In the RiboMethSeq approach, such relative protection compared to neighbors is calculated using different scoring schemes, and the presence/absence of a 2′-O-methylation is then deduced on this basis. An alternative calculated score (ScoreC) also allows precisely measuring the methylation ratio at a given nucleotide.

In this work, we report the comprehensive optimization of every step of the bioinformatic treatment used for the detection and quantification of ribose 2′-O-methylation by the RiboMethSeq protocol. We systematically evaluated the importance and the impact of 5′- and 3′-trimming strategies, parameters for alignment to the reference sequence, as well as the use of specific calculated scores for 2′-O-Me mapping and quantification. Our results demonstrate that a reduced calculation interval is favorable for the general discrimination of 2′-O-Me signals from potential false-positive hits. We propose new, optimized scores (ScoreMEAN2, ScoreA2, and MethScore2) that provide better FDR values and also improve the relative quantification of 2′-O-methylation in RNA.

MATERIALS AND METHODS

Biological Material

To optimize the RiboMethSeq scores, we used previously published datasets obtained for wild-type yeast *Saccharomyces cerevisiae* and human HeLa cell rRNA 2′-O-methylation analysis, as well as additional samples for hTERT immortalized human mammary epithelial cell line (HME) (Marchand et al., 2016; Erales et al., 2017; Sharma et al., 2017); accession numbers PRJEB43738, PRJEB34951 and PRJEB35565.

RiboMethSeq Protocol

The RiboMethSeq protocol (Marchand et al., 2016; Marchand et al., 2017b) consists of random RNA fragmentation under alkaline conditions (96°C, pH 9.3, ~12–14 min for rRNA), an end-repair step consisting of de-phosphorylation of the 3′- ends of the RNA fragments and 5′-end phosphorylation, library preparation using 3′ -end and 5′-end ligation of adapters, an RT-step and PCR amplification coupled with barcoding. The resulting library is sequenced in paired-end PE 2x75 or, more commonly, in single-end mode (SR50) using an Illumina sequencing device (MiSeq or HiSeq1000).

Trimming and Alignment

Adapter removal in this study was performed using the Trimmomatic utility (Bolger et al., 2014). With the default trimming parameters, the recognition of adapter sequences requires at least a 16 nt length. Shorter fragments of adapters are not recognized and thus are not removed. However, with a stringency of 7 (instead of 10), the adapter recognition requires only 10 nt. Considering this, 3′-end counting was carried out only for reads shorter than 40 nt after trimming. The alignment of raw reads was conducted by Bowtie2 (Langmead and Salzberg, 2012) in end-to-end mode.

Comparing the Performance of 2′-O-Methylated Site Detection

For the selected datasets, we first applied the RiboMethSeq pipeline under standard conditions, and the previously described scores (ScoreMAX6, ScoreA, B and C) were calculated (Birkedal et al., 2015; Marchand et al., 2016). Score values were sorted in descending order, and the known modification status of every nucleotide (2′-O-methylated residue, pseudouridine, other modified residue or unmodified nucleotide) was attributed. Receiver operating characteristics (ROC) curves were plotted using these data, together with associated Matthews correlation coefficient (MCC) values. Other associated parameters of the ROC curves were calculated for maximal MCC value, taking into account true positive/false positive/true negative/false negative (TP/FP/TN/FN) hits. The performance of the treatment was evaluated on the basis of both the maximal MCC value and the associated FDR.

RESULTS AND DISCUSSION

Brief Overview of RiboMethSeq Experiment

As described above, the high-throughput mapping of 2′-O-Me residues in RNA is based on random fragmentation of the phosphodiester bonds under mild alkaline conditions. The presence of a 2′-O-Me group protects the 3′-adjacent phosphodiester bond from nucleolytic cleavage, thus generating the characteristic gap in the 5′-end (as well as 3′-end) coverage profile of the sequencing library prepared from the fragmented RNA (**Supplementary Figure S1**). This enhanced protection is used as a signature for 2′-O-methylation and protection (and thus the gap's depth) and is supposed to be proportional to the level of 2′-O-Me at a given position.

In previously published studies (Birkedal et al., 2015; Marchand et al., 2016; Erales et al., 2017; Sharma et al., 2017), we and others used rather standard parameters for read trimming and alignment, and calculations arbitrarily used cleavage efficiency for 12 neighboring nucleotides (+/−6 from the methylation site). Scores allowing 2′-O-Me detection (ScoreMAX6 and ScoreA, called here ScoreA6), were calculated. The 2′-O-methylation level was assessed by calculating the MethScore (identical to the previously reported ScoreC, called ScoreC6 here).

Selection of Representative Datasets for Optimization

Initial screening and optimization of the RiboMethSeq bioinformatic pipeline was performed with >40 available human rRNA RiboMethSeq datasets obtained under standard, previously described (Marchand et al., 2016; Erales et al., 2017; Sharma et al., 2017), conditions of RNA fragmentation, sequencing, trimming, alignment and score calculation. We used a cumulative list of human modified rRNA positions reported in a 3D rRNA modification database (Piekna-Przybylska et al., 2008) and in the LBME snoRNA database (Lestrade and Weber, 2006), including two new positions that were recently reported (Krogh et al., 2016). Altogether, we considered 40 sites in 18S rRNA, 64 sites in 28S rRNA, and 2 positions in 5.8S rRNA (see **Supplementary Table S1**). Some of these positions are probably variably modified, or even not modified at all in some human cell lines or tissues (Krogh et al., 2016; Erales et al., 2017; Sharma et al., 2017); therefore, these incomplete modifications necessarily affect the number of FN hits and the max MCC values in the RiboMethSeq analysis. For each dataset, calculations of the RiboMethSeq scores were performed, and the performance of each dataset was evaluated for the detection of known rRNA 2′-O-methylated positions. Based on the preanalysis of available RiboMethSeq human rRNA samples, we selected three representative human datasets corresponding to two different cell lines (HUVEC and HeLa), as well as cultured bone marrow stem cells for further, more extensive analysis and optimization of the whole treatment pipeline (respectively named Sample 1 – HUVEC, 2 – BMSC, and 3 – HeLa (**Figure 1** and **Supplementary Figure S2**, **Supplementary Table S2**).

To allow a performance comparison between datasets, we selected samples with similar numbers of raw RiboMethSeq sequencing reads (>20 mln, see **Supplementary Table S2**).

Minimal Number of Reads Required for Analysis

An optimized volume of sequencing reads required for complete RNA analysis is highly important, since it allows to obtain reliable results with a minimal sequencing cost per sample. From the analysis of the yeast rRNA samples (Marchand et al., 2016), sufficient coverage was evaluated to be ~ 750–1,000 reads/RNA position. For a more detailed analysis, we applied human rRNA datasets. Despite a similar number of raw sequencing reads (~20 mln), these datasets behaved differently regarding the precision of 2′-O-Me detection. Notably, the fraction of 5S rRNA reads varied substantially from one sample to another (**Figure 1**), probably also reflecting the different RNA extraction protocols used. However, there was no correlation between the total coverage and the prediction quality for 2′-O-Me. To define the minimal amount of raw sequencing information required for the successful application of RiboMethSeq analysis of human rRNA, we compared the performance of the method using a variable number of input reads for the same sample.

As anticipated, human 28S rRNA was the most difficult target to get full representation for all positions in the sequence. A comparison of missing positions in 28S rRNA in relation to the sequenced population is given in **Table 1**. A low number of raw reads (4 mln) can still be used, but numerous positions of 28S rRNA have zero raw read 5′-/3′-end counts in the final set. Despite this, the analysis of known 28S rRNA 2′-O-Me was not affected, because underrepresented regions are far away from these modified positions. Increasing the read number (8–12 mln, > 1,000 reads/nt) improves representativity, with only a marginal number of uncovered nucleotides, while 15–20 mln reads is recommended to achieve full coverage.

Minimal Trimming Length

The minimal trimming length used in the treatment pipeline may affect 2′-O-Me detection. Trimming parameters considerably influence the precision of 3′-end mapping for SR50 reads and the alignment quality. We thus tested variable minimal trimming lengths keeping alignment parameters constant. The calculated max MCC values for the tested human datasets showed no influence of these parameters on the final results, even if the number of ambiguously aligned short reads increased with a decreased minimal trimming length (**Supplementary Figure S3**). Depending on the length and complexity of the target RNA sequence, we recommend adapting the minimal length and Bowtie2 seed length (see below); the optimal seems to be 10 or 12 for human or yeast rRNA, or even lower for shorter RNAs (e.g., tRNA).

Variation of Alignment Parameters

The original RiboMethSeq protocols (Marchand et al., 2016; Marchand et al., 2017b) used rather strict alignment parameters in Bowtie2: end-to-end mode, a minimal seed length of 22 nt and zero mismatches allowed in the seed (preset option "– sensitive"). The influence of the alignment mode (end-to-end versus local) was previously evaluated, and the soft read trimming performed in the local mode was found to be unsuitable for precise mapping of the read ends (Marchand et al., 2016). However, the seed length and number of mismatches allowed may also influence the quality of the alignment, since human rRNA has variations in nucleotide sequence and contains other modified nucleotides, which alter cDNA sequences, thus perturbing alignment to the reference.

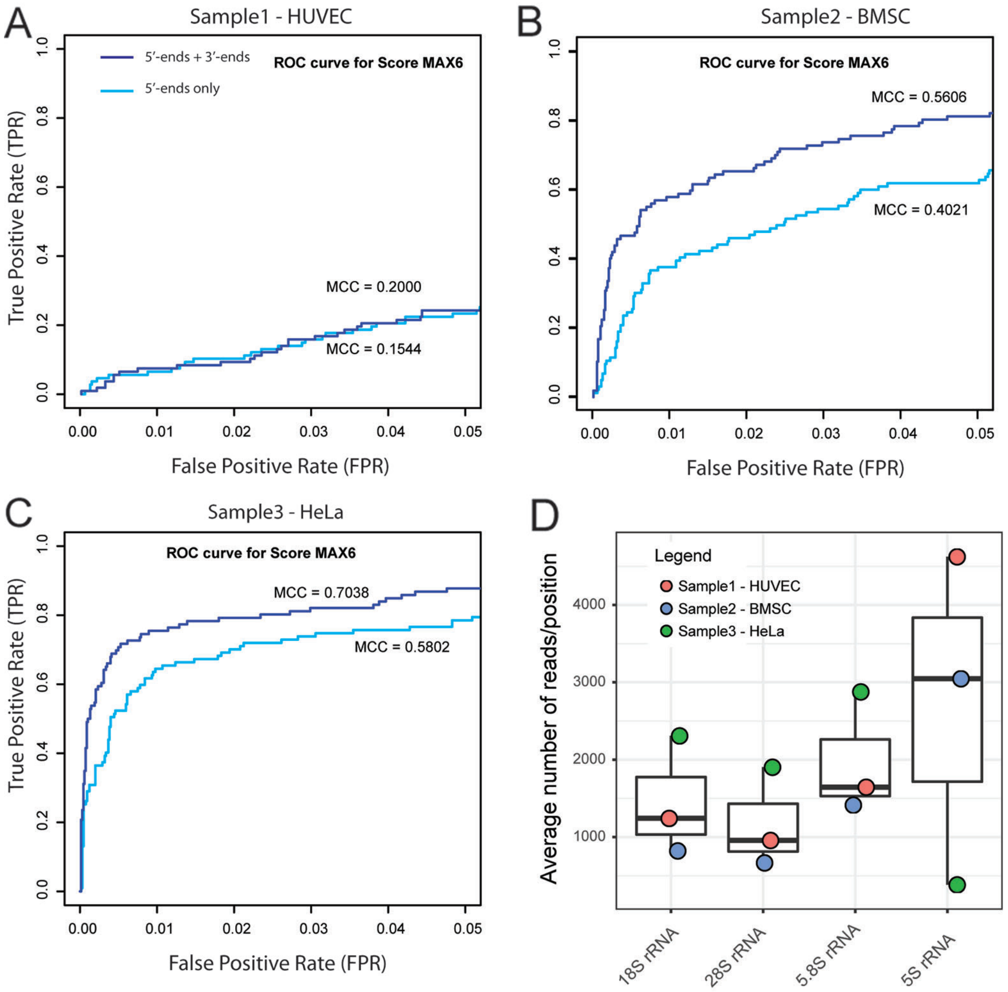


FIGURE 1 | Selection of RiboMethSeq datasets for optimization. Three human datasets providing representative performance of 2'-O-Me detection (Sample 1 – HUVEC, 2 – BMSC, and 3 – HeLa) were selected on the basis of receiver operating characteristics (ROC) curves and the associated max Matthews correlation coefficient (MCC) values for ScoreMAX6 **(A–C)**. Graphs represent zoom to ROC curve 0–0.05 for false positive rate (FPR) and 0–1 for true positive rate (TPR). It was previously shown (Marchand et al., 2016) that 5'-end coverage (light blue curve) is sufficient for reliable construction of the RNA protection profile, but cumulated 5'- and 3'-end coverage (violet curve) provides better discrimination between methylated positions and false positive (FP) hits. **(D)** shows the read coverage per position for human rRNAs. 5S rRNA shows quite variable coverage, probably due to variations in 5S rRNA content in the total rRNA fraction due to biased extraction.

To evaluate the importance of the alignment parameters, we proceeded with treatment using a reduced seed length and allowing (or not) mismatched nucleotides in the seed. The following parameter combinations were tested: seed lengths of 22 nt (default value for "– sensitive" preset option, and used previously), 16 nt, 12 nt, and 8 nt, allowing (or not) mismatched nucleotides in the seed.

The data in **Supplementary Figure S4A** show that the total proportion of aligned reads (unique or multiple alignments) vary only very slightly as a function of seed length and allowed mismatches; a seed length of 22 nt and mismatches in the seed allows only a slightly better alignment (69.41% vs. 65.41% of aligned reads). Variation of the Bowtie2 seed length does not much affect the max MCC value for both scores used for optimization (**Supplementary Figure S4B**). Similarly, allowing mismatches in the seed also has no influence on the final results. This shows that the alignment to the reference sequence is quite robust and mostly depends on the quality of sequencing data. Based on these observations, we recommend the use of 8–12 nt seed length, depending on the complexity and the length of the target RNA sequence. For better performance, the seed length can be coordinated with the minimal trimming size for sequencing reads (**Supplementary Figure S5**).

Importance of the Calculation Window With Neighboring Nucleotides

In the originally published RiboMethSeq protocols (Birkedal et al., 2015; Marchand et al., 2016), 12 neighboring nucleotides

TABLE 1 | Alignment statistics and uncovered rRNA positions in samples used for analysis.

Sample	number of raw reads used	4 mln	8 mln	12 mln	16 mln	20 mln
Sample 1 HUVEC	trimmed reads	3873996	7741257	11607806	15471803	19341955
	short reads for alignment	1761248	3581209	5338558	7104102	8909438
	aligned to rRNA reference	1513287	3077770	4587105	6104305	7657111
	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	**4**	**1**	0	0	0
	uncovered pos 28S rRNA	**100**	**37**	**21**	**11**	**7**
Sample 2 BMSC	trimmed reads	3878093	7752516	11628210	15494852	19371588
	short reads for alignment	1473330	2986750	4455805	5927702	7428697
	aligned to rRNA reference	999714	2027867	3023042	4022365	5042133
	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	**2**	**1**	0	0	0
	uncovered pos 28S rRNA	**18**	**5**	0	0	0
Sample 3 HeLa	trimmed reads	3882713	7764523	11644031	15516647	19387461
	short reads for alignment	2582027	5182220	7776132	10353556	12934621
	aligned to rRNA reference	2222928	4460722	6693085	8910528	11132036
	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	0	0	0	0	0
	uncovered pos 28S rRNA	0	0	0	0	0

The bold font highlights the number of uncovered positions in the different datasets.

(+/−6 nt window) were taken into account to calculate the 2′-O-Me scores (ScoreA, B and C). Since the size of this window was arbitrarily selected, we explored the influence of the window's size on the discrimination of 2′-O-Me signals from background. We compared the maximal MCC and the FDR values for the calculation interval from +/−2 nt up to +/−8 nt (window of 4 to 16 nt). The graph in **Figure 2A** shows that the ScoreMAX is nearly insensitive to the size of the calculation window, while ScoreA shows the best performance (and the lowest FDR) with the smallest window size (+/−2 nt). A larger window size has a detrimental effect for both scores. On the basis of these observations we suggest reducing the calculation window size to four neighboring nucleotides (Score 2 calculation scheme, +/−2 nt); the scores calculated with this window are referred to as ScoreMAX2 and ScoreA2.

Quantification of 2′-O-Methylation With MethScore (ScoreC)

In the original RiboMethSeq protocol (Marchand et al., 2016), the MethScore [identical to ScoreC (Birkedal et al., 2015; Krogh et al., 2016)] was used for quantification of the 2′-O-Me level because the MethScore demonstrates a linear dependence on the depth of the gap in a cleavage profile, which is supposed to represent the protection and thus the degree of 2′-O-Me. In an ideal situation of a homogeneous cleavage profile, the MethScore varies from 0 to 1.0 and thus can be used as a measure of the degree of RNA methylation. In the case of yeast rRNA studied previously, the MethScore varied from negative values to 1.0, while conserving the linear dependence on the methylation rate (Marchand et al., 2016).

We noticed that RiboMethSeq detection scores behave better with a reduced calculation interval (+/−2 nt); therefore, we also explored variations of the MethScore. To select the best calculation interval, we compared the MethScore for intervals from two to eight neighboring nucleotides for naturally modified yeast rRNA and for unmodified synthetic rRNA transcripts. The MethScore values were expected to reach maximum for naturally modified rRNA and minimum for the unmodified counterpart. **Figure 2B** shows the difference between the MethScore values calculated for variable intervals for rRNA and synthetic transcript. For MethScore6 (+/−6 nt), we also tested different weight contributions of neighboring nucleotides.

The MethScore demonstrates the maximal cumulative difference between rRNA and synthetic transcript for the shortest interval of +/−2 nt; the other tested intervals gave roughly the same results. On the basis of this observation, we suggest calculating the MethScore for two neighboring nucleotides (referred to as MethScore2).

A detailed analysis, position by position, for 18S and 25S rRNA (**Figures 2C, D**) shows that methylated yeast rRNA displays MethScore2 values close to 0.9–1.0 for almost all modified positions (blue dots and line for average value), while the average level for synthetic unmodified transcript (red dots and line) is rather low. However, it is notable that the MethScore2 values for unmodified transcript are extremely variable, ranging from −1 to almost 0.9. For a limited subset of sites (Gm1428 in SSU 18S-rRNA and four positions in LSU 25S-rRNA), the difference of MethScore2 between modified and unmodified RNA is as low as 0.1. Precise measuring of the 2′-O-Me level variations at these rRNA positions is thus extremely difficult. However, over 90% of methylation sites display considerable MethScore2 differences between the modified and unmodified state, thus validating relative quantification of the methylation rate. Absolute values of the 2′-O-methylation

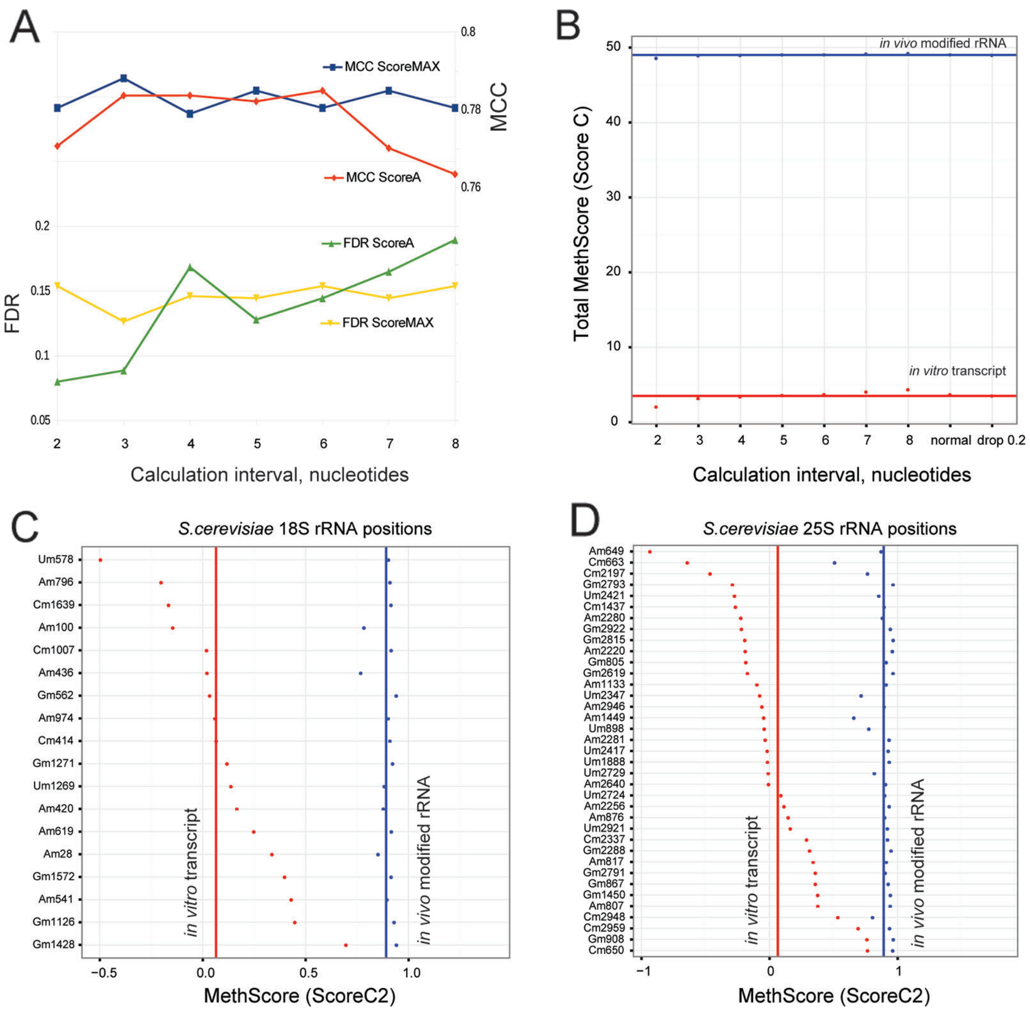


FIGURE 2 | (A) Performance of ScoreMAX and ScoreA calculated using variable numbers of neighboring nucleotides (from +/−2 to +/−8). The standard RiboMethSeq protocol uses a +/−6 interval. Values for FDR and max Matthews correlation coefficient (MCC) are given. The scale on the left corresponds to false discovery rate (FDR), and on the right to MCC. Sample 2 – BMSC was used here for all calculations; other datasets gave similar trends. **(B)** shows global values for MethScore (ScoreC) calculated for modified yeast rRNA and *in vitro* rRNA transcripts using different neighboring intervals. The total number of "2'-O-Me groups" in rRNA is given (red - *in vitro* transcript, blue - modified rRNA). **(C, D)** MethScores2 (ScoreC2) for individual 2'-O-methylated positions in 18S **(C)** and 25S rRNA [**(D)**, red - *in vitro* transcript, blue - modified rRNA]. Lines correspond to average values.

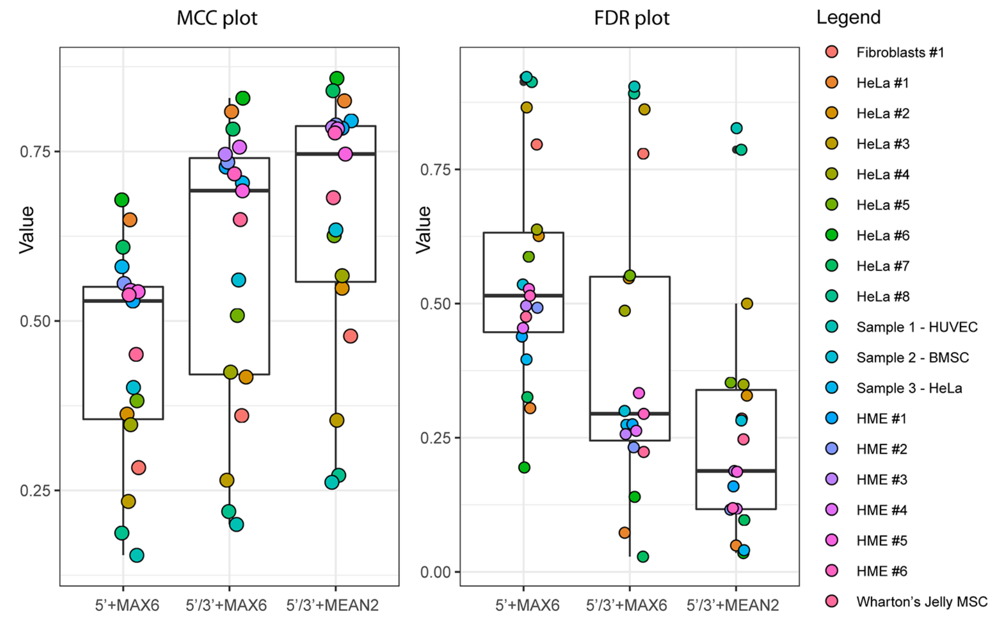
cannot be directly measured using the RiboMethSeq approach; however, for a limited subset of sites the values of MethScore2 are comparable with independent measurements of 2′-O-Me rates assessed by LC-MS/MS (Buchhaupt et al., 2014; Taoka et al., 2016).

Optimization of the ScoreMAX

The originally used ScoreMAX6 (Marchand et al., 2016) was designed to favor directional 5'- > 3′ gap depth compared to the opposite orientation. This design was based on the assumption that directional 5'- > 3′ drop is more informative for measuring the protection of the phosphodiester bond, while the drop in the opposite direction (3′- > 5') may represent nonspecific RNA structural effects. To verify the validity of such assumption, we tested different variants of ScoreMAX and their performance in the detection of 2′-O-Me. We compared the ScoreMAX2 (+/−2 nt window) with two modified versions. The first modified score retained the maximal value of normalized 5'- > 3′ and 3′- > 5' drop (ScoreMAX-MAX), while the second version calculated the average value between the two (ScoreMEAN2).

Calculation of the max MCC value demonstrated that ScoreMAX-MAX is less performant than the original ScoreMAX2, while ScoreMEAN2 shows better discrimination of FP hits. For the same dataset, Sample 3 – HeLa application of the ScoreMEAN2 allows attainment of the maximal MCC value of 0.7954 and reduces the FDR from 27% to 4%. A similar tendency was observed for two other datasets (**Supplementary Table S3**).

An important source of FP hits in RiboMethSeq analysis is the reduced ligation efficiency observed when modified RNA



Legend
● Fibroblasts #1
● HeLa #1
● HeLa #2
● HeLa #3
● HeLa #4
● HeLa #5
● HeLa #6
● HeLa #7
● HeLa #8
● Sample 1 - HUVEC
● Sample 2 - BMSC
● Sample 3 - HeLa
● HME #1
● HME #2
● HME #3
● HME #4
● HME #5
● HME #6
● Wharton's Jelly MSC#1

FIGURE 3 | Improvement of ScoreMAX/MEAN (MAX6 and MEAN2) with 5'/3'-counts and reduced calculation window (Score 2 calculation scheme). Boxplot shows max Matthews correlation coefficient (MCC) values (left) and associated false discovery rate (FDR) (right) for all 19 RiboMethSeq datasets used for validation. Identity of the RiboMethSeq datasets is given on the right.

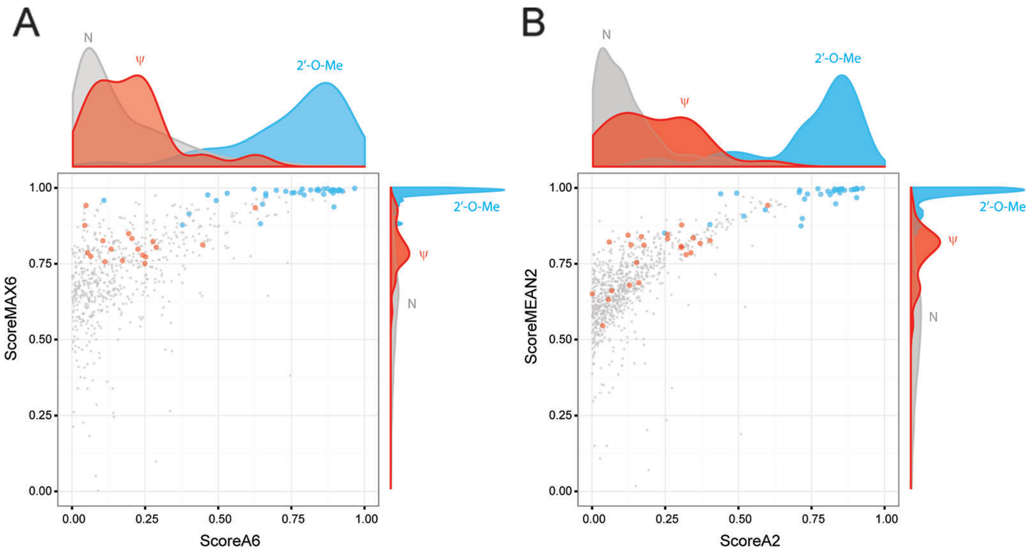


FIGURE 4 | Validation of ScoreMEAN2 and ScoreA2 with the *S. cerevisiae* rRNA RiboMethSeq dataset. Comparative distribution of ScoreA6/ScoreMAX6 signals **(A)** and ScoreA2/ScoreMEAN2 signals **(B)** for the same *S. cerevisiae* rRNA dataset. Graphs represent scatter plots for two scores, with the associated density plot on top (ScoreA6 or ScoreA2) and on the right (ScoreMAX6 and ScoreMEAN2). RiboMethSeq signals for 2'-O-Me positions (light blue), pseudouridines (red) and unmodified nucleotides (gray) are shown.

nucleotides are present at the 5'-end extremity of the fragment. This was experimentally observed for pseudouridine and other rare RNA modifications (Birkedal et al., 2015; Marchand et al., 2016). Since this reduced ligation efficiency also generates a "gap" in the sequencing profile, such FP signals are difficult to discriminate from undermethylated 2′-O-Me sites. Both scores (MEAN2 and A2) show a fair separation of values for 2′-O-Me nucleotides (blue) and unmodified residues (gray), but the peak for pseudouridines (red) partially overlaps with that of 2′-O-methylation. ScoreA2 shows better separation of 2′-O-Me and pseudouridine signals, while ScoreMEAN2 demonstrates a higher MCC and thus better performance for 2′-O-methylation detection (**Supplementary Figure S6**).

We also attempted to cumulate values for ScoreA2 and ScoreMEAN2 together by calculating their normalized sum (ScoreD). Despite the fact that ScoreA2 and ScoreMEAN2 generally pick out different FP hits, ScoreD does not further improve the performance (max MCC and FDR) compared to ScoreMEAN2 alone. In conclusion, the best results for detecting 2′-O-methylated residues were obtained with a calculation window of 4 nt (+/−2 nt) using ScoreA2 and ScoreMEAN2.

Validation of ScoreMEAN2 and ScoreA2 With Human and *S. cerevisiae* rRNA Datasets

To compare improvements associated with the use of ScoreMEAN2 and ScoreA2, we used independent human RiboMethSeq datasets obtained for other HeLa samples, human mammary epithelial cells (HME), human fibroblasts and Wharton's jelly MSC (19 altogether). Comparison of max MCC and FDR values obtained for Scores MAX6/MEAN2 and Scores A6/A2 calculated with different scoring schemes is shown on **Figure 3** and **Supplementary Figure S7**). In all cases, calculation scheme Score 2 improved detection of rRNA 2′-O-methylation (**Supplementary Table S3**). Additional validation was also performed with yeast *S. cerevisiae* rRNA RiboMethSeq dataset. **Figures 4A, B** show a side-by-side comparison of ScoreMAX6 and ScoreA6 values with their respective distributions (panel A), as well as ScoreMEAN2 with ScoreA2 (panel B). The new scoring scheme provides better separation of 2′-O-Me signals from those for pseudouridine and unmodified nucleotides, which is also confirmed by a better FDR and MCC values (**Supplementary Figure S8**). These data validate the newly proposed ScoreMEAN2 and ScoreA2, which can now be used for the detection of 2′-O-methylation in various RNA types.

CONCLUSION

The results of this optimization suggest that RiboMethSeq analysis of rRNA can be performed using short trimming lengths (10–12 nt) and adapted Bowtie2 alignment parameters, which allows a gain in sequencing information. Calculations of RiboMethSeq scores for short calculation intervals (4 neighboring nucleotides) and ScoreMEAN2 and ScoreA2 can be used for mapping of the modified positions. Quantification of 2′-O-Me is accomplished using MethScore2, also calculated for four neighbors. A minimal of 4–6 mln of raw reads (~400 reads/nt on average) can be used to evaluate the methylation level for known rRNA methylation sites, but at least 15 mln reads (~1,500 reads/nt on average) should be used to discover new methylation candidates. For different RNA species such as tRNAs, the trimming/seed length can be further reduced (up to 8 nt), but the calculation of RiboMethSeq scores with four neighbors can be maintained. The optimal read coverage is somewhat similar (15–20 mln raw reads for yeast/human tRNAs, respectively).

DATA AVAILABILITY STATEMENT

We used previously published datasets obtained for wild type yeast *S. cerevisiae* and human HeLa cells rRNA 2′-O-methylation analysis as well additional samples (Marchand et al., 2016; Erales et al., 2017; Sharma et al., 2017), accession numbers PRJEB43738, PRJEB35565 and PRJEB34951.

AUTHOR CONTRIBUTIONS

FP—designed and optimized treatment pipeline. VM, LA and VB-I—performed RiboMethSeq library preparation and analysis. VM, MH and YM—wrote the manuscript.

ACKNOWLEDGMENTS

We thank Team 05 of IMoPA UMR7365 for generous gift of HUVEC and bone marrow stem cells used for RNA extraction. This work was supported by FRCR EpiARN project from Lorraine Region to YM and SPP1784, HE 3397/13-2, HE 3397/14-1 from DFG to MH. This work was performed in the framework of EPITRAN COST initiative (CA16120).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00038/full#supplementary-material

REFERENCES

Birkedal, U., Christensen-Dalsgaard, M., Krogh, N., Sabarinathan, R., Gorodkin, J., and Nielsen, H. (2015). Profiling of ribose methylations in RNA by high-throughput sequencing. *Angew. Chem. Int. Ed. Engl.* 54, 451–455. doi: 10.1002/ange.201408362

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Buchhaupt, M., Sharma, S., Kellner, S., Oswald, S., Paetzold, M., Peifer, C., et al. (2014). Partial methylation at Am100 in 18S rRNA of baker's yeast reveals ribosome heterogeneity on the level of eukaryotic rRNA modification. *PloS One* 9, e89640. doi: 10.1371/journal.pone.0089640

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 515, 143–146. doi: 10.1038/nature13802

Dai, Q., Moshitch-Moshkovitz, S., Han, D., Kol, N., Amariglio, N., Rechavi, G., et al. (2017). Nm-seq maps 2'-O-methylation sites in human mRNA with base precision. *Nat. Methods* 14, 695–698. doi: 10.1038/nmeth.4294

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Erales, J., Marchand, V., Panthu, B., Gillot, S., Belin, S., Ghayad, S. E., et al. (2017). Evidence for rRNA 2'-O-methylation plasticity: control of intrinsic translational capabilities of human ribosomes. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12934–12939. doi: 10.1073/pnas.1707674114

Freund, I., Buhl, D. K., Boutin, S., Kotter, A., Pichot, F., Marchand, V., et al. (2019). 2'-O-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. *RNA* 25, 869–880. doi: 10.1261/rna.070243.118

Hauenschild, R., Tserovski, L., Schmid, K., Thüring, K., Winz, M.-L., Sharma, S., et al. (2015). The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* 43, 9950–9964. doi: 10.1093/nar/gkv895

Krogh, N., Jansson, M. D., Häfner, S. J., Tehler, D., Birkedal, U., Christensen-Dalsgaard, M., et al. (2016). Profiling of 2'-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic Acids Res.* 44, 7884–7895. doi: 10.1093/nar/gkw482

Krogh, N., Kongsbak-Wismann, M., Geisler, C., and Nielsen, H. (2017). Substoichiometric ribose methylations in spliceosomal snRNAs. *Org. Biomol. Chem.* 15, 8872–8876. doi: 10.1039/C7OB02317K

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lestrade, L., and Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* 34, D158–D162. doi: 10.1093/nar/gkj002

Marchand, V., Blanloeil-Oillo, F., Helm, M., and Motorin, Y. (2016). Illumina-based RiboMethSeq approach for mapping of 2'-O-Me residues in RNA. *Nucleic Acids Res.* 44, e135. doi: 10.1093/nar/gkw547

Marchand, V., Pichot, F., Thüring, K., Ayadi, L., Freund, I., Dalpke, A., et al. (2017a). Next-generation sequencing-based RiboMethSeq protocol for analysis of tRNA 2'-O-Methylation. *Biomolecules* 7 (1), 13. doi: 10.3390/biom7010013

Marchand, V., Ayadi, L., El Hajj, A., Blanloeil-Oillo, F., Helm, M., and Motorin, Y. (2017b). High-throughput mapping of 2'-O-Me residues in RNA using next-generation sequencing (Illumina RiboMethSeq Protocol). *Methods Mol. Biol.* 1562, 171–187. doi: 10.1007/978-1-4939-6807-7_12

Marchand, V., Ayadi, L., Ernst, F. G. M., Hertler, J., Bourguignon-Igel, V., Galvanin, A., et al. (2018). AlkAniline-Seq: profiling of m7 G and m3 C RNA modifications at single nucleotide resolution. *Angew. Chem. Int. Ed. Engl.* 57, 16785–16790. doi: 10.1002/anie.201810946

Piekna-Przybylska, D., Decatur, W. A., and Fournier, M. J. (2008). The 3D rRNA modification maps database: with interactive tools for ribosome analysis. *Nucleic Acids Res.* 36, D178–D183. doi: 10.1093/nar/gkm855

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., León-Ricardo, B. X., et al. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* 159, 148–162. doi: 10.1016/j.cell.2014.08.028

Schwartz, S. (2018). m1A within cytoplasmic mRNAs at single nucleotide resolution: a reconciled transcriptome-wide map. *RNA* 24, 1427–1436. doi: 10.1261/rna.067348.118

Sharma, S., Marchand, V., Motorin, Y., and Lafontaine, D. L. J. (2017). Identification of sites of 2'-O-methylation vulnerability in human ribosomal RNAs by systematic mapping. *Sci. Rep.* 7, 11490. doi: 10.1038/s41598-017-09734-9

Squires, J. E., Patel, H. R., Nousch, M., Sibbritt, T., Humphreys, D. T., Parker, B. J., et al. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* 40, 5023–5033. doi: 10.1093/nar/gks144

Taoka, M., Nobe, Y., Yamaki, Y., Yamauchi, Y., Ishikawa, H., Takahashi, N., et al. (2016). The complete chemical structure of Saccharomyces cerevisiae rRNA: partial pseudouridylation of U2345 in 25S rRNA by snoRNA snR9. *Nucleic Acids Res.* 44, 8951–8961. doi: 10.1093/nar/gkw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 06 March 2020
doi: 10.3389/fgene.2020.00194

Quantifying RNA Editing in Deep Transcriptome Datasets

Claudio Lo Giudice[1]†, Domenico Alessandro Silvestris[2]†, Shalom Hillel Roth[3], Eli Eisenberg[4,5], Graziano Pesole[1,6,7], Angela Gallo[2] and Ernesto Picardi[1,6,7]**

[1] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy, [2] RNA Editing Lab, Oncohaematology Department, IRCCS Ospedale Pediatrico "Bambino Gesù," Rome, Italy, [3] The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, [4] School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel, [5] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel, [6] Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, Bari, Italy, [7] National Institute of Biostructures and Biosystems, Rome, Italy

OPEN ACCESS

Edited by:
*Mattia Pelizzola,
Italian Institute of Technology, Italy*

Reviewed by:
*Meng How Tan,
Nanyang Technological University,
Singapore
Igor B. Rogozin,
National Institutes of Health (NIH),
United States*

***Correspondence:**
*Angela Gallo
angela.gallo@opbg.net
Ernesto Picardi
ernesto.picardi@uniba.it*

† These authors have contributed equally to this work

Specialty section:
*This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics*

Received: *03 December 2019*
Accepted: *18 February 2020*
Published: *06 March 2020*

Citation:
Lo Giudice C, Silvestris DA, Roth SH, Eisenberg E, Pesole G, Gallo A and Picardi E (2020) Quantifying RNA Editing in Deep Transcriptome Datasets. Front. Genet. 11:194. doi: 10.3389/fgene.2020.00194

Massive transcriptome sequencing through the RNAseq technology has enabled quantitative transcriptome-wide investigation of co-/post-transcriptional mechanisms such as alternative splicing and RNA editing. The latter is abundant in human transcriptomes in which million adenosines are deaminated into inosines by the ADAR enzymes. RNA editing modulates the innate immune response and its deregulation has been associated with different human diseases including autoimmune and inflammatory pathologies, neurodegenerative and psychiatric disorders, and tumors. Accurate profiling of RNA editing using deep transcriptome data is still a challenge, and the results depend strongly on processing and alignment steps taken. Accurate calling of the inosinome repertoire, however, is required to reliably quantify RNA editing and, in turn, investigate its biological and functional role across multiple samples. Using real RNAseq data, we demonstrate the impact of different bioinformatics steps on RNA editing detection and describe the main metrics to quantify its level of activity.

Keywords: RNA editing, transcriptome, RNAseq, deep sequencing, *Alu* editing index

INTRODUCTION

Eukaryotic organisms exhibit quite complex and dynamic transcriptomes whose regulation is essential for all cellular processes and for maintaining the homeostatic state (Mele et al., 2015). The complexity and dynamicity of transcriptomes depends on highly controlled and modulated post-transcriptional mechanisms such as alternative splicing and RNA modifications (Pan et al., 2008; Meyer and Jaffrey, 2014; Roundtree et al., 2017). The latter are now emerging as key players in promoting transcriptome diversity and fine tuning gene expression (Helm and Motorin, 2017; Roundtree et al., 2017). Transient and non-transient RNA modifications belong to the epitranscriptome world (Schwartz, 2016; Tajaddod et al., 2016; Boccaletto et al., 2018). Non-transient modifications occurring in a variety of RNA molecules and organisms through base insertions/deletions or substitutions are referred to as RNA editing changes (Gott and Emeson, 2000). In mammals, the most common RNA editing event involves the deamination of adenosine (A) into inosine (I), carried out by members of the ADAR family of enzymes acting on double stranded RNA (dsRNA) (Nishikura, 2016; Eisenberg and Levanon, 2018).

Deep transcriptome sequencing, through the RNAseq technology, has greatly promoted identification of RNA editing events at genomic scale, revealing the extent of A-to-I editing in humans, with more than 4.6 million modification sites identified so far. The majority of RNA editing modifications (>95%) resides in *Alu* repetitive elements that are widespread in human genes (accounting for around 10% of the human genome) (Levanon et al., 2004). Transcripts harboring two such elements with inverted orientations may fold to form dsRNA structures targeted by ADARs. In contrast, only a minute fraction of RNA editing events occurs in protein-coding genes and can lead to recoding, i.e., non-synonymous substitutions that generate novel protein isoforms. Recoding sites are enriched in neural tissues and over-represented in transcripts encoding proteins linked to the nervous system function (Rosenthal and Seeburg, 2012).

Accumulating evidence indicates that A-to-I RNA editing in mammals modulates the innate immune response (Mannion et al., 2014) and its deregulation has been observed in various human diseases including autoimmune and inflammatory tissue injury (Gallo and Locatelli, 2012; Roth et al., 2018; Shallev et al., 2018; Vlachogiannis et al., 2019), neurodegenerative and psychiatric disorders (Khermesh et al., 2016; Breen et al., 2019; Tran et al., 2019), and tumors (Gallo, 2013; Han et al., 2015; Paz-Yaacov et al., 2015; Silvestris et al., 2019).

An important property of RNA editing is that its levels vary across different tissues and cell types. Both the edited and unedited versions of transcripts co-exist in the same tissue or cell and the ratio between the unedited and edited variants is regulated by a variety of factors depending on tissue type or developmental stage. Consequently, quantifying RNA editing, detecting levels of edited variants or measuring the overall editing activity, are crucial for investigating its functional involvement and biological role.

A variety of bioinformatics programs and workflows have been released to profile RNA editing in deep transcriptome datasets (Picardi et al., 2015a; Diroma et al., 2019; Lo Giudice et al., 2020). Although based on different algorithms, all of them predict RNA editing candidates mitigating biases mainly due to sequencing errors, mapping errors, and genomic SNPs (Diroma et al., 2019). Hereafter, we describe a number of important metrics to quantify RNA editing in RNAseq experiments, enabling comparative analysis of whole inosinomes across multiple samples. Using real RNAseq data, we elaborate on different bioinformatics steps that have an impact on the profiling of RNA editing. These include pre-processing of raw reads or the specific strategy for alignment to the genome. As of to date no single computational methodology guarantees detection of all real editing events occurring in a sample, and the specific procedures for RNA editing detection and quantification in a given RNAseq dataset should be carefully selected, bearing in mind that the same procedure should be applied to all samples of a study to allow comparison of the results.

METHODS

RNAseq Samples, Pre-processing, and Alignment

RNAseq data from four tissues and 10 "body sites" (**Table 1** and **Supplementary Table S1**) were downloaded from Genotype-Tissue Expression (GTEx) Project through the dbGAP accession phs000424. Raw data were initially inspected using FASTQC and reads were trimmed using FASTP. Then, high quality reads were aligned onto the human genome (hg19 assembly from UCSC) using STAR v.2.5.2b (Dobin et al., 2013), providing a list of known gene annotations from GENCODE (Derrien et al., 2012). In addition, human cerebellum reads (accession SRR607967) were aligned to the human genome (hg19 and hg38 primary assemblies) using BWA v.0.7.17 (Li and Durbin, 2009) and HISAT2 v.2.1.0 (Kim et al., 2015) with known splice sites and exons from GENCODE.

RNA Editing Detection

A list of *de novo* RNA editing candidates per sample was generated using REDItools, following the filtering procedure described in Picardi et al. (2015b) and Lo Giudice et al. (2020). Aligned reads from run SRR607967 were also analyzed by JACUSA (Piechotta et al., 2017) using common basic filters. Hyper-edited reads were identified using the computational procedure described by (Porath et al., 2014).

RNA Editing Quantification

The overall RNA editing level *per* sample was calculated using a custom python script, taking as input a list of positions inferred by REDItools. The same program was also used to quantify RNA editing levels at known positions, downloaded from REDIportal database (including more than 4.5 million events in humans). The robustness of the overall editing metric over the number of RNA editing positions was tested selecting randomly growing numbers of positions from the REDIportal collection and calculating the overall editing *per* each sampling and "body site." Then, we measured the Pearson correlation between the overall editing calculated *per* each group of positions and the same metric detected using the whole database collection, by means of a custom script (pearsonr function from scipy python module).

Recoding index was also calculated using a custom python script working on REDItools tables. We considered as recoding sites all 1585 editing positions in REDIportal that are marked as non-synonymous in all three gene annotations available in the database (RefSeq, UCSC, and GENCODE). *Alu* editing index (AEI) was calculated using the methodology by Roth et al. (2019).

TABLE 1 | Summary table of experiments used.

Tissue	Body site	N. samples
Artery	Aorta	14
Artery	Tibial	14
Brain	Amygdala	13
Brain	Cerebellum	12
Brain	Frontal cortex	13
Brain	Hippocampus	11
Brain	Hypothalamus	14
Brain	Spinal cord	10
Lung	Lung	9
Muscle	Skeletal	13

Differential RNA Editing

Differential RNA editing at REDIportal recoding sites was identified using the non-parametric Mann–Whitney (MW) *U*-test. Recoding sites were collected *per* each artery tibial and cerebellum sample from REDItools tables. The comparison was carried out by a custom python script taking into account sites covered by at least 10 RNAseq reads in at least 50% of the samples *per* group. *p*-values were corrected for multiple testing using the Benjamini–Hochberg method.

Software, command lines, and scripts used in this work are available at the following GitHub repository https://github.com/BioinfoUNIBA/QEdit.

RESULTS AND DISCUSSION

Pre-processing and Alignment of RNAseq Experiments

Profiling RNA editing in whole transcriptome data is yet a challenging task, due to sequencing errors, read-mapping errors, genome-encoded polymorphisms (SNPs), somatic mutations, and spontaneous RNA chemical changes. SNPs and somatic mutations may be partly filtered out using genomic reads from matched whole genome sequencing (WGS) or whole exome sequencing (WXS) experiments, as well as tables of known SNPs from public databases. Alignment and sequencing errors may be partly removed using stringent filters of read and base quality. All of these aforementioned issues require careful design and tuning of computational pipelines to detect RNA editing candidates, as each step or procedure or software can affect the yield and quality of predictions.

Here we demonstrate the effects of pre-processing and genome alignment steps on RNA editing calling using a single GTEx RNAseq experiment from human cerebellum (run accession SRR607967). Raw reads were initially inspected using FASTQC and their low quality regions were removed by means of FASTP. Two datasets were generated, the first containing original raw reads and the second including trimmed reads. Both datasets were aligned onto the hg19 and hg38 reference chromosomes of the human genome using three different aligners, BWA designed for unspliced reads and STAR and HISAT2 optimized for handling spliced reads. Resulting multi-alignments were processed with REDItools in order to provide the distribution of single RNA variants according to a common basic filtering scheme. Known SNPs from the WGS of the same individual (run accession SRR2165704) were removed. In all tested cases, we achieved quite similar distributions, in which A-to-G and T-to-C changes (putative editing events on the direct or reverse strand) are over-represented, suggesting enrichment in true RNA editing events (**Figure 1**). However, the number of detected sites varied

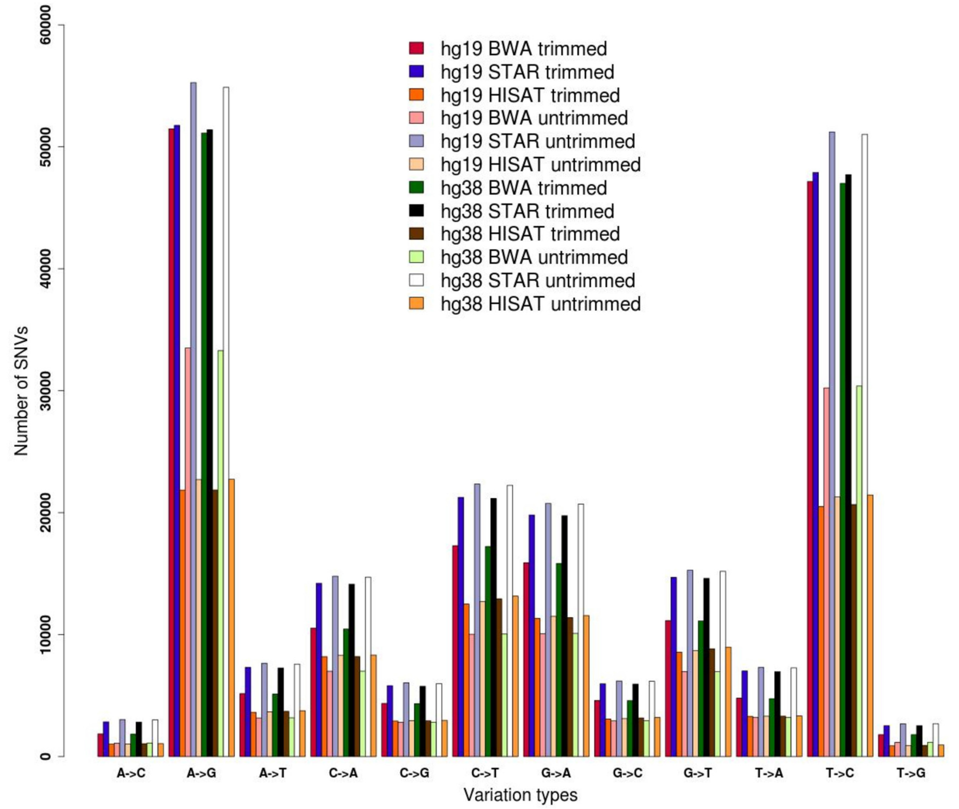


FIGURE 1 | Distribution of single nucleotide variants detected by REDItools on trimmed and untrimmed reads (from accession SRR607967) aligned by means of BWA, STAR, and HISAT2 onto hg19 and hg38 human genome assemblies.

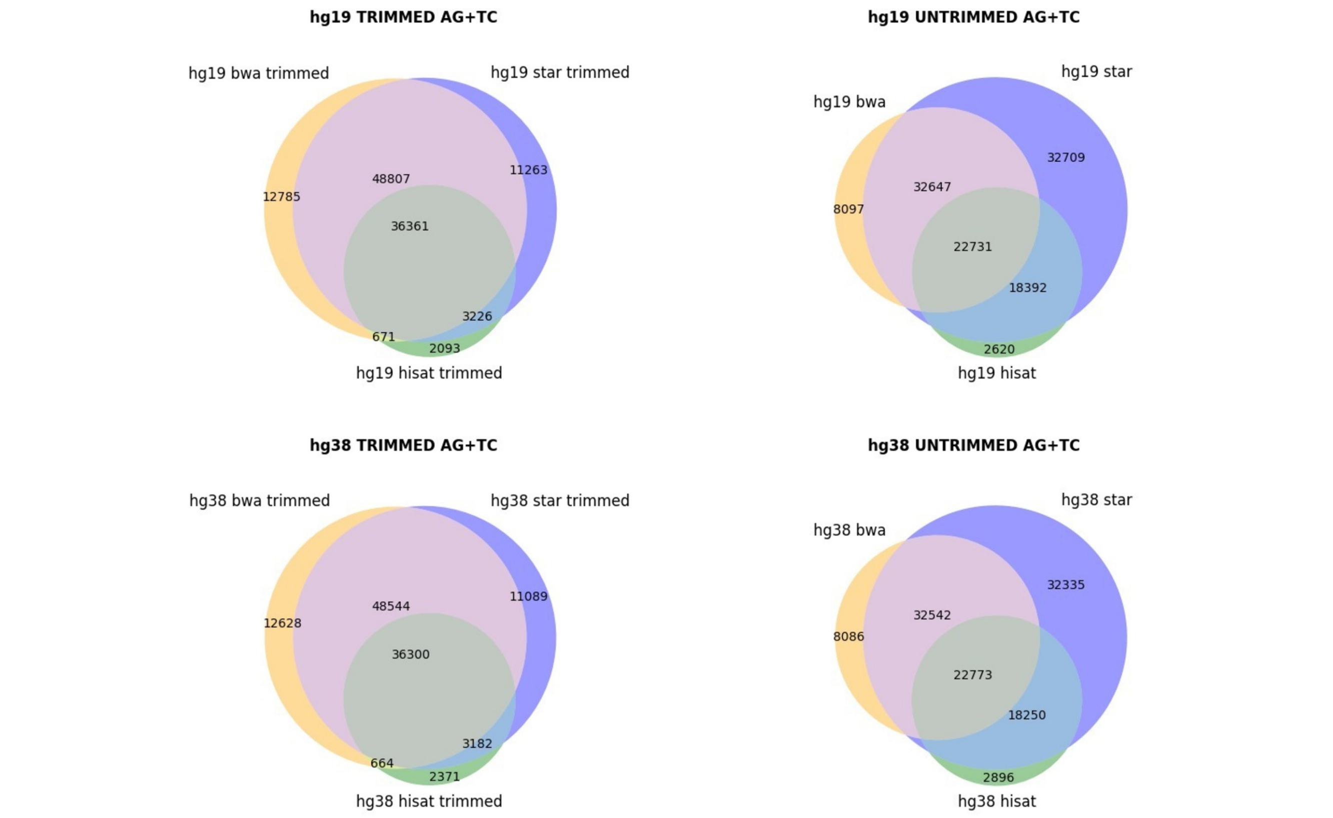


FIGURE 2 | Venn diagrams, showing the AG/TC overlapping positions for BWA, STAR, and HISAT2 aligners. The comparison is made for trimmed and untrimmed reads mapping onto hg19 and hg38 assemblies, respectively.

depending on the processing steps, suggesting that the trimming procedure as well as the aligner type affect the detection of RNA editing. The three different aligners resulted in different results, reflecting the slightly different algorithms. STAR has returned the highest number of candidates. Surprisingly, HISAT2 yielded the lowest number of variants, even though it is splice-aware and did align the same proportion of reads as STAR (**Figure 1** and **Supplementary Table S2**).

The genome version used (hg19 and hg38 human genome assemblies) did not make an appreciable difference (**Figure 1**), but the alignment of raw or trimmed reads did have an aligner-dependent effect (**Figure 1**). Although deviations in all checked cases do not appear graphically marked, they do influence the final list of candidates (**Figure 2**). We thus see that simple computational steps or the adoption of specific software can dramatically change the final results and impact commonly used metrics for quantification of global or local RNA editing activity in a sample. Adopting the same computational pipeline to analyze multiple samples or compare results from already published works is highly recommended.

RNA Editing Detection

Once trimming and alignment steps have been performed, the final list of RNA editing candidates strongly depends on the methodology used to call them. In general, two types of approaches can be pursued, *de novo* or "known". The former aims to identify all potential RNA editing events of a sample or the hyper-edited regions only without relying on previously known sets of editing positions, while the latter focuses on a restricted number of known changes from literature or well-established databases.

De novo Approach

Several software packages to detect *de novo* RNA editing events in deep transcriptome data have been released to date. They all suffer from some level of false positives, and each tool requires fine tuning of a variety of parameters that can strongly affect the quality of results and, thus, sensitivity and specificity of predictions (Diroma et al., 2019). The behavior of several RNA editing detection programs has been recently assessed (Diroma et al., 2019). Here we analyze comparatively two *de novo* approaches for RNA editing identification, REDItools (Picardi and Pesole, 2013) and JACUSA (Piechotta et al., 2017), using the same aligned human cerebellum reads. The two methods require traversing multiple alignments of reads through a pileup function. REDItools detect events applying different empirical filters while JACUSA implements a statistical model for variant calling. Both tools were applied to trimmed reads aligned onto the hg19 genome by STAR, followed by common basic filters such as the removal of sites in homopolymeric stretches longer than five residues or falling in the first and last six bases of a read, the exclusion of positions covered by less than 10 reads and showing a phred quality score less than 30.

The two programs return a similar number of variants, but with different precision. REDItools yielded 99,657 putative editing sites (49.56% of all observed modification sites) while JACUSA predicted 91,955 putative editing sites (75.23% of all observed modification sites) (**Figure 3**). In this specific example, JACUSA appeared more stringent than REDItools showing a higher signal-to-noise ratio, likely due to its statistical model and further filtering step by a companion R script, the JacusaHelper. This example demonstrates that RNA editing calling tools should be used with care, paying attention in advance to the various combinations of parameters and the experimental features of samples. A good practice is to estimate the false discovery rate comparing the A-to-G fraction (and T-to-C for unstranded reads) with the noise due to other base changes not expected to be edited, and then tune the parameters accordingly. Indeed, multiple filters can greatly improve the quality of final results. For example, to mitigate mapping errors (by Blat re-alignment) and other spurious changes occurring near splice sites or in genomic regions containing poorly aligned reads we applied more stringent filters to REDItools (Lo Giudice et al., 2020). Doing so, the number of variants detected in the same sample dropped down to only 52,400 sites including about 99% (51,888 positions) of potential RNA editing events (A-to-G and T-to-C changes) with a very low estimated false discovery rate, <1%. The effect of the different filtering steps on the distribution of RNA variants is shown in **Figure 4**. Importantly, the third step (coverage cut-off) results in a sizable drop in the number of excess AG/TC mismatches. While this step is necessary in order to achieve a good signal-to-noise ratio, one should bear in mind that the vast majority of the signal is lost during this step.

Note that in other species, e.g., mice, *Alu* elements are not present and the number of expected RNA editing candidates is much lower compared to humans (Neeman et al., 2006; Ramaswami and Li, 2014). This might require re-tuning the alignment and calling parameters. Furthermore, in case multiple samples from biological replicates are available, these may be used to further improve final results, looking only at putative RNA editing candidates common to all replicates.

"Known" Approach

The *de novo* approach generates a list of candidate sites likely to be edited in the specific RNAseq dataset. Sometimes, however, it could be useful to focus on a set of known events in order to better investigate RNA editing dynamics in different experimental contexts. For example, RNA editing could be profiled in known recoding events of neurotransmitter receptors to study its involvement in synaptic function or its deregulation in neurological/psychiatric disorders or cancer (Gallo, 2013; Han et al., 2015; Paz-Yaacov et al., 2015; Khermesh et al., 2016; Silvestris et al., 2019). REDItools package is the most suitable tool for this task (Picardi and Pesole, 2013). Providing a list of genomic positions and a pre-aligned file of RNAseq reads, it recovers the exact site and the corresponding RNA editing level. The "known" approach has been successfully applied also to large scale genomic projects. In the specialized database REDIportal (Picardi et al., 2016), for example, REDItools have been used to interrogate multiple read alignments from 2660 GTEx RNAseq experiments employing a large collection of known RNA editing sites from the ATLAS repository (Picardi et al., 2015b) and DARNED database (Kiran et al., 2013). Another example is The Cancer RNA Editome Atlas (TCEA) (Lin and Chen, 2019), where REDIportal positions (4,656,896) have been explored in more than 11,000 RNAseq data from the TCGA project (Cancer Genome Atlas Research Network et al., 2013).

Hyper-Editing

ADAR enzymes are known to have the ability to deaminate clusters of adjacent adenosines leading to hyper-edited RNA molecules (Eisenberg, 2016). Many RNA editing calling programs, however, fail to discover hyper-editing events because of the high number of mismatches *per* read that avoids its correct alignment on the genome (Porath et al., 2014). Heavily edited reads can be detected through a specific computational protocol in which not aligned sequences are rescued and mapped again onto a transformed genome replacing As with Gs (Porath et al., 2014). Since hyper-editing occurs mainly in *Alu* repetitive elements, it could lead to altered AEI values with a trend to underestimate the RNA editing activity *per* sample. As an example, we applied the computational strategy by Porath et al. (2014) to the above cerebellum RNAseq experiment (run accession SRR607967) using 3,490,661 unmapped reads by STAR. The alignment onto the transformed human genome yielded 19,377 reads enriched in A-to-G clusters, corresponding to 124,546 RNA editing changes. Of these, only 3586 were present in the filtered list of candidates by REDItools. Consequently, more than 120,000 A-to-G RNA editing events, missed by REDItools in the previous analysis, have been *de novo* identified in hyper-edited regions. So, events falling in hyper-edited reads should not be excluded *a priori* since they may represent a considerable fraction of sites. Large scale investigations based on TCGA samples have proven that the number of unique editing sites identified in hyper-edited regions follows the same trend as the AEI index calculated excluding hyper-edited reads (Paz-Yaacov et al., 2015). These findings suggest that the expected AEI underestimation does not significantly affect the global RNA editing activity measured at *Alu* level.

Metrics for RNA Editing Quantification

Once RNA editing has been detected in RNAseq samples, quantification is the next step that allows comparing values across samples and study of the potential role of RNA editing in different experimental conditions, such as its involvement in human disorders. Quantification of RNA editing is also important to identify molecular markers that could be the target for engineered ADAR enzymes or modified CRISPR-Cas systems, according to the recent paradigm of the precision medicine. Quantification of RNA editing has always been a challenging task, especially in the last few years in which deep

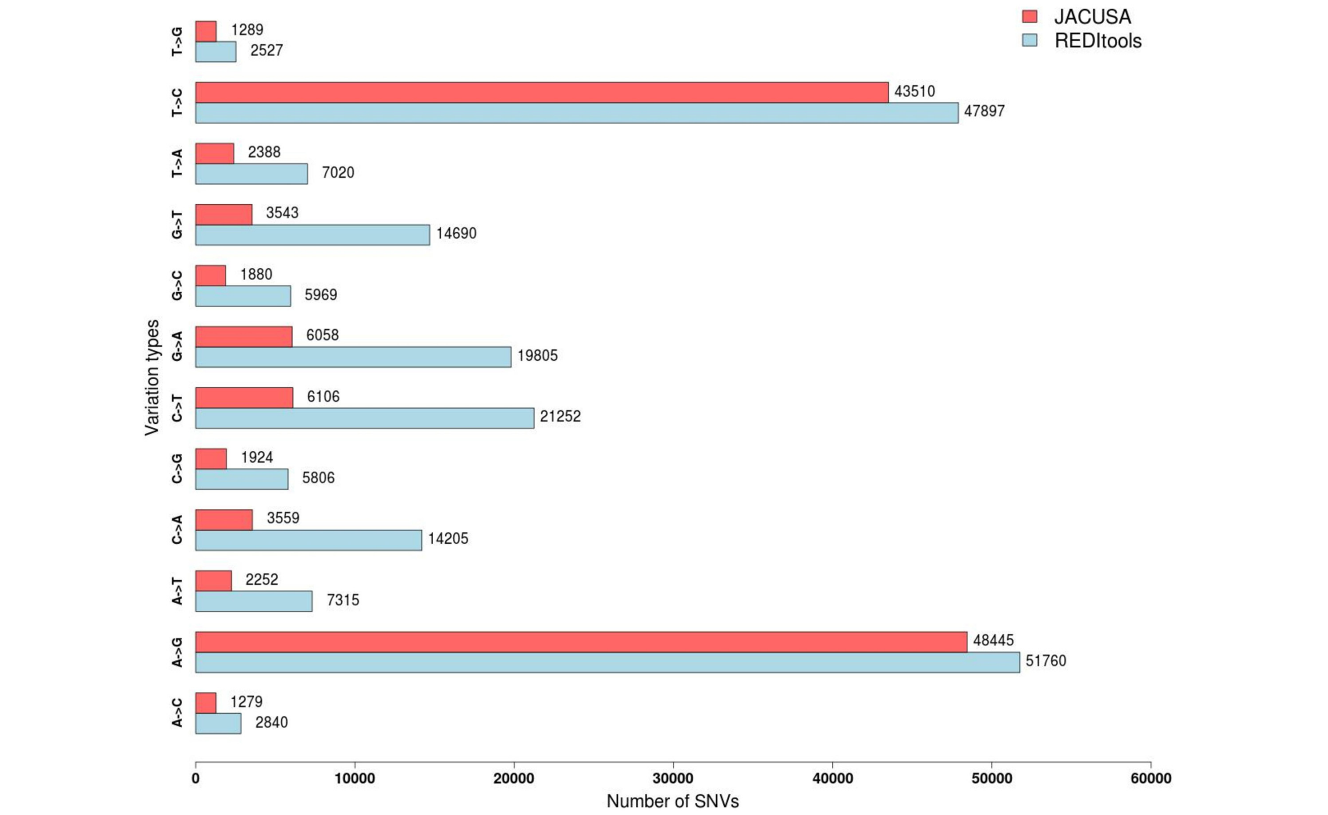


FIGURE 3 | Distribution of single nucleotide variants detected by JACUSA vs REDItools on trimmed reads SRR607967 aligned by STAR on hg19 human genome assembly.

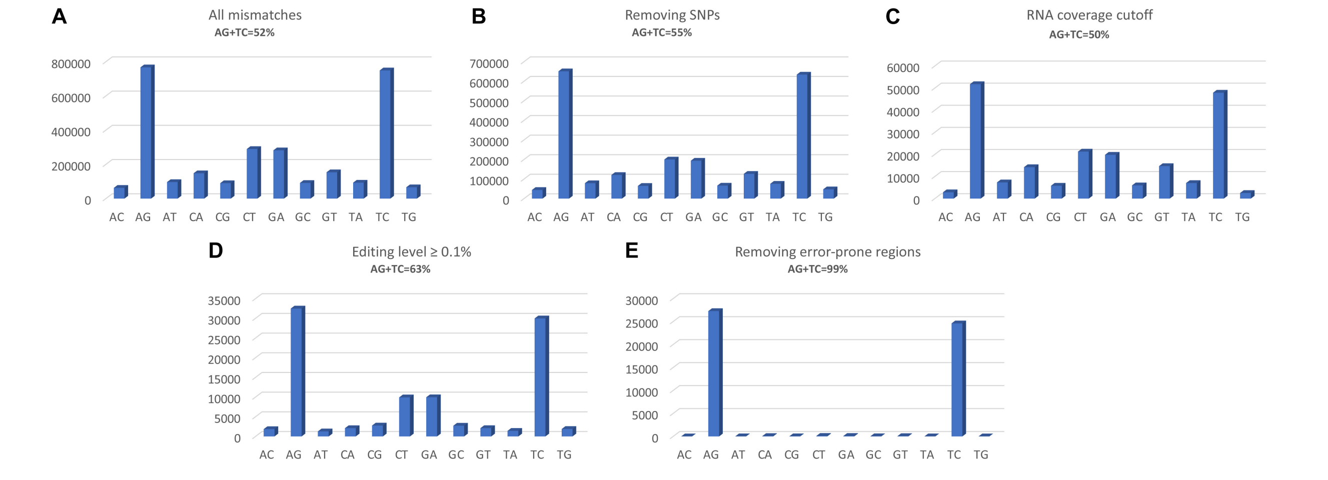


FIGURE 4 | Distributions of RNA variants detected by REDItools obtained following the different filtering steps: **(A)** all mismatches found following mapping, with a phred quality score of at least 30; **(B)** selecting only sites supported by at least 10 WGS reads and removing positions in dbSNP; **(C)** selecting sites covered by at least 10 reads and not falling in homopolymeric stretches longer than five residues or in the first and last six bases of a read; **(D)** selecting sites with an editing frequency of at least 0.1; **(E)** excluding sites in mis-mapped reads (by Blat correction) or near splice sites or in genomic regions containing poorly aligned reads.

transcriptome sequencing has enabled large scale investigations. Several metrics have been proposed, some of them take into account the global RNA editing activity (Tan et al., 2017; Roth et al., 2019), while other approaches focus on specific sites only (Khermesh et al., 2016; Silvestris et al., 2019). Below, we illustrate the main metrics using GTEx RNAseq data from four tissues and ten "body sites" (see section "Methods" for further details).

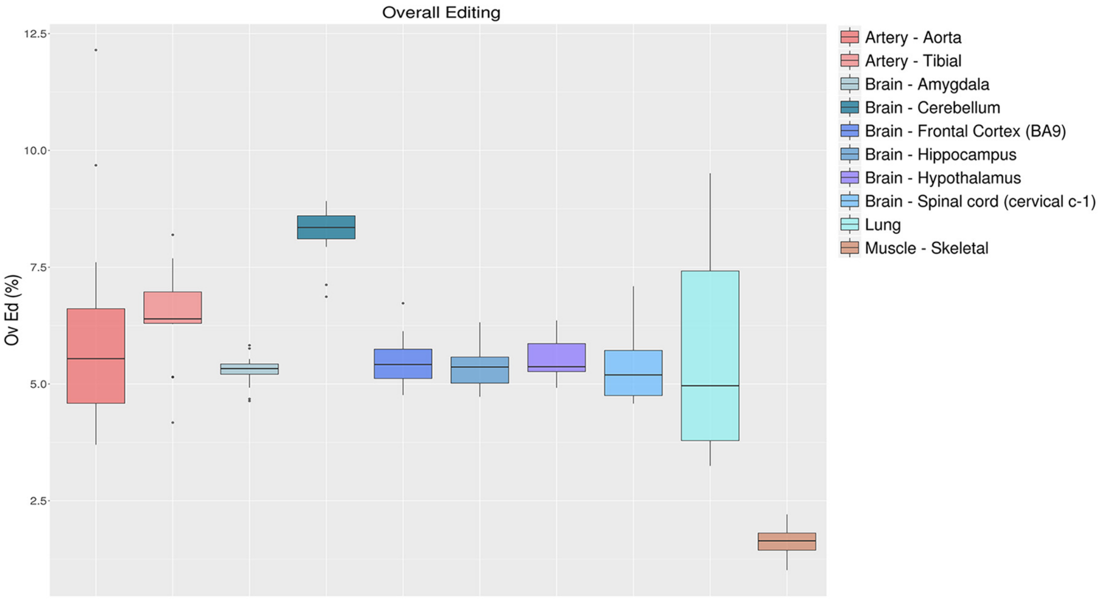


FIGURE 5 | Overall editing levels in 10 selected "body sites" from the GTEx project. Each box plot represents samples from one tissue type. The overall editing level is defined as the percentage of edited nucleotides at all known editing sites. Cerebellum and skeletal muscle emerge, respectively, as the most edited tissue and the less-edited tissue among the analyzed tissues.

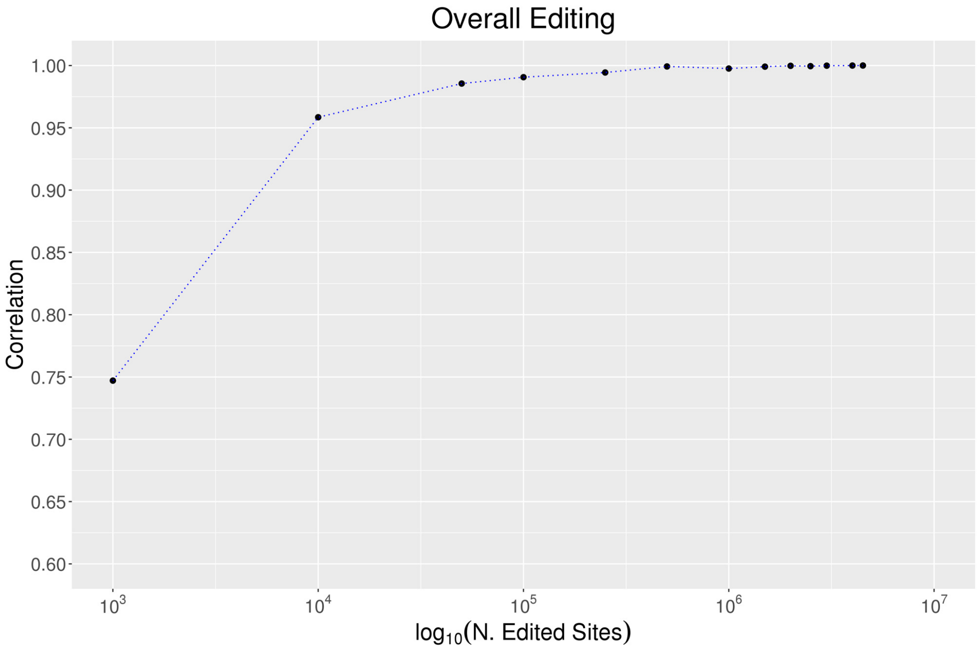


FIGURE 6 | The effect of the number of sites on the overall editing. We calculated the overall editing calculated in all 123 GTEx samples using a growing number of positions randomly selected from REDIportal database. The Pearson correlations between the overall editing measured per each group of positions and the same metric on the entire REDIportal collection are depicted.

Overall Editing Level

To quantify the global RNA editing in a sample, one can average the editing levels measured over the sites detected previously, or by *de novo* methods (Tan et al., 2017). This metric, referred to as the overall editing, is determined as the total number of reads with G at all known editing positions over the number of all reads covering the positions without imposing specific sequencing coverage criteria (Tan et al., 2017). The overall editing depends on the number of known editing sites included in the analysis that have to be the same for all samples analyzed. Using *de novo* editing events for this purpose is not recommended, as the number of detected sites is unevenly distributed across samples and strongly depends on the amount of raw reads input and the bioinformatics procedure (Picardi et al., 2015b; Diroma et al., 2019). Even merging *de novo* candidates from all samples of interest does not remove the coverage bias altogether. Alternatively, one may calculate the overall editing employing known events stored in public databases such as REDIportal (Picardi et al., 2016), RADAR (Ramaswami and Li, 2014), or DARNED (Kiran et al., 2013). To illustrate the behavior of the overall editing index, we calculated this metric in 123 GTEx RNAseq experiments from 10 "body sites" employing REDIportal as it stores the largest public collection of human RNA editing annotations (4,665,677 sites in its last release). As shown in **Figure 5**, RNA editing appeared reduced in skeletal muscle compared to other tissues, as already observed in previous studies (Picardi et al., 2015b; Tan et al., 2017). On the contrary, cerebellum displayed the highest RNA editing level. These results are consistent with the *Alu* editing level among "body sites" (Roth et al., 2019) with cerebellum emerging as the top tissue carrying the highest editing level, higher that other brain regions including cortex. It has been estimated that there are about 3.6 times as many neurons in the cerebellum as in the cortex (Herculano-Houzel, 2010). Possibly, the higher level in cerebellum is merely a result of a higher fraction of neurons in this tissue, as neurons are highly edited compared to other brain cells (Gal-Mark et al., 2017).

To evaluate the effect of the number of RNA editing positions on the robustness of the overall editing metric, we randomly selected growing numbers of positions from the REDIportal collection and calculated the overall editing *per* each sampling and "body site." Assuming the highest accuracy when all REDIportal positions are used, we measured the correlation between the overall editing calculated *per* each group of positions and the same metric detected using the whole database collection. As reported in **Figure 6**, 100,000 RNA editing positions are sufficient to obtain a very high correlation (Pearson $R = 0.99$ Pval $<< 10^{-4}$) with the entire REDIportal database. Using the RNA editing sites from DARNED (333,215 sites) and RADAR (2,576,459 sites), we obtained a

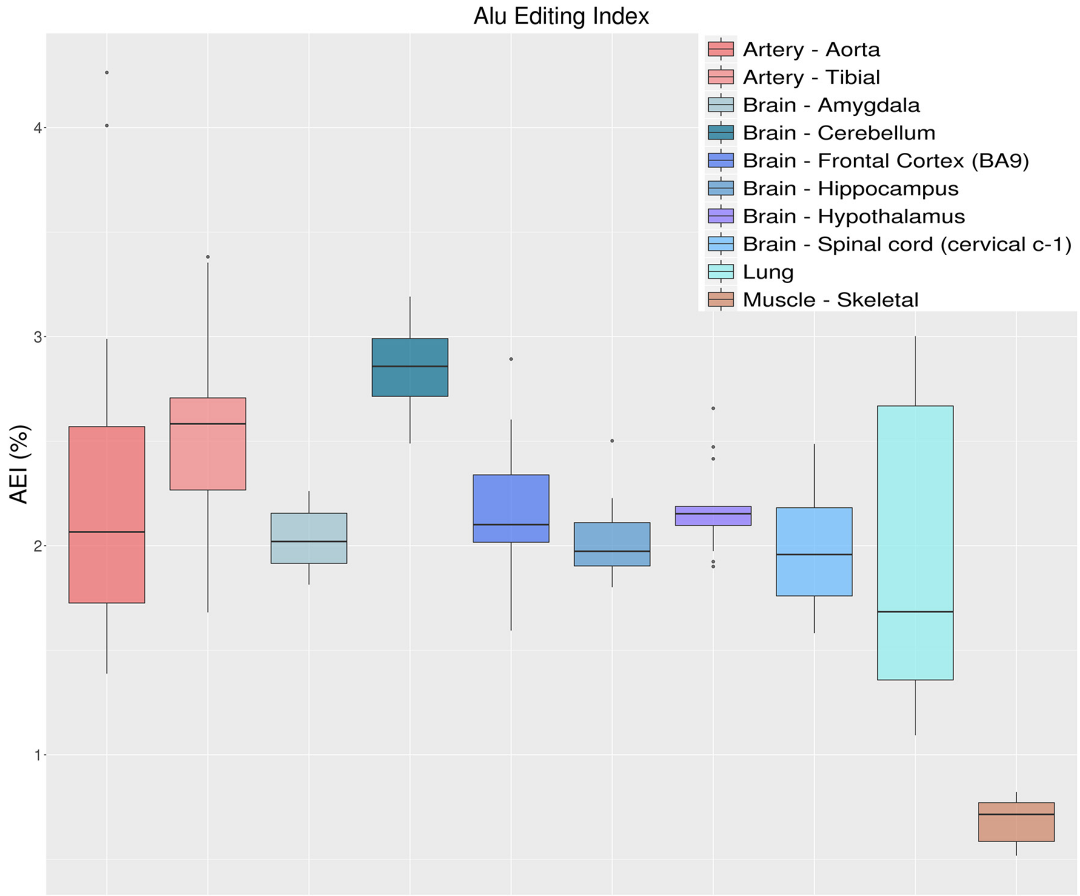


FIGURE 7 | Distributions of *Alu* editing index (AEI) values over 10 selected tissue types from the GTEx project. AEI represents the weighted average editing level across all expressed *Alu* sequences. Distributions are presented as box-plots. AEI clearly recapitulates the same trend as overall editing thus confirming that the sites in *Alu* regions are those that have the greatest impact on the global editing activity.

correlation with REDIportal of 95% (Pval $<< 10^{-4}$) and 99% (Pval $<< 10^{-4}$), respectively.

Alu Editing Index

Another metric to quantify the global RNA editing activity is to calculate the weighted average of editing events occurring in all adenosines within *Alu* elements, defined as the AEI. As mentioned above, the vast majority of editing activity takes place within *Alu* elements, with almost every adenosine in the ADAR-targeted *Alu* repeats being edited to some extent (Bazak et al., 2014a). The AEI is defined to be the ratio (for convenience in percentage) of the number observed A-to-G mismatches to the total coverage of adenosines (both A-A matches and presumed editing events, A-to-G mismatches). It is therefore the weighted average of the measured editing levels weighted by the coverage of each site (Bazak et al., 2014b). The AEI avoids the quantification of editing rates per-sites, while accounting for editing in lowly covered regions. It also frees the user from dependence on public databases that might be continuously changing (or even unavailable for other species). Since the AEI is calculated over millions of positions it is highly robust to the number of input raw reads, and as few as one million input reads already provide a consistent and almost invariable signal (Roth et al., 2019). It is, however, affected by the alignment process (i.e., aligner and read lengths), but preserves the relative rank of each sample. As an example, **Figure 7** shows the distribution of AEI values for 123 GTEx samples, calculated as described in Roth et al. (2019). Results indicate a general agreement between the measured AEI and the overall editing index depicted above (**Figure 5**). It should be noted that this approach is not limited to the human genome. One can use the index for any organism, as long as a large set of highly editable elements (often, SINE elements) is available and the editing is strong enough to result in a sufficiently large signal-to-noise ratio.

Recoding Index

Similarly to the overall editing, recoding activity due to RNA editing could be quantified, focusing on editing levels at recoding positions (residing in coding protein genes). For example, one may calculate the weighted average over all known recoding sites, known as the recoding editing index (REI) (Silvestris et al., 2019). This measure is well correlated with ADAR2 expression, at least in normal brain (Silvestris et al., 2019), and may be a good indicator of ADAR2 deaminase activity. Interestingly, REI may be utilized to investigate RNA editing deregulation in different brain regions or neurological disorders (Khermesh et al., 2016) or cancer (Silvestris et al., 2019). REI is simply defined as the number of reads with G at recoding positions over the number

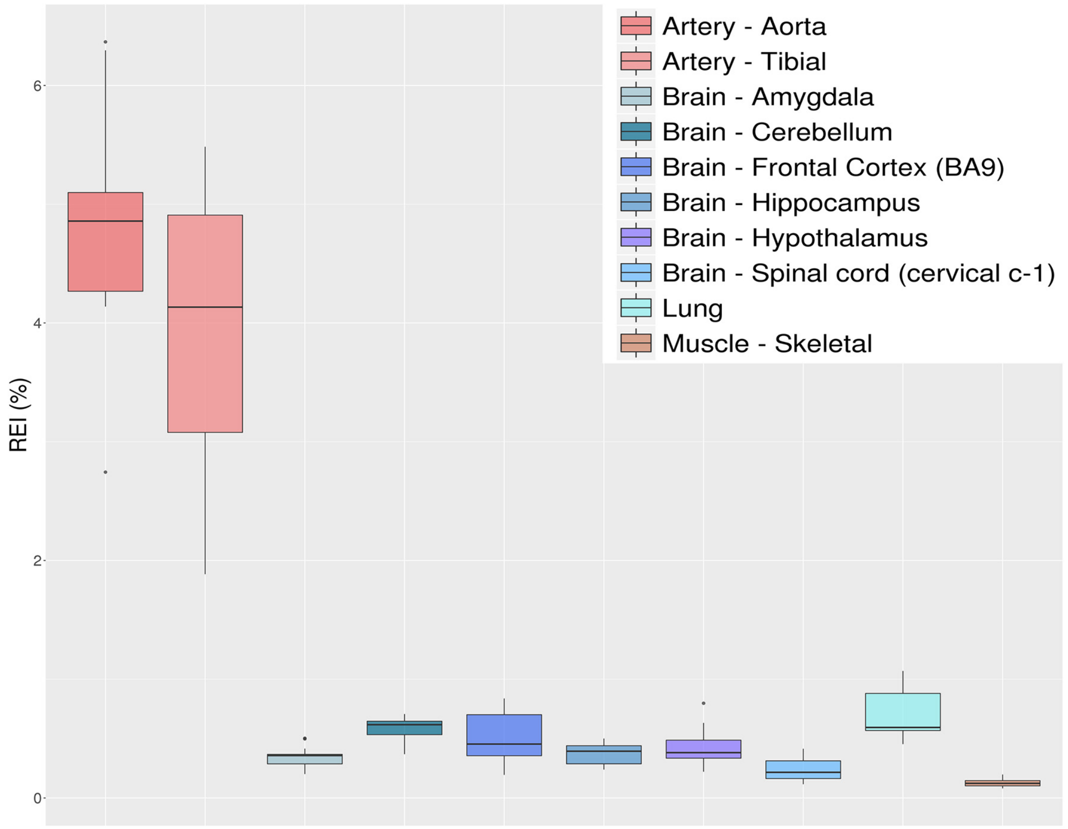


FIGURE 8 | Distributions of recoding editing index (REI) values over 10 selected tissues from the GTEx project reported as box-plots. REI is calculated as the weighted average of editing levels over all known recoding sites from the REDIportal database. Most brain sub-tissues show similar levels of recoding editing. A remarkable exception is represented by the aorta and tibial artery showing a surprisingly high editing level.

of all reads covering the same positions (same as AEI, but for the recoding sites). As in the case for the overall editing, the reliability of REI depends on the number of recoding sites to assay. Indexing over very small numbers, e.g., the 35 recoding sites known to be conserved across the mammalian lineage (Pinto et al., 2014), could lead to biased values and misleading conclusions. The list of recoding sites can be obtained from databases such as REDIportal (Picardi et al., 2016), RADAR (Ramaswami and Li, 2014), or DARNED (Kiran et al., 2013). However, one should bear in mind that the false positive level of recoding sites in these public collections is notoriously high.

Here, we show the REI results using 1585 non-synonymous RNA editing events from REDIportal (see selection criteria in section "Methods") for the above GTEx RNAseq experiments (**Figure 8**). Our results, similarly to those by Tan et al. (2017) from the complete GTEx dataset, show a very high recoding activity at arteries compared to other tissues, with lung and brain being at similar levels and skeletal muscle showing the lowest REI levels. Of note, the ADAR2 expression level (as shown by GTEx in **Supplementary Figure S1**) overlaps well the results shown in **Figure 8**. So far, many studies, including ours, have underlined the important role played by RNA editing at recoding sites in the central nervous system (CNS). In contrast, the role of A-to-I RNA editing in angiogenesis, artery, endothelium, and vascular disease was only recently explored (Stellos et al., 2016; Jain et al., 2018). While Stellos et al. (2016) have pointed to ADAR1 activity within the 3′ untranslated region (3′ UTR) of cathepsin S mRNA (*CTSS*), Jain et al. (2018) reported that recoding at *FLNA* (Q/R) is an important regulator of vascular contraction and blood pressure. Our data and a previous study (Jain et al., 2018) indicated the presence of some almost fully edited sites in artery, similar to the *GRIA2* Q/R in CNS, and extended the list of important recoding sites in artery that may play a crucial role in vascular physiology and diseases (**Figure 9**). Indeed, among the top edited genes in arteries, there is the Insulin-like growth factor-binding protein 7 (IGFBP7). IGFBP7 is a secreted protein involved in diverse biological functions, from apoptosis to inhibition/stimulation of growth and angiogenesis (Brahmkhatri et al., 2015). Proteolytic processing of IGFBP7 modulates its biological activity as it can stimulate growth of DLD−1 colon carcinoma cells in synergy with insulin/IGF−I but, if cleaved, IGFBP7 completely abolishes this growth-stimulatory activity (Ahmed et al., 2003). Interestingly, editing of *IGFBP7* transcripts (K/R site) affects the protein's susceptibility to proteolytic cleavage, thus providing a

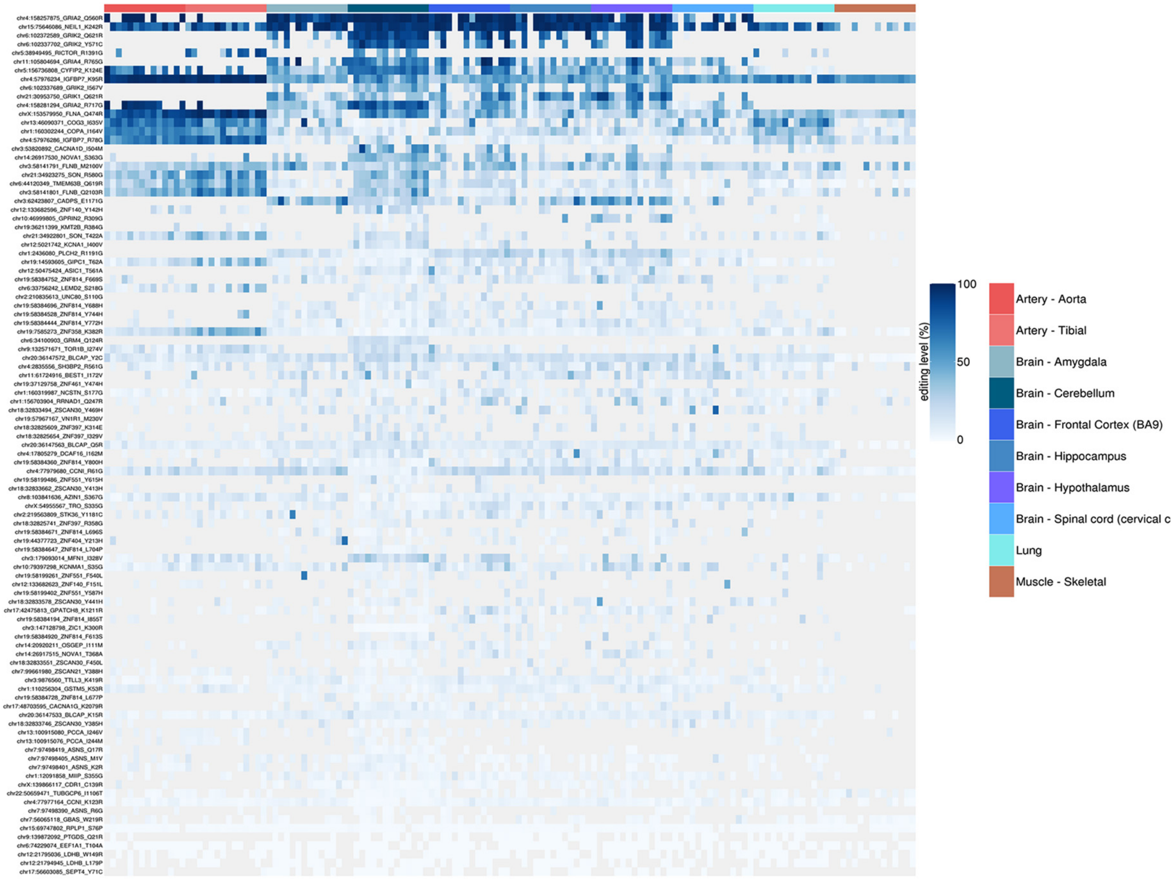


FIGURE 9 | Heatmap representing RNA editing levels at 99 selected recoding events. Body sites are reported in the same order as in the previous box-plots and follow the same color code. The hierarchical clustering (dendrogram not shown) of the recoding sites shows how the artery (both aorta and tibial) are characterized by a very peculiar and specific set of strongly (>90%) edited sites, thus suggesting a possible key functional role of these sites in the vascular system.

means for a cell to modulate its multiple activity through A-to-I RNA editing (Godfried Sie et al., 2012).

The REI is a measure of global RNA editing activity at recoding sites. However, one should bear in mind that recoding activity is often unevenly distributed across the different sites. High REI values could mean overall high recoding activity, but might also occur at a few highly expressed and highly edited sites only. In the aforementioned artery samples, for instance, three recoding events in *IGFBP7* and *FLNA* transcripts account for more than 90% of all edited Gs, and for the high value of the REI as compared to other tissues. In case one is interested in the distribution, a common practice is to look at graphical visualizations of editing levels through all sites of interest, using, for example, a heatmap plot (**Figure 9**).

Differential RNA Editing

An important question related to the RNA editing profiling is the identification of differentially edited sites. A variety of statistical tests have been proposed so far, but reliable, consistent, and reproducible detection of dysregulated RNA editing is still a major task. The observed A-to-I levels at individual sites depend strongly on the methodology used to call them, sequencing depth and coverage. Events residing in repetitive elements, comprising the majority of A-to-I changes, exhibit low levels (typically lower than 0.01), requiring ultra-high coverage for reliable detection and quantification. A given position could appear edited in some samples but unedited in others (because of limited coverage), a fact that is often ignored in the statistical testing. Sometimes, when the number of samples is sufficiently high, missing editing levels could be imputed using methods based on the principal component analysis (Josse and Husson, 2016), chained equations (Buuren and Groothuis-Oudshoorn, 2011), or random forest (Stekhoven and Bühlmann, 2012).

Finally, the large number of editing sites requires an aggressive multiple-testing correction, and severely limits the statistical power. This leads to an underestimate of the number of differentially edited sites.

Identification of differential RNA editing is most relevant at recoding sites, where altered A-to-I levels could lead to different protein isoforms. Editing dysregulation at recoding sites between two groups of samples is often assayed applying the two-tailed MW *U*-test followed by Benjamin–Hochberg multiple test

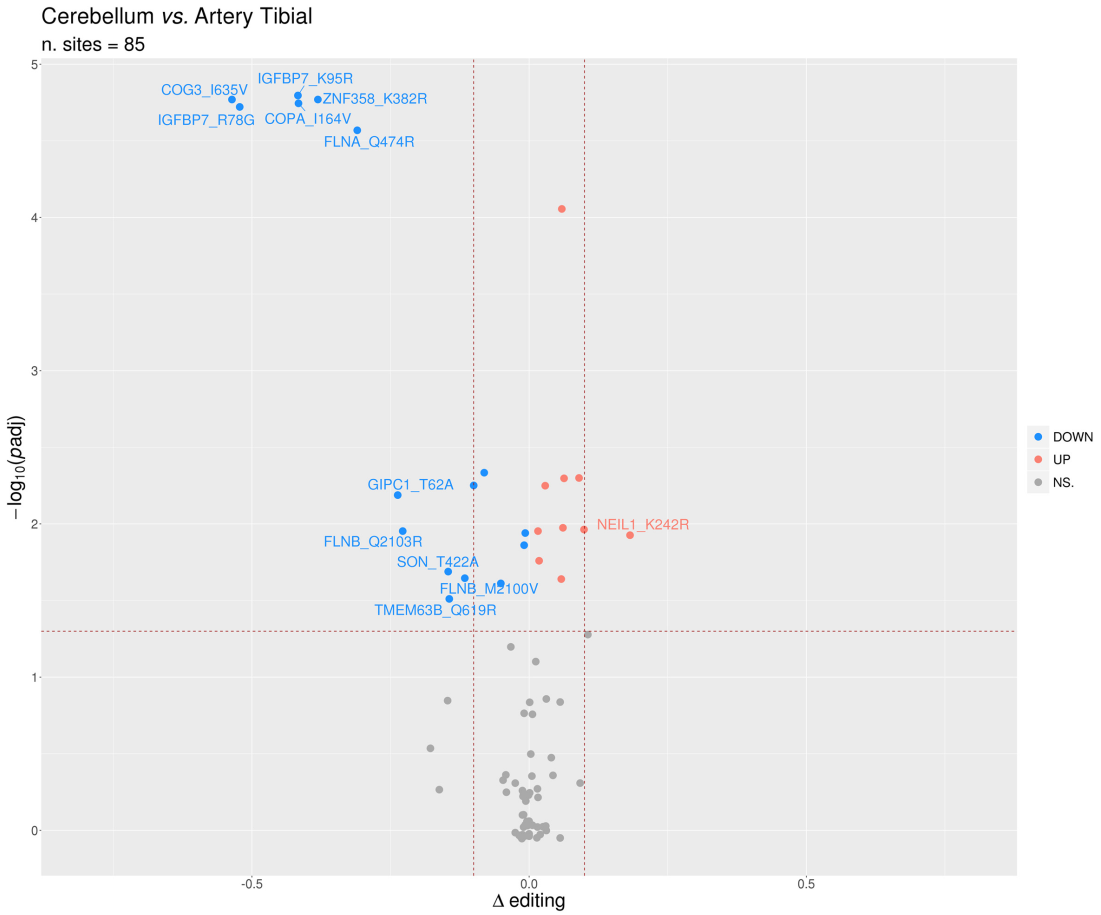


FIGURE 10 | Volcano plot reporting the differentially edited sites between cerebellum and tibial artery. The horizontal dotted line marks a multiple test-corrected level of significance (adjusted padj < 0.05, Mann–Whitney with Benjamini–Hochberg correction). The vertical dotted lines indicate a Delta editing of 0.1 and -0.1. Red, blue, and gray points indicate, respectively, over-edited (UP) sites, under-edited (DOWN) sites, and non-significative sites (NS.).

corrections. For example, such an approach was used to identify many recoding sites differentially edited in cancer compared with normal samples (Maas et al., 2001; Paz et al., 2007; Cenci et al., 2008; Chen et al., 2013; Qin et al., 2014; Han et al., 2015; Paz-Yaacov et al., 2015; Hu et al., 2016; Lin and Chen, 2019; Silvestris et al., 2019). Here, we demonstrate this approach by detecting statistically significant differentially recoded sites between 14 artery tibial and 12 cerebellum samples, looking at 1585 non-synonymous REDIportal positions quantified using REDItools. We considered only sites supported by at least 10 RNAseq reads in at least the three samples per group, thus obtaining 85 positions to test for differential RNA editing levels (**Figure 10**). Of these, 26 sites, residing in 21 target genes, were statistically significant (**Table 2**). Sixteen positions appeared more edited in artery tibial than cerebellum while 10 appeared more edited in cerebellum than in artery tibial (**Table 2**). Sites showing higher differences in RNA editing levels belonged to well-characterized target genes such as *COG3* (Han et al., 2015; Peng et al., 2018; Silvestris et al., 2019), *IGFBP7* (Chen et al., 2017), *COPA* (Han et al., 2015; Peng et al., 2018), *FLNA* (Riedmann et al., 2008; Jain et al., 2018), and *ZNF358* (Zhang et al., 2016; Lee et al., 2017). The functional impact of RNA editing at these substrates is mostly unknown.

As an alternative to MW *U*-test, deregulated A-to-I editing has been identified using the statistical pipeline proposed by Tran et al. (2019) to detect dysregulated RNA editing in brains of autistic individuals. In this case, differential RNA editing sites are defined as positions having significantly different average editing levels between autistic donors and controls, or observed at significantly different population frequencies (Tran et al., 2019). Editing candidates are ranked by read coverage and the Wilcoxon rank-sum test is used if at least five samples in both control and donor groups have the required depth (Tran et al., 2019). By applying this pipeline to the above data, we found 10 differentially edited sites, eight of them already detected by the MW *U*-test (**Table 2**).

To date performance of statistical tests for differential RNA editing has never been tested and systematically assessed. Typically, the tests applied ignore the inherent noise introduced by the limited reads' coverage. Generally, tests assuming a normal distribution of RNA editing levels (such as the *t*-test) should be avoided. Indeed, accumulating evidence from large scale projects indicates that RNA editing levels seem to follow a beta distribution rather than a normal distribution (Picardi et al., 2015b). Further investigations are, in any case, needed to better understand the statistical properties of RNA editing levels.

TABLE 2 | Statistically significant differential recoding sites.

Chr:position	Gene	AA change	Δ editing	Pval (MW)	Padj (BH)
chr4:57976234*	IGFBP7	K95R	−0.417	0.000015	0.000016
chr13:46090371	COG3	I635V	−0.536	0.000016	0.000017
chr19:7585273	ZNF358	K382R	−0.381	0.000016	0.000017
chr1:160302244*	COPA	I164V	−0.416	0.000017	0.000018
chr4:57976286*	IGFBP7	R78G	−0.522	0.000019	0.000019
chrX:153579950*	FLNA	Q474R	−0.310	0.000027	0.000027
chr4:17805279	DCAF16	I162M	0.059	0.000088	0.000088
chr8:103841636	AZIN1	S367G	−0.081	0.000218	0.004633
chr20:36147572*	BLCAP	Y2C	0.090	0.000295	0.005015
chr20:36147563	BLCAP	Q5R	0.063	0.000178	0.005043
chr4:77979680	CCNI	R61G	−0.100	0.000132	0.005610
chr20:36147533	BLCAP	K15R	0.029	0.000464	0.005634
chr19:14593605	GIPC1	T62A	−0.237	0.000458	0.006488
chr12:133682596	ZNF140	Y142H	0.061	0.000998	0.010604
chr14:26917530	NOVA1	S363G	0.099	0.000128	0.010880
chr3:9876560	TTLL3	K419R	0.016	0.001442	0.011143
chr3:58141801*	FLNB	Q2103R	−0.228	0.001180	0.011144
chr5:156736808	CYFIP2	K124E	−0.007	0.001350	0.011475
chr15:75646086	NEIL1	K242R	0.182	0.001673	0.011850
chr4:77977164	CCNI	K123R	−0.009	0.002107	0.013777
chr1:12091858	MIIP	S355G	0.018	0.002867	0.017407
chr21:34922801*	SON	T422A	−0.146	0.003614	0.020479
chr3:58141791	FLNB	M2100V	−0.116	0.004256	0.022610
chr18:32825609	ZNF397	K314E	0.058	0.004582	0.022910
chr10:79397298	KCNMA1	S35G	−0.051	0.005179	0.024456
chr6:44120349*	TMEM63B	Q619R	−0.144	0.006899	0.030864

*Per each position we report the target gene and amino-acid change induced by RNA editing, the difference between mean editing levels of groups, the Mann–Whitney p-value, and the adjusted p-value by Benjamin–Hochberg. Positive Δ values indicate higher editing in cerebellum than artery, while negative Δs are associated to lower editing in cerebellum than artery. Positions marked by * are differentially edited by also the Tran et al. statistical pipeline.*

CONCLUSION

RNAseq is currently the technology of choice for large-scale studies of transcriptional and co-/post-transcriptional mechanisms. In the last few years, several computational tools have been developed to profile A-to-I editing in a variety of RNAseq data. Yet, RNA editing prediction is still not a fully solved bioinformatics task. However, noise and biases due to sequencing errors, read-mapping errors, and SNPs can be partly mitigated pre-processing reads and fine tuning program parameters depending on the selected algorithm.

The accurate detection of A-to-I editing is indispensable to systematically quantify RNA editing and facilitate comparative investigations across multiple samples. Similarly, A-to-I quantification metrics should be carefully selected. Indeed, measuring RNA editing activity across samples counting *de novo* detected sites or averaging over *de novo* sites leads to very noisy and confounding results. RNA editing is unevenly distributed across samples and different intrinsic (read quality, coverage, or depth) and extrinsic (mapping tool, read pre-processing, RNA editing calling software) factors affect the *de novo* detection that is far from being exhaustive. Averaging over millions of known sites from public databases can help but it requires estimated RNA editing levels that are dependent on a prefixed coverage cut-off that, in turn, drastically reduces the number of usable sites and leads to unreliable, often irreproducible, measures. The weighted average (or an index) over millions of known sites from public database, named here as the overall editing, is a much better solution. However, using this approach one has to rely on a specific set of sites from a given database, a set that might be continuously being modified. In contrast, the AEI is calculated over all tens of millions of genomic adenosines located within *Alu* sequences and accounts for the editing activity in low covered regions, while avoiding the need to quantify the editing level per-site. An index similar to AEI can be determined for recoding events. However, as the number of recoding sites is much lower, and the current set is known to be very noisy, the REI, while informative in some cases, should be used with care.

Identification of differential RNA editing is an important task. Although many studies have been employing various parametric and non-parametric approaches, further investigations are required. Given the non-normal distribution of RNA editing levels, and the strong (yet, usually ignored) effect of variable coverage, *ad hoc* models may be probably required to better perform this task.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: dbGAP accession phs000424.

AUTHOR CONTRIBUTIONS

CL and DS performed main bioinformatics analyses. SR carried out AEI computations. EE and GP supervised the work. AG and EP conceived the study and designed the analyses. EP drafted the manuscript. All authors approved the final version of the manuscript.

FUNDING

This work was supported by ELIXIR-IIB to GP, ISF grants 2673/17 and 1945/18 to EE, PRACE projects 2016163924 and 2018194670 to EP, AIRC (Associazione Italiana Ricerca sul Cancro) IG grant no. 22080 to AG, and Fondazione Mia Neri to AG. Funding for open access charge: EPITRAN COST initiative (CA16120).

ACKNOWLEDGMENTS

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424. This article is based upon work from COST Action EPITRAN (CA16120), supported by COST (European Cooperation in Science and Technology).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00194/full#supplementary-material

REFERENCES

Ahmed, S., Yamamoto, K., Sato, Y., Ogawa, T., Herrmann, A., Higashi, S., et al. (2003). Proteolytic processing of IGFBP-related protein-1 (TAF/angiomodulin/mac25) modulates its biological activity. *Biochem. Biophys. Res. Commun.* 310, 612–618. doi: 10.1016/j.bbrc.2003.09.058

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., et al. (2014a). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* 24, 365–376. doi: 10.1101/gr.164749.113

Bazak, L., Levanon, E. Y., and Eisenberg, E. (2014b). Genome-wide analysis of Alu editability. *Nucleic Acids Res.* 42, 6876–6884. doi: 10.1093/nar/gku414

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Brahmkhatri, V. P., Prasanna, C., and Atreya, H. S. (2015). Insulin-like growth factor system in cancer: novel targeted therapies. *BioMed. Res. Int.* 2015:538019. doi: 10.1155/2015/538019

Breen, M. S., Dobbyn, A., Li, Q., Roussos, P., Hoffman, G. E., Stahl, E., et al. (2019). Global landscape and genetic regulation of RNA editing in cortical samples from individuals with schizophrenia. *Nat. Neurosci.* 22, 1402–1412. doi: 10.1038/s41593-019-0463-7

Buuren, S. V., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Cenci, C., Barzotti, R., Galeano, F., Corbelli, S., Rota, R., Massimi, L., et al. (2008). Down-regulation of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation. *J. Biol. Chem.* 283, 7251–7260. doi: 10.1074/jbc.M708316200

Chen, L., Li, Y., Lin, C. H., Chan, T. H., Chow, R. K., Song, Y., et al. (2013). Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat. Med.* 19, 209–216. doi: 10.1038/nm.3043

Chen, Y.-B., Liao, X.-Y., Zhang, J.-B., Wang, F., Qin, H.-D., Zhang, L., et al. (2017). ADAR2 functions as a tumor suppressor via editing IGFBP7 in esophageal squamous cell carcinoma. *Int. J. Oncol.* 50, 622–630. doi: 10.3892/ijo.2016.3823

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Diroma, M. A., Ciaccia, L., Pesole, G., and Picardi, E. (2019). Elucidating the editome: bioinformatics approaches for RNA editing detection. *Brief. Bioinform.* 20, 436–447. doi: 10.1093/bib/bbx129

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Eisenberg, E. (2016). Proteome diversification by genomic parasites. *Genome Biol.* 17:17. doi: 10.1186/s13059-016-0875-6

Eisenberg, E., and Levanon, E. Y. (2018). A-to-I RNA editing – immune protector and transcriptome diversifier. *Nat. Rev. Genet.* 19, 473–490. doi: 10.1038/s41576-018-0006-1

Gallo, A. (2013). RNA editing enters the limelight in cancer. *Nat. Med.* 19, 130–131. doi: 10.1038/nm.3072

Gallo, A., and Locatelli, F. (2012). ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1. *Biol. Rev. Camb. Philos. Soc.* 87, 95–110. doi: 10.1111/j.1469-185X.2011.00186.x

Gal-Mark, N., Shallev, L., Sweetat, S., Barak, M., Billy Li, J., Levanon, E. Y., et al. (2017). Abnormalities in A-to-I RNA editing patterns in CNS injuries correlate with dynamic changes in cell type composition. *Sci. Rep.* 7:43421. doi: 10.1038/srep43421

Godfried Sie, C., Hesler, S., Maas, S., and Kuchka, M. (2012). IGFBP7's susceptibility to proteolysis is altered by A-to-I RNA editing of its transcript. *FEBS Lett.* 586, 2313–2317. doi: 10.1016/j.febslet.2012.06.037

Gott, J. M., and Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34, 499–531. doi: 10.1146/annurev.genet.34.1.499

Han, L., Diao, L., Yu, S., Xu, X., Li, J., Zhang, R., et al. (2015). The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell* 28, 515–528. doi: 10.1016/j.ccell.2015.08.013

Helm, M., and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.* 18, 275–291. doi: 10.1038/nrg.2016.169

Herculano-Houzel, S. (2010). Coordinated scaling of cortical and cerebellar numbers of neurons. *Front. Neuroanat.* 4:12. doi: 10.3389/fnana.2010.00012

Hu, J., Xu, J., Pang, L., Zhao, H., Li, F., Deng, Y., et al. (2016). Systematically characterizing dysfunctional long intergenic non-coding RNAs in multiple brain regions of major psychosis. *Oncotarget* 7, 71087–71098. doi: 10.18632/oncotarget.12122

Jain, M., Mann, T. D., Stulić, M., Rao, S. P., Kirsch, A., Pullirsch, D., et al. (2018). RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood pressure. *EMBO J.* 37:e94813. doi: 10.15252/embj.201694813

Josse, J., and Husson, F. (2016). missMDA: a package for handling missing values in multivariate data analysis. *J. Stat. Softw.* 70, 1–31. doi: 10.18637/jss.v070.i01

Khermesh, K., D'Erchia, A. M., Barak, M., Annese, A., Wachtel, C., Levanon, E. Y., et al. (2016). Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA* 22, 290–302. doi: 10.1261/rna.054627.115

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kiran, A. M., O'Mahony, J. J., Sanjeev, K., and Baranov, P. V. (2013). Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.* 41, D258–D261. doi: 10.1093/nar/gks961

Lee, S. H., Kim, H. P., Kang, J. K., Song, S. H., Han, S. W., and Kim, T. Y. (2017). Identification of diverse adenosine-to-inosine RNA editing subtypes in colorectal cancer. *Cancer Res. Treat.* 49, 1077–1087. doi: 10.4143/crt.2016.301

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* 22, 1001–1005. doi: 10.1038/nbt996

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Lin, C.-H., and Chen, S. C.-C. (2019). The cancer editome atlas: a resource for exploratory analysis of the adenosine-to-inosine RNA editome in cancer. *Cancer Res.* 79, 3001–3006. doi: 10.1158/0008-5472.CAN-18-3501

Lo Giudice, C., Tangaro, M. A., Pesole, G., and Picardi, E. (2020). Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat. Protoc.* doi: 10.1038/s41596-019-0279-7

Maas, S., Patt, S., Schrey, M., and Rich, A. (2001). Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14687–14692. doi: 10.1073/pnas.251531398

Mannion, N. M., Greenwood, S. M., Young, R., Cox, S., Brindle, J., Read, D., et al. (2014). The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Rep.* 9, 1482–1494. doi: 10.1016/j.celrep.2014.10.041

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Meyer, K. D., and Jaffrey, S. R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* 15, 313–326. doi: 10.1038/nrm3785

Neeman, Y., Levanon, E. Y., Jantsch, M. F., and Eisenberg, E. (2006). RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA* 12, 1802–1809. doi: 10.1261/rna.165106

Nishikura, K. (2016). A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.* 17, 83–96. doi: 10.1038/nrm.2015.4

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Paz, N., Levanon, E. Y., Amariglio, N., Heimberger, A. B., Ram, Z., Constantini, S., et al. (2007). Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res.* 17, 1586–1595. doi: 10.1101/gr.6493107

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H. T., Danan-Gotthold, M., Knisbacher, B. A., et al. (2015). Elevated RNA editing activity is a major contributor to transcriptomic diversity in tumors. *Cell Rep.* 13, 267–276. doi: 10.1016/j.celrep.2015.08.080

Peng, X., Xu, X., Wang, Y., Hawke, D. H., Yu, S., Han, L., et al. (2018). A-to-I RNA editing contributes to proteomic diversity in cancer. *Cancer Cell* 33, 817–828.e7. doi: 10.1016/j.ccell.2018.03.026

Picardi, E., D'Erchia, A. M., Gallo, A., and Pesole, G. (2015a). Detection of post-transcriptional RNA editing events. *Methods Mol. Biol.* 1269, 189–205. doi: 10.1007/978-1-4939-2291-8_12

Picardi, E., D'Erchia, A. M., Lo Giudice, C., and Pesole, G. (2016). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* 45, D750–D757. doi: 10.1093/nar/gkw767

Picardi, E., Manzari, C., Mastropasqua, F., Aiello, I., D'Erchia, A. M., and Pesole, G. (2015b). Profiling RNA editing in human tissues: towards the inosinome atlas. *Sci. Rep.* 5:14941. doi: 10.1038/srep14941

Picardi, E., and Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. *Bioinformatics* 29, 1813–1814. doi: 10.1093/bioinformatics/btt287

Piechotta, M., Wyler, E., Ohler, U., Landthaler, M., and Dieterich, C. (2017). JACUSA: site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics* 18:7. doi: 10.1186/s12859-016-1432-8

Pinto, Y., Cohen, H. Y., and Levanon, E. Y. (2014). Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol.* 15:R5. doi: 10.1186/gb-2014-15-1-r5

Porath, H. T., Carmi, S., and Levanon, E. Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* 5:4726. doi: 10.1038/ncomms5726

Qin, Y. R., Qiao, J. J., Chan, T. H., Zhu, Y. H., Li, F. F., Liu, H., et al. (2014). Adenosine-to-inosine RNA editing mediated by ADARs in esophageal squamous cell carcinoma. *Cancer Res.* 74, 840–851. doi: 10.1158/0008-5472.CAN-13-2545

Ramaswami, G., and Li, J. B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113. doi: 10.1093/nar/gkt996

Riedmann, E. M., Schopoff, S., Hartner, J. C., and Jantsch, M. F. (2008). Specificity of ADAR-mediated RNA editing in newly identified targets. *RNA* 14, 1110–1118. doi: 10.1261/rna.923308

Rosenthal, J. J., and Seeburg, P. H. (2012). A-to-I RNA editing: effects on proteins key to neural excitability. *Neuron* 74, 432–439. doi: 10.1016/j.neuron.2012.04.010

Roth, S. H., Danan-Gotthold, M., Ben-Izhak, M., Rechavi, G., Cohen, C. J., Louzoun, Y., et al. (2018). Increased RNA editing may provide a source for autoantigens in systemic lupus Erythematosus. *Cell Rep.* 23, 50–57. doi: 10.1016/j.celrep.2018.03.036

Roth, S. H., Levanon, E. Y., and Eisenberg, E. (2019). Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. *Nat. Methods* 16, 1131–1138. doi: 10.1038/s41592-019-0610-9

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Schwartz, S. (2016). Cracking the epitranscriptome. *RNA* 22, 169–174. doi: 10.1261/rna.054502.115

Shallev, L., Kopel, E., Feiglin, A., Leichner, G. S., Avni, D., Sidi, Y., et al. (2018). Decreased A-to-I RNA editing as a source of keratinocytes' dsRNA in psoriasis. *RNA* 24, 828–840. doi: 10.1261/rna.064659.117

Silvestris, D. A., Picardi, E., Cesarini, V., Fosso, B., Mangraviti, N., Massimi, L., et al. (2019). Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome Biol.* 20:33. doi: 10.1186/s13059-019-1647-x

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Stellos, K., Gatsiou, A., Stamatelopoulos, K., Perisic Matic, L., John, D., Lunella, F. F., et al. (2016). Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. *Nat. Med.* 22, 1140–1150. doi: 10.1038/nm.4172

Tajaddod, M., Jantsch, M. F., and Licht, K. (2016). The dynamic epitranscriptome: A to I editing modulates genetic information. *Chromosoma* 125, 51–63. doi: 10.1007/s00412-015-0526-9

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550, 249–254. doi: 10.1038/nature24041

Tran, S. S., Jun, H.-I., Bahn, J. H., Azghadi, A., Ramaswami, G., Van Nostrand, E. L., et al. (2019). Widespread RNA editing dysregulation in brains from autistic individuals. *Nat. Neurosci.* 22:25. doi: 10.1038/s41593-018-0287-x

Vlachogiannis, N. I., Gatsiou, A., Silvestris, D. A., Stamatelopoulos, K., Tektonidou, M. G., Gallo, A., et al. (2019). Increased adenosine-to-inosine RNA editing in rheumatoid arthritis. *J. Autoimmun* 106:102329. doi: 10.1016/j.jaut.2019.102329

Zhang, L., Yang, C.-S., Varelas, X., and Monti, S. (2016). Altered RNA editing in 3′ UTR perturbs microRNA-mediated regulation of oncogenes and tumor-suppressors. *Sci. Rep.* 6:23226. doi: 10.1038/srep23226

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 12 March 2020
doi: 10.3389/fgene.2020.00088

PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation

Bowen Song[1†], *Yujiao Tang*[1,2†], *Zhen Wei*[1,3], *Gang Liu*[4], *Jionglong Su*[4], *Jia Meng*[1,2] *and Kunqi Chen*[1,3*]

[1] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China, [2] Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom, [3] Institute of Ageing & Chronic Disease, University of Liverpool, Liverpool, United Kingdom, [4] Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China

OPEN ACCESS

Edited by:
Pavel V. Baranov,
University College Cork,
Ireland

Reviewed by:
Katsura Asano,
Kansas State University,
United States
Ivan V. Kulakovskiy,
Engelhardt Institute of Molecular
Biology (RAS), Russia
Claudio Lo Giudice,
National Research Council,
Italy

***Correspondence:**
Kunqi Chen
kunqi.chen@xjtlu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *11 November 2019*
Accepted: *27 January 2020*
Published: *12 March 2020*

Citation:
Song B, Tang Y, Wei Z, Liu G, Su J, Meng J and Chen K (2020) PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation. Front. Genet. 11:88. doi: 10.3389/fgene.2020.00088

Known as the "fifth RNA nucleotide", pseudouridine (Ψ or psi) is the first-discovered and most abundant RNA modification occurring at the Uridine site, and it plays a prominent role in a number of biological processes. Thousands of Ψ sites have been identified within different biological contexts thanks to the advancement in high-throughput sequencing technology; nevertheless, the transcriptome-wide distribution, biomolecular functions, regulatory mechanisms, and disease relevance of pseudouridylation are largely elusive. We report here a web server—PIANO—for **p**seudouridine site (Ψ) **i**dentification **a**nd fu**n**ctional ann**o**tation. PIANO was built upon a high-accuracy predictor that takes advantage of both conventional sequence features and 42 additional genomic features. When tested on six independent datasets generated from four independent Ψ-profiling technologies (Ψ-seq, RBS-seq, Pseudo-seq, and CeU-seq) as benchmarks, PIANO achieved an average AUC of 0.955 and 0.838 under the full transcript and mature mRNA models, respectively, marking a substantial improvement in accuracy compared to the existing *in silico* Ψ-site prediction methods, i.e., PPUS (0.713 and 0.707), iRNA-PseU (0.713 and 0.712), and PseUI (0.634 and 0.652). Besides, PIANO web server systematically annotates the predicted Ψ sites with post-transcriptional regulatory mechanisms (miRNA-targets, RBP-binding regions, and splicing sites) in its prediction report to help the users explore potential machinery of Ψ. Moreover, a concise query interface was also built for 4,303 known Ψ sites, which is currently the largest collection of experimentally validated human Ψ sites. The PIANO website is freely accessible at: http://piano.rnamd.com.

Keywords: pseudouridine sites, genome-derived feature, RNA modification, Web-server, functional annotation

INTRODUCTION

Pseudouridine (5-ribosyluracil, Ψ, and psi) is the first-discovered (Cohn and Volkin, 1951) and most abundant RNA modification occurring at the Uridine site catalyzed by 13 pseudouridine synthase (PUS) (Chen and Patton, 2000; Zhao et al., 2004; McCleverty et al., 2007; Shaheen et al., 2016; Jacob et al., 2017). Ψ is present in many classes of RNA within all organisms, such as messenger RNA (mRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), and ribosomal RNA (rRNA) (Ge and Yu, 2013). Ψ was termed as "the fifth nucleotide" with an estimated Ψ/U ratio of 7–9% (Jacob et al., 2017), and it is considered to be the most prevalent of the mRNA modifications (Meyer and Jaffrey, 2017). Ψ plays a prominent role in many biological processes. The presence of Ψ in tRNA and rRNA regulates the entry site binding process in ribosomal RNA (Jack et al., 2011) and RNA structure stabilization (Kierzek et al., 2014). A recent study also demonstrated that Ψ is related to transcript stability (Schwartz et al., 2014), environmental signal response (Carlile et al., 2014), and genetic code switching in mRNA (Karijolich and Yu, 2011; Fernández et al., 2013). Ψ deficiency may be associated with various diseases. It has been found that the dysregulation of Ψ modification of mitochondrial tRNA acts as an etiology of mitochondrial myopathy and sideroblastic anemia (MLASA) (Bykhovskaya et al., 2004). Furthermore, mutations in pseudouridine are also involved in diseases like lung cancer and duykeratosis congenita (Mei et al., 2012).

Several high-throughput sequencing approaches have been developed for profiling the transcriptome-wide distribution of Ψ, including Pseudo-seq (Carlile et al., 2014), Ψ-seq (Schwartz et al., 2014), PSI-seq (Lovejoy et al., 2014), and CeU-seq (Li X,et al., 2015). These approaches all share the same principle, in which RNA is treated with the N-cyclohexyl-N'-(2-morpholinoethyl)-carbodiimide-metho-*p*-toluenesulfonate (CMC) to leave a bulky group on Ψ and stop reverse transcription. Since the bulky adduct on the Ψ may reduce the sensitivity in the detection of Ψ, Vahid et al. recently developed a new approach, RBS-seq, which is based on a modification of RNA bisulfite sequencing and claims better sensitivity (Khoddami et al., 2019). Currently, the experiment-validated Ψ sites in human, mouse, and a few other model organisms are available from RMBase database (Xuan et al., 2017), and the regulation pathways of Ψ were more explicitly explained in MODOMICS database (Boccaletto et al., 2017).

Wet-lab approaches are surely effective for the study of transcriptome pseudouridylation with respect to a specific biological context; however, they are also laborious and offer only limited coverage, i.e., the reported RNA Ψ sites by wet-lab experiments are still restricted to the transcripts more readily expressed under a specific cell/tissue condition. Alternatively, computational efforts may provide a more cost-effective avenue (Chen X, et al., 2017). To date, many computational efforts have been made to facilitate the study of RNA epigenetics (Boccaletto et al., 2017; Chen X, et al., 2017; Chen Z, et al., 2019; Xue et al., 2020; Liu et al., 2020) in terms of both experimental data collection and site prediction works. For predictors related to the identification of Ψ RNA modification, PseUI (He et al., 2018), XG-PseU (Liu et al., 2019), and iRNA-PseU (Chen et al., 2016) allow for prediction of putative Ψ sites from an RNA sequence, and PPUS (Li Y.H, et al., 2015) can predict the Ψ sites regulated by a specific pseudouridine synthase. However, these three predictors are all based on sequence-derived features only without considering other genomic features (such as conservation, gene annotation, and miRNA binding) that may contribute to the prediction, and thus their performance is limited (Chen K, et al., 2019). Moreover, their prediction results are not functionally annotated with potential post-transcriptional regulation machineries that may explain the functional consequences of the predicted Ψ sites.

We present here a web server—PIANO—for **p**seudouridine site **i**dentification **a**nd fu**n**ctional an**n**otation. Inspired by the WHISTLE framework (Chen K, et al., 2019), PIANO took advantage of both the conventional sequence features and 42 additional genomic features. Using six independent datasets generated from four different technologies, we showed that PIANO adds a marked improvement to the accuracy of existing Ψ-site prediction. Moreover, the PIANO web server accepts both genomic location and RNA sequence format as input file when making predictions, and the putative Ψ sites returned are also annotated with various post-transcriptional regulations, including miRNA-targets, RBP-binding regions, and splicing sites, to unveil potential functional mechanisms of Ψ. The PIANO website is freely accessible at: http://piano.rnamd.com.

MATERIALS AND METHODS

Training and Testing Data for Ψ-Site Prediction

To construct the Ψ-site prediction model, we used the known human Ψ sites detected from four different base-resolution Ψ profiling techniques, including Ψ-Seq, RBS-Seq, CeU-Seq, and Pseudo-Seq (see **Table 1**). The Ψ sites at base-resolution were directly downloaded from Gene Expression Omnibus (GEO).

In the beginning of the performance evaluation, dataset H1 (see **Table 1**) was used as the testing data, while dataset H2-H4 were used as for training. Specifically, the base-resolution Ψ sites in training datasets were used as the positive training data. The negative sites used in model training were randomly selected from unmodified U sites located on the same transcripts of positive sites (see **Figure 1**). To make the best use of the limited volume of positive data, we randomly selected 10 negative sites for

TABLE 1 | Base-resolution dataset used for Ψ-site prediction.

Dataset	Cell line	Treatment	Technique	Site #	Source
H1	HEK293		Ψ-Seq	652	(Schwartz et al., 2014)
H2	Hela		RBS-Seq	322	(Khoddami et al., 2019)
H3	HEK293T		CeU-Seq	1555	(Li X, et al., 2015)
H4	HEK293T	H_2O_2		460	
H5	HEK293T	Heat Shock (HS)		421	
H6	Hela		Pseudo-Seq	156	(Carlile et al., 2014)

The experimentally validated human Ψ sites used in this project are also available from the PIANO website of this project (http://piano.rnamd.com), annotated with various post-transcriptional regulations.

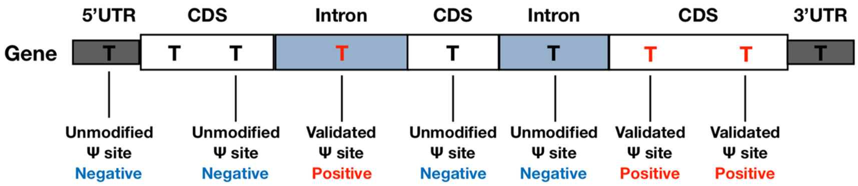

5'UTR
CDS
Intron
CDS
Intron
CDS
3'UTR
Gene
T
T
T
T
T
T
T
T
T
Unmodified Ψ site Negative
Unmodified Ψ site Negative
Validated Ψ site Positive
Unmodified Ψ site Negative
Unmodified Ψ site Negative
Validated Ψ site Positive
Validated Ψ site Positive

FIGURE 1 | Negative and Positive Data. Negative sites were randomly selected from un-modified U sites located on the same transcripts of the positive sites.

each of the positive sites. To balance the positive-to-negative ratio, the negative sites were then randomly split into 10 subsets, and 10 separate predictors were generated with a 1:1 positive-to-negative ratio. The negative sites of testing data were generated following the same procedure. Consequently, 10 separate predictors were generated, and their prediction results were averaged.

Following the experimental design of WHISTLE framework (Chen K, et al., 2019), we performed dataset level leave-one-out validation over the H1-H5 base-resolution datasets; four samples from H1–H5 were used as training, while the other was used for testing. Subsequently, the sites from the datasets H1–H5 (generated from Ψ-Seq, RBS-Seq, and CeU-Seq) were used to establish a predictor, whose performance was evaluated on the dataset H6, which was generated from an independent technology (Pseudo-Seq).

Features Used for Ψ-Site Prediction

Sequence-Derived Features

The length of 41bp was widely used to extracted sequence information in many previous studies, which was determined as a suitable flanking window by relevant tests, i.e., iRNA-m7G (Liu et al., 2019), iRNA-2OM (Yang et al., 2018), and MethyRNA (Chen W, et al., 2017). Consequently, the sequence-derived information of 41 bp flanking window of Ψ and non-Ψ (U) sites as central was generated using the chemical properties of nucleotides, position-specific nucleotide propensity (PSNP), and cluster information.

In the first encoding method, the nucleotides are classified into three categories based on three distinct structural chemical properties. Ring structures of nucleotides are the first to be considered; here, adenosine and guanosine have two rings, while cytidine and uridine only have one ring. In addition, the guanosine and cytidine have stronger hydrogen bonding than adenosine and uridine. Furthermore, adenosine and cytidine can be classified as the amino group, while guanosine and uridine contain the keto group. Based on these chemical properties defined above, the i -th nucleotide from sequence S may be encoded by a vector $S_i = (x_i, y_i, z_i)$:

$$x_i = \begin{cases} 1 \text{ if } s_i \in \{A,G\} \\ 0 \text{ if } s_i \in \{C,U\} \end{cases}, \; y_i = \begin{cases} 1 \text{ if } s_i \in \{A,C\} \\ 0 \text{ if } s_i \in \{G,U\} \end{cases}, \; z_i = \begin{cases} 1 \text{ if } s_i \in \{A,U\} \\ 0 \text{ if } s_i \in \{C, G\} \end{cases} \quad (1)$$

Thus, the A, C, G, and U may be encoded as a vector (1,1,1), (0,1,0), (1,0,0), and (0,0,1), respectively.

The position-specific nucleotide propensity (PSNP) stands for the differences of the frequency of nucleotides calculated in specific locations between RNA sequences of positive and negative data. The frequency of occurrence of A, U, G, and C in the i -position were calculated for both positive and negative data, respectively, to obtain two matrices with 4×41 dimension as Z_{plus} and Z_{minus}, where Z_{plus} was extracted from sequence of all positive data, and Z_{minus} was extracted from sequence of all negative data. The position-specific nucleotide propensity (PSNP) matrices was defined as Z_{PSNP}:

$$Z_{PSNP} = Z_{plus} = Z_{minus} \quad (2)$$

For the cluster information, the average relative position of the closest k *(k=1,2 and 3)* nucleotide to center Ψ/non-Ψ was calculated for each nucleic acid (A, G, C, and U). The k was considered as 1 to 3. Using sequence 'AGCUAGCCAUC CUACGGUACAGCAU' as an example, the center U is at the ninth positive. For encoding the cluster information of adenine, the average relative position of the closest 1 (k=1) adenine to center U is 1 (1/1); when k equals to 2, the relative position of the second closest adenine to center U is 4, and, therefore, the average relative position of the closest 2 (k=2) adenine to center U is 2.5 (5/2) and 3.7 (11/3) when k equals to 3. Similarly, the cluster information of guanosine in this example sequence is 3 (3/1), 3.5(7/2), and 4.7(14/3) when k equals to 1, 2, and 3, respectively.

The sequence-derived encoding methods employed by the three previously published predictors were used to reproduce the PPUS, iRNA-PseU, and PseUI with the same training data of PIANO, respectively, and their performances were compared with PIANO using independent datasets.

Genome-Derived Features

In the original WHISTE approach, 35 additional genomic features that might contribute to the prediction of m^6A RNA methylation sites were considered (Chen K, et al., 2019). In PIANO, seven new genomic features were added to the prediction model, the details of the 42 genomic features considered in the prediction were summarized in **Supplementary Table S1**. Specifically, genomic Features 1–16 are dummy variable features indicating whether the uridine sites shall fall within the transcript regions that satisfy certain topological properties. All the features in this category are generated by the GenomicFeatures R/Bioconductor package using the transcript annotations hg19 TxDb package (Lawrence et al., 2013). To remove the ambiguity caused by transcript isoforms, only the primary (longest) transcripts of each gene were kept for the extraction of the transcript subregions. The longest transcript isoform was used to unambiguously assign m6A peak regions to mRNAs (Ke et al., 2017) and contributed to a better performance in accuracy compared with using the average value of multiple transcripts. Genomic Features 17–20 are real valued features defining the relative position of the transcript regions (3'UTR, 5'UTR, CDS, and whole transcript), i.e., the distance from the adenine to the 5' end divided by the width of the region. The values are also set to zero for sites that do not belong to the region. Genomic features 21–25 represent the length of the transcript region containing the modification site. The values are also set to zero for sites that do not belong to the region. Features 26–27 captured the distance from the adenine sites to the 5'end or 3'end of the splicing junctions. Additionally, the distance to the nearest neighboring ψ sites in the training data is generated to measure the clustering effect of the ψ RNA modification sites. Evolutionary conservation score of the uridine sites and its flanking regions are measured by PhastCons (Siepel et al., 2005) score, and the fitness consequence (Gulko et al., 2015) scores were presented in features 28–31. To consider the RNA secondary structures around the uridine site, the RNA secondary structures are predicted using RNAfold from the Vienna RNA package (Lorenz et al., 2011) and presented in features 32–33. Genomic properties of transcripts containing the Ψ sites were presented in features 34–38. Finally, features 39–42 represent omics information, such as microRNA target sites (Chou et al., 2017) and HNRNPC binding sites (2012).

Machine Learning Approach Used for Ψ-Site Prediction

As a high-efficiency machine learning algorithm in computational biology, the SVM (Support Vector Machine) has been widely applied in microRNA target prediction (Liu et al., 2010), protein phosphorylation prediction (Wong et al., 2007), and m^6A RNA methylation site prediction (Chen W, et al., 2017). In this project, the R language interface of LIBSVM (Chang and Lin, 2011) was used to build our model with the radial basis function as kernel, and the other parameters were set at the default.

Performance Evaluation of Ψ-Site Prediction

To evaluate the performance of PIANO, a 5-fold cross-validation was employed on training datasets using the SVM classifier, and the independent testing dataset was used to measure the final performance of PIANO. There is no overlap between the training sites and testing sites, as only the Ψ sites not previously used as training data were considered during performance evaluation; the performance evaluation result should thus directly reflect the capability of the algorithm to identify previously unknown Ψ sites. To evaluate the performance, the ROC (receiver operating characteristic) curve (sensitivity against 1-specificity) was used, and the area under ROC curve (AUROC) was calculated as the main performance evaluation metric.

Estimate the Probability of Ψ

The likelihood ratio (LR) of a Ψ site is calculated to estimate the probability of Ψ RNA methylation:

$$LR = \frac{P(observation|\Psi)}{P(observation|\mathrm{U})} \tag{3}$$

In the PIANO web server, a site was predicted to be a putative Ψ site if its predictive value was above 0.5 with a minimum LR value of 1. A site with a larger LR value suggests that it is more likely to be a Ψ site. The machine learning classifiers usually obtain the lowest empirical rate with the value of 0.5 as cutoff. The statistical significance of LR is assessed by an upper bound of the p-value, indicating how extreme the observed LR is among all the transcriptome U sites. It is calculated from the relative ranking of the putative Ψ sites among all the transcriptome U sites, i.e., if only 0.1% of U sites have a LR score larger than a specific U site, then the upper bound of the p-value of this site is 0.001. In the report of PIANO web server, a putative Ψ site is considered to be of high confidence if its LR within the top 0.5% of all transcriptome Us (corresponding to an upper bound of the p-value < 0.005) of all the transcriptome U sites, followed by medium confidence (0.005 < upper bound of the p-value ≤ 0.05) and low confidence (p-value > 0.05).

Functional Annotation of Putative Ψ Site

The gene symbol, Ensembl gene ID, gene region, and gene type for each putative Ψ site were annotated using ANNOVAR package (Wang et al., 2010). Furthermore, we annotated the putative Ψ sites with three kinds of post-transcriptional regulation, including RNA-binding proteins (RBPs) regions, miRNA-RNA targets, and splicing sites. We first found the intersection between the computational predicted Ψ sites and POSTAR2-derived RBP binding regions (Zhu et al., 2018). For miRNA targets, we obtained the information from miRanda (Agarwal et al., 2015) and starBase2 (Li et al., 2013), and we found the Ψ sites within the miRNA targets regions to explore the potential influence of Ψ on miRNA-target interactions. Finally, we obtained the Canonical splice sites (GT-AG) from UCSC (Lawrence et al., 2013) annotations, 100 bp upstream region from 5' splicing sites and 100 bp downstream region from 3' splicing sites were extracted for the subsequent analysis of Ψ sites on splicing sites. The detailed information of the post-transcriptional regulation association analysis can be found in **Supplementary Table S2**.

RESULTS

Although the genome-derived features alone are already very effective for predicting Ψ sites, the best performance was achieved when the sequence features and genomic features were combined. Consequently, our PIANO predictor was established based on both the genome-derived features and sequence-derived features. When designing the encoding methods for sequence features used for the PIANO approach, the chemical properties of nucleotides, position-specific nucleotide propensity (PSNP), and cluster information were considered. We found that this combination (sequence and genomic features) achieved the best performance in accuracy compared with combining genome-derived features with other basic sequence encoding methods (i.e., one-hot encoding method).

The performance of the predictor was evaluated under two modes. For the full transcript mode, the positive and negative Ψ sites located in both exonic and intronic regions are all considered to construct the predictor. In the mature mRNA mode, only positive and negative Ψ sites located on mature mRNA transcripts are considered; this is because existing experimental datasets overwhelmingly relied on polyA selection in RNA-seq library preparation, and intronic Ψ sites are likely to be underrepresented in the data, which may lead to an over-estimation of accuracy under the full transcript mode.

To avoid potential over-fitting and to identify the most significant subset of genomic features, feature selection was implemented; the datasets H2–H5 were used as training data, while dataset H1 was used for the independent testing data. The relative importance of each genome-derived feature were measured by the Perturb method (Gevrey et al., 2003). According to the rank of importance, the top N most important features were reserved in the prediction and were evaluated with a 5-fold cross-validation. For the predictor under full transcript model, the top 17 genomic features led to the best predictor performance, with fitCons scores, exons containing stop codons, and number of exons as the top three most important genomic features for prediction. Similarly, the top 20 genome-derived features were selected under the mature mRNA model. The length of the mature transcript plays the most important role under this model, and the exons containing stop codons and an miRNA target won the second and third significance. Consequently, to obtain the most robust performance, only the top 17 and 20 genomic features were used under full transcript model and mature mRNA model for Ψ site prediction, respectively. Please see **Supplementary Figure S1** for more details.

We showed that the newly developed method PIANO substantially outperformed competing approaches on cross-validation (**Supplementary Table S3**) when tested on independent datasets (**Supplementary Table S3**) or benchmarked by an independent technique (**Supplementary Table S4**). To sum up, by testing independent datasets generated from four different Ψ profiling technologies (Ψ-seq, RBS-seq, Pseudo-seq, and CeU-seq), the newly developed method PIANO achieved an average AUC of 0.955 and 0.838 under full transcript and mature mRNA modes, respectively (see **Table 2**), representing a marked improvement compared to PPUS (0.713 and 0.707), iRNA-PseU (0.713 and 0.712), and PseUI (0.634 and 0.652).

TABLE 2 | Performance evaluation of Ψ-site predictors.

Mode	Method	Benchmarking data (AUC)				Average AUC
		Ψ-Seq	RBS-Seq	CeU-Seq	Pseudo-Seq	
Full transcript	PIANO	0.957	0.978	0.914	0.972	0.955
	iRNA-PseU	0.679	0.727	0.721	0.708	0.713
	PPUS	0.700	0.721	0.724	0.705	0.713
	PseUI	0.631	0.710	0.610	0.585	0.634
Mature mRNA	PIANO	0.859	0.770	0.864	0.857	0.838
	iRNA-PseU	0.753	0.582	0.760	0.751	0.712
	PPUS	0.749	0.575	0.757	0.748	0.707
	PseUI	0.666	0.651	0.652	0.639	0.652

*The table presents the performance of different Ψ site predictors achieved on independent human datasets with different technologies as a benchmark, and it is summarized from **Supplementary Table S3** and **S4**. Only the Ψ sites not previously used as training data were considered during performance evaluation, so the training sites and testing sites did not overlap. Because existing datasets overwhelmingly relied on polyA selection in RNA library preparation and intronic Ψ sites are likely to be underrepresented in the data, the performances were evaluated under two modes: full transcript and mature mRNA modes. In the mature mRNA mode, only positive and negative Ψ sites located on mature mRNA transcripts are considered, as previously described (Chen K,et al., 2019). Our new approach PIANO substantially outperformed competing approaches in accuracy.*

The performance of the purposed predictor was further evaluated by separating the training and testing datasets between the cell type in which datasets H3–H5 generated from HEK293T were used for training, while datasets H2 and H6 from Hela were used for independent testing. Consistent with previous validation results, our method PIANO achieved a marked improvement in prediction accuracy compared with existing predictors, using the AUROC (area under ROC curve) and AUPRC (area under precision-recall curve) as an evaluation metric, when tested on independent dataset with a 1:1 positive to negative ratio (**Supplementary Table S5**) and 1:10 positive to negative ratio (**Supplementary Table S6**), respectively, suggesting the reliability of our newly proposed approach. Besides, the comparison between different algorithms indicated that SVM (Support Vector Machine) was a quite effective machine learning approach and achieved the best performance in our study (**Supplementary Table S5**). In addition, to further evaluate different approaches, we also considered the prediction of PUS-specific Ψ sites. In this experiment, TruB1, PSU7, and TruB2 were considered, and the goal was to predict their specific substrates (Safra et al., 2017). Consistent with previous results in Ψ-site prediction, the PIANO method again substantially outperformed competing approaches under both the full transcript and mature mRNA model (**Table 3**), suggesting the effectiveness of the approach.

Construction of the PIANO Website

A website PIANO, which stands for **p**seudouridine site **i**dentification **a**nd fu**n**ctional ann**o**tation, was built for the convenience of academic users. Hyper Text Markup Language

TABLE 3 | PUS-specific substrate prediction.

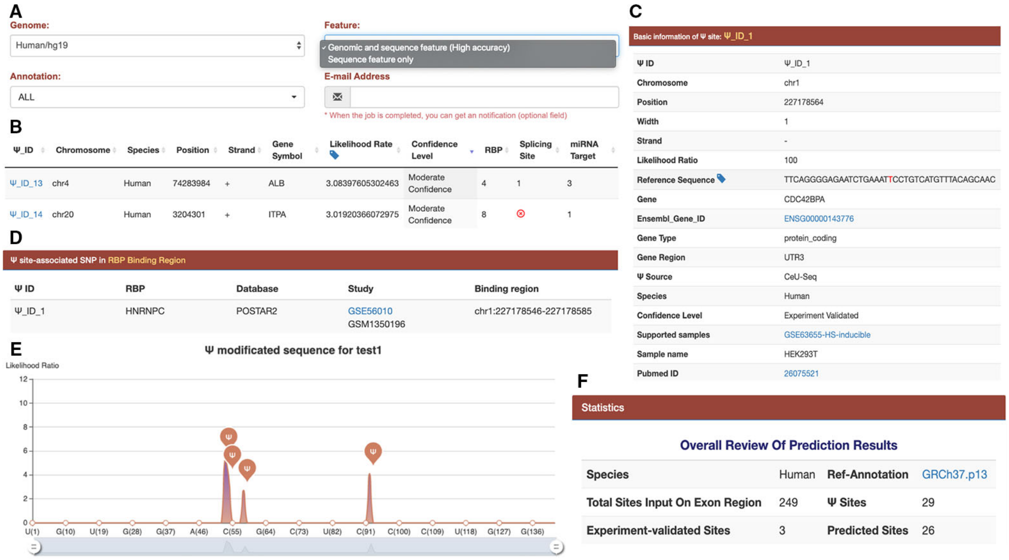
Method	Full transcript model			Mature mRNA model		
	TruB2	PSU7	TruB1	TruB2	PSU7	TruB1
PIANO	0.981	0.966	0.973	0.837	0.960	0.910
iRNA-PseU	0.812	0.829	0.838	0.719	0.812	0.731
PPUS	0.806	0.824	0.824	0.733	0.816	0.739
PseUI	0.853	0.870	0.840	0.805	0.861	0.786

(HTML), Cascading Style Sheets (CSS), and Hypertext Preprocessor (PHP) were used to construct the PIANO web interface. This included a database containing 4,303 experimentally validated Ψ sites reported from four different high-throughput Ψ profiling techniques, which is so far the most complete collection of Ψ in humans. Among those experimentally validated Ψ sites, we found Ψ was distributed most often along coding DNA sequence and 3'UTR, but it was relatively rare in 5'UTR (**Supplementary Figure S2**). Secondly, a web server for putative Ψ-site identification from the user-defined genomic ranges or provided FASTA sequences (detailed in **Figure 2**) was used. The help document of the PIANO web server is provided in the **Supplementary Materials**. Both experimentally validated Ψ sites and the predicted putative Ψ sites are functionally annotated with various post-transcriptional regulations to unveil potential functional mechanism concerning Ψ. The data and prediction results may be conveniently downloaded and visualized with web browser. The PIANO website is freely accessible from: http://piano.rnamd.com.

CONCLUSION

With recent advancements that unveiled various biomolecular functions of Ψ under different biological contexts, Ψ starts to capture broader interests of the scientific community (Schwartz et al., 2014; Carlile et al., 2014; Li X, et al., 2015; Karijolich et al., 2015; Dominissini et al., 2016; Penzo et al., 2017; Guzzi et al., 2018; Adachi et al., 2018; Shaheen et al., 2019). To date, a number of high-throughput approaches have been developed for profiling the transcriptome-wide distribution of Ψ (Adachi et al., 2019), including Pseudo-seq (Carlile et al., 2014), Ψ-seq (Schwartz et al., 2014), PSI-seq (Lovejoy et al., 2014), CeU-seq (Li X, et al., 2015), and RBS-seq (Khoddami et al., 2019). These technologies all reported the widespread occurrence of Ψ on mRNA and lncRNA in human cells. Four Ψ site predictors have been built, including PseUI (He et al., 2018), XG-PseU (Liu et al., 2019), iRNA-PseU (Chen et al., 2016), and PPUS (Li Y.H, et al., 2015); however, all of them are based on sequence-derived features only without considering other genomic features that may contribute to the prediction and thus limited their performance.

Here, by integrating 42 genomic features together with conventional sequence-derived features, we have developed the (so far) most accurate Ψ-site predictor. Our new method (PIANO) substantially outperformed competing approaches when using four different Ψ profiling protocols as the benchmarks (with 0.24 and 0.12 improvement in terms of AUC under full transcript and mature mRNA modes, respectively) and supports functional annotation for the putative Ψ sites. A web site—PIANO—was also developed, including (1) a database hosting currently the largest collection of 4,303 experimentally validated human Ψ sites; and (2) a web



Genome: Human/hg19

Annotation: ALL

Feature:
✓ Genomic and sequence feature (High accuracy)
Sequence feature only

E-mail Address

* When the job is completed, you can get an notification (optional field)

Ψ_ID	Chromosome	Species	Position	Strand	Gene Symbol	Likelihood Rate	Confidence Level	RBP	Splicing Site	miRNA Target
Ψ_ID_13	chr4	Human	74283984	+	ALB	3.08397605302463	Moderate Confidence	4	1	3
Ψ_ID_14	chr20	Human	3204301	+	ITPA	3.01920366072975	Moderate Confidence	8	⊗	1

Basic information of Ψ site: Ψ_ID_1

Ψ ID	Ψ_ID_1
Chromosome	chr1
Position	227178564
Width	1
Strand	-
Likelihood Ratio	100
Reference Sequence	TTCAGGGGAGAATCTGAAATTCCTGTCATGTTTACAGCAAC
Gene	CDC42BPA
Ensembl_Gene_ID	ENSG00000143776
Gene Type	protein_coding
Gene Region	UTR3
Ψ Source	CeU-Seq
Species	Human
Confidence Level	Experiment Validated
Supported samples	GSE63655-HS-inducible
Sample name	HEK293T
Pubmed ID	26075521

Ψ site-associated SNP in RBP Binding Region

Ψ ID	RBP	Database	Study	Binding region
Ψ_ID_1	HNRNPC	POSTAR2	GSE56010 GSM1350196	chr1:227178546-227178585

Ψ modificated sequence for test1

Statistics

Overall Review Of Prediction Results

Species	Human	Ref-Annotation	GRCh37.p13
Total Sites Input On Exon Region	249	Ψ Sites	29
Experiment-validated Sites	3	Predicted Sites	26

FIGURE 2 | Interface and output of the PIANO web server for Ψ-site prediction and functional annotation. **(A)** When predicting human Ψ sites, the PIANO web server supports two types of input: the genomic ranges of human genome assembly and the FASTA sequences. As the prediction process may take quite some time, it is highly recommended that the user should provide an email address, where an email notification will be sent when the job is finished. **(B)** The basic information of each putative Ψ site, such as gene symbol, likelihood ratio, confidence level, and the number of related post-transcriptions associated with the putative site. **(C)** The source and detailed information of each putative Ψ site. If the input file contains any experimental validated Ψ sites collected in PIANO, the sites will be annotated with additional information. **(D)** The details of the site-relevant RBP information. **(E)** A graph to visualize the position of predicted Ψ sites on a user-provided FASTA sequence. **(F)** An overall review of the prediction result.

server enabling the prediction of novel Ψ sites from given genomic ranges or FASTA sequences. Users may query and download their predicted results with clear and simple instructions (see **Supplementary Materials**). The scripts used to generate genomic and sequence features considered in PIANO's framework, the training and testing data, and datasets related to the construction of the PIANO database were provided in the download page of PIANO website. In conclusion, our work will serve as a useful resource for researchers who are interested in Ψ and its role concerning various post-transcriptional regulations.

Nevertheless, it is worth noting that there exist significant discrepancies in the Ψ sites reported by different technologies (Zaringhalam and Papavasiliou, 2016; Adachi et al., 2018). Although the discrepancy may be due to the context-specificity of pseudouridylation and technology preferences, our PIANO predictor achieved reasonable consensus with all the four high-throughput profiling Ψ techniques; Ψ is, however, considered as the most prevalent mRNA modifications (Meyer and Jaffrey, 2017) with an estimated Ψ/U ratio of 7–9% (Jacob et al., 2017). Currently, only a small number of Ψ sites have been reported; we are therefore not able to calculate a reasonable number for the real-life estimate of class imbalance. This may due to the limited detection power of existing experimental approaches. With an estimated real-life Ψ/U ratio as 8%, we can expect at least 10 times the number of negative sites. Under this assumption, we tested the stability of our method by assigning 1:10 and 1:1 positive-to-negative ratio for the training and testing data. The result showed that the performance generated by the 1:10 class were more stable than the 1:1 class (**Supplementary Figure S3**). We further calculated the value of FDR, FPR, and TPR in this setting, using different LRs as cutoff (**Supplementary Table S7**). To sum up, we cannot rule out the possibility of experimental bias, and the training data (gold standard data) may be further optimized in the future as more experimental evidence is accumulated. To make the PIANO method more practically useful, the predictor should be used by combining with other experimental evidence and knowledge, e.g., the Us within a binding site of PUS. The performance of PIANO method is much better than all existing approaches, and it can provide the most reliable putative Ψ sites for users.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE60047, GSE58200, GSE63655, GSE90963.

AUTHOR CONTRIBUTIONS

KC, JM, GL, and JS initialized the project. KC and BS designed the research plan. ZW constructed the genomic features considered in human Ψ site prediction. BS performed the development of the Ψ site web server. YT and BS built the website. BS and KC drafted the manuscript. All authors read, critically revised, and approved the final manuscript.

FUNDING

This work has been supported by the National Natural Science Foundation of China [31671373]; XJTLU Key Program Special Fund [KSF-T-01].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00088/full#supplementary-material

REFERENCES

(2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74.

Adachi, H., De Zoysa, M. D., and Yu, Y.-T. (2018). Post-transcriptional pseudouridylation in mRNA as well as in some major types of noncoding RNAs. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech*. 1862, 230–239.

Adachi, H., DeZoysa, M. D., and Yu, Y.-T. (2019). "Detection and Quantification of Pseudouridine in RNA," in *Epitranscriptomics* (Springer), 219–235.

Agarwal, V., Bell, G. W., Nam, J. -W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4, e05005.

Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2017). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307.

Bykhovskaya, Y., Casas, K., Mengesha, E., Inbal, A., and Fischel-Ghodsian, N. (2004). Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *Am. J. Hum. Genet.* 74 (6), 1303–1308. doi: 10.1086/421530

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 515 (7525), 143–146.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2 (3), 1–27. doi: 10.1145/1961189.1961199

Chen, J., and Patton, J. R. (2000). Pseudouridine synthase 3 from mouse modifies the anticodon loop of tRNA. *Biochemistry* 39 (41), 12723–12730. doi: 10.1021/bi001109m

Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K. C.. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* 5, e332.

Chen, X., Sun, Y. Z., Liu, H., Zhang, L., Li, J. Q., and Meng, J. (2017). RNA methylation and diseases: experimental results, databases, Web servers and computational models. *Brief Bioinform.* 20, 896–917.

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dynamics* 35 (3), 683–687. doi: 10.1080/07391102.2016.1157761

Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A. I., Webb, G. I., Akutsu, T., et al. (2019). Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Briefings Bioinf*. bbz112 doi: 10.1093/bib/bbz112

Chen, K., Wu, Q., Zhang, Z., Wei, R., Rong, Z., Lu, J., Meng, J. P., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res*. 47, e41 doi: 10.1093/nar/gkz074

Chou, C. -H., Shrestha, S., Yang, C. -D., Chang, N. -W., Lin, Y. -L., Liao, K.-W., et al. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46 (D1), D296–D302.

Cohn, W. E., and Volkin, E. (1951). Nucleoside-5′-Phosphates from Ribonucleic Acid. *Nature* 167 (4247), 483–484.

Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., et al. (2016). The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature* 530 (7591), 441–446.

Fernández, I. S., Ng, C. L., Kelley, A. C., Wu, G., Yu, Y. -T., and Ramakrishnan, V. (2013). Unusual base pairing during the decoding of a stop codon by the ribosome. *Nature* 500 (7460), 107.

Ge, J., and Yu, Y. T. (2013). RNA pseudouridylation: new insights into an old modification. *Trends Biochem. Sci.* 38 (4), 210–218. doi: 10.1016/j.tibs.2013.01.002

Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* 160 (3), 249–264. doi: 10.1016/S0304-3800(02)00257-0

Gulko, B., Hubisz, M. J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* 47 (3), 276–283. doi: 10.1038/ng.3196

Guzzi, N., Cieśla, M., Ngoc, P. C. T., Lang, S., Arora, S., Dimitriou, M., et al. (2018). Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. *Cell* 173 (5), 1204–1216 e26. doi: 10.1016/j.cell.2018.03.008

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

Jack, K., Bellodi, C., Landry, D. M., Niederer, R. O., Meskauskas, A., Musalgaonkar, S., et al. (2011). rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. *Mol. Cell* 44 (4), 660–666. doi: 10.1016/j.molcel.2011.09.017

Jacob, R., Zander, S., and Gutschner, T. (2017). The dark side of the epitranscriptome: chemical modifications in long non-coding RNAs. *Int. J. Mol. Sci.* 18 (11), 2387. doi: 10.3390/ijms18112387

Karijolich, J., and Yu, Y. T. (2011). Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature* 474 (7351), 395–398.

Karijolich, J., Yi, C., and Yu, Y.-T. (2015). Transcriptome-wide dynamics of RNA pseudouridylation. *Nat. Rev. Mol. Cell Biol.* 16 (10), 581. doi: 10.1038/nrm4040

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vagbo, C. B., Geula, S., et al. (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 (10), 990–1006. doi: 10.1101/gad.301036.117

Khoddami, V., Yerra, A., Mosbruger, T. L., Fleming, A. M., Burrows, C. J., and Cairns, B. R. (2019). Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U. S. A.* 116 (14), 6784–6789. doi: 10.1073/pnas.1817334116

Kierzek, E., Malgowska, M., Lisowiec, J., Turner, D. H., Gdaniec, Z., and Kierzek, R. (2014). The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res.* 42 (5), 3492–3501. doi: 10.1093/nar/gkt1330

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PloS Comput. Biol.* 9 (8), e1003118. doi: 10.1371/journal.pcbi.1003118

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42 (D1), D92–D97.

Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F., et al. (2015). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* 11 (8), 592–597. doi: 10.1038/nchembio.1836

Li, Y. H., Zhang, G., and Cui, Q. (2015). PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics* 31 (20), 3362–3364. doi: 10.1093/bioinformatics/btv366

Liu, H., Yue, D., Chen, Y., Gao, S. J., and Huang, Y. (2010). Improving performance of mammalian microRNA target prediction. *BMC Bioinf.* 11, 476. doi: 10.1186/1471-2105-11-476

Liu, K., Chen, W., and Lin, H. (2019). XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. *Mol. Genet. Genomics.* 295, 13–21.

Liu, L., Lei, X., Meng, J., and Wei, Z. (2020). WITMSG: Large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Curr. Genomics*. 21, 67–76.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6 (1), 26.

Lovejoy, A. F., Riordan, D. P., and Brown, P. O. (2014). Transcriptome-wide mapping of pseudouridines: Pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PloS One* 9 (10), e110799. doi: 10.1371/journal.pone.0110799

McCleverty, C. J., Hornsby, M., Spraggon, G., and Kreusch, A. (2007). Crystal structure of human Pus10, a novel pseudouridine synthase. *J. Mol. Biol.* 373 (5), 1243–1254. doi: 10.1016/j.jmb.2007.08.053

Mei, Y. P., Liao, J. P., Shen, J., Yu, L., Liu, B. L., Liu, L., et al. (2012). Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* 31 (22), 2794–2804. doi: 10.1038/onc.2011.449

Meyer, K. D., and Jaffrey, S. R. (2017). Rethinking m(6)A readers, writers, and erasers. *Annu. Rev. Cell Dev. Biol.* 33, 319–342. doi: 10.1146/annurev-cellbio-100616-060758

Penzo, M., Guerrieri, A., Zacchini, F., Treré, D., and Montanaro, L. (2017). RNA Pseudouridylation in physiology and medicine: for better and for worse. *Genes* 8 (11), 301.

Safra, M., Nir, R., Farouq, D., Vainberg Slutskin, I., and Schwartz, S. (2017). TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res.* 27 (3), 393–406. doi: 10.1101/gr.207613.116

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., Leon-Ricardo, B. X., et al. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* 159 (1), 148–162. doi: 10.1016/j.cell.2014.08.028

Shaheen, R., Han, L., Faqeih, E., Ewida, N., Alobeid, E., Phizicky, E. M., et al. (2016). A homozygous truncating mutation in PUS3 expands the role of tRNA modification in normal cognition. *Hum. Genet.* 135 (7), 707–713. doi: 10.1007/s00439-016-1665-7

Shaheen, R., Tasak, M., Maddirevula, S., Abdel-Salam, G. M., Sayed, I. S., Alazami, A. M., et al. (2019). PUS7 mutations impair pseudouridylation in humans and cause intellectual disability and microcephaly. *Hum. Genet.* 138 (3), 231–239. doi: 10.1007/s00439-019-01980-3

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15 (8), 1034–1050. doi: 10.1101/gr.3715005

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164–e164. doi: 10.1093/nar/gkq603

Wong, Y.-H., Lee, T.-Y., Liang, H.-K., Huang, C.-M., Wang, T.-Y., Yang, Y.-H., et al. (2007). KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* 35 (suppl_2), W588–W594.

Xuan, J.-J., Sun, W.-J., Lin, P.-H., Zhou, K.-R., Liu, S., Zheng, L.-L., et al. (2017). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* 46 (D1), D327–D334.

Xue, H., Wei, Z., Chen, K., Tang, Y., Wu, X., Su, J., et al. (2020). Prediction of RNA methylation status from gene expression data using classification and regression methods. *Evol. Bioinf.*

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-Methylation sites in homo sapiens. *J. Comput. Biol.* 25 (11), 1266–1277. doi: 10.1089/cmb.2018.0004

Zaringhalam, M., and Papavasiliou, F. N. (2016). Pseudouridylation meets next-generation sequencing. *Methods* 107, 63–72. doi: 10.1016/j.ymeth.2016.03.001

Zhao, X., Patton, J. R., Davis, S. L., Florence, B., Ames, S. J., and Spanjaard, R. A. (2004). Regulation of nuclear receptor activity by a pseudouridine synthase through posttranscriptional modification of steroid receptor RNA activator. *Mol. Cell* 15 (4), 549–558. doi: 10.1016/j.molcel.2004.06.044

Zhu, Y., Xu, G., Yang, Y. T., Xu, Z., Chen, X., Shi, B., et al. (2018). POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* 47 (D1), D203–D211.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

TECHNOLOGY AND CODE
published: 17 March 2020
doi: 10.3389/fgene.2020.00211

MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets

Luca Cozzuto[1], Huanle Liu[1], Leszek P. Pryszcz[1,2], Toni Hermoso Pulido[1], Anna Delgado-Tejedor[1,3], Julia Ponomarenko[1,3] and Eva Maria Novoa[1,3,4,5]**

[1] Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain, [2] International Institute of Molecular and Cell Biology, Warsaw, Poland, [3] Universitat Pompeu Fabra, Barcelona, Spain, [4] Department of Neuroscience, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia, [5] St Vincent's Clinical School, UNSW Sydney, Darlinghurst, NSW, Australia

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Italian Institute of Technology, Italy

Reviewed by:
Cuncong Zhong,
University of Kansas, United States
Sarath Chandra Janga,
Indiana University, Purdue University Indianapolis, United States

***Correspondence:**
Julia Ponomarenko
julia.ponomarenko@crg.eu
Eva Maria Novoa
eva.novoa@crg.eu;
evamaria.novoa@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

***Received:** 25 October 2019*
***Accepted:** 21 February 2020*
***Published:** 17 March 2020*

Citation:
Cozzuto L, Liu H, Pryszcz LP, Pulido TH, Delgado-Tejedor A, Ponomarenko J and Novoa EM (2020) MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. Front. Genet. 11:211. doi: 10.3389/fgene.2020.00211

The direct RNA sequencing platform offered by Oxford Nanopore Technologies allows for direct measurement of RNA molecules without the need of conversion to complementary DNA, fragmentation or amplification. As such, it is virtually capable of detecting any given RNA modification present in the molecule that is being sequenced, as well as provide polyA tail length estimations at the level of individual RNA molecules. Although this technology has been publicly available since 2017, the complexity of the raw Nanopore data, together with the lack of systematic and reproducible pipelines, have greatly hindered the access of this technology to the general user. Here we address this problem by providing a fully benchmarked workflow for the analysis of direct RNA sequencing reads, termed *MasterOfPores*. The pipeline starts with a pre-processing module, which converts raw current intensities into multiple types of processed data including FASTQ and BAM, providing metrics of the quality of the run, quality-filtering, demultiplexing, base-calling and mapping. In a second step, the pipeline performs downstream analyses of the mapped reads, including prediction of RNA modifications and estimation of polyA tail lengths. Four direct RNA MinION sequencing runs can be fully processed and analyzed in 10 h on 100 CPUs. The pipeline can also be executed in GPU locally or in the cloud, decreasing the run time fourfold. The software is written using the NextFlow framework for parallelization and portability, and relies on Linux containers such as Docker and Singularity for achieving better reproducibility. The *MasterOfPores* workflow can be executed on any Unix-compatible OS on a computer, cluster or cloud without the need of installing any additional software or dependencies, and is freely available in Github (https://github.com/biocorecrg/master_of_pores). This workflow simplifies direct RNA sequencing data analyses, facilitating the study of the (epi)transcriptome at single molecule resolution.

Keywords: Nextflow, direct RNA sequencing, nanopore, Docker, singularity

INTRODUCTION

Next generation sequencing (NGS) technologies have revolutionized our understanding of the cell and its biology. However, NGS technologies are heavily limited by their inability to sequence long reads, thus requiring complex bioinformatic algorithms to assemble back the DNA pieces into a full genome or transcriptome. Moreover, NGS technologies require a PCR amplification step, and as such, they are typically blind to DNA or RNA modifications (Novoa et al., 2017).

The field of epitranscriptomics, which studies the biological role of RNA modifications, has experienced an exponential growth in the last few years. Systematic efforts coupling antibody immunoprecipitation or chemical treatment with next-generation sequencing (NGS) have revealed that RNA modifications are much more widespread than originally thought, are reversible (Jia et al., 2011), and can play major regulatory roles in determining cellular fate (Batista et al., 2014), differentiation (Lin et al., 2017; Furlan et al., 2019; Lee et al., 2019) and sex determination (Haussmann et al., 2016; Lence et al., 2016; Kan et al., 2017), among others. However, the lack of selective antibodies and/or chemical treatments that are specific for a given modification have largely hindered our understanding of this pivotal regulatory layer, limiting our ability to produce genome-wide maps for 95% of the currently known RNA modifications (Jonkhout et al., 2017; Boccaletto et al., 2018).

Third-generation sequencing (TGS) platforms, such as the one offered by Oxford Nanopore Technologies (ONT), allow for direct measurement of both DNA and RNA molecules without prior fragmentation or amplification (Brown and Clarke, 2016), thus putting no limit on the length of DNA or RNA molecule that can be sequenced. In the past few years, ONT technology has revolutionized the fields of genomics and (epi)transcriptomics, by showing its wide range of applications in genome assembly (Jain et al., 2018), study of structural variations within genomes (Cretu Stancu et al., 2017), 3′ poly(A) tail length estimation (Krause et al., 2019; Workman et al., 2019), accurate transcriptome profiling (Bolisetty et al., 2015; Sessegolo et al., 2019), identification of novel isoforms (Byrne et al., 2017; Križanovic et al., 2018) and direct identification of DNA and RNA modifications (Carlsen et al., 2014; Simpson et al., 2017; Garalde et al., 2018; Leger et al., 2019; Liu et al., 2019; Parker et al., 2020). Thus, not only this technology overcomes many of the limitations of short-read sequencing, but importantly, it also can directly measure RNA and DNA modifications in their native molecules. Although ONT can potentially address many problems that NGS technologies cannot, the lack of proper standardized pipelines for the analysis of ONT output has greatly limited its reach to the scientific community.

To overcome these limitations, workflow management systems together with Linux containers offer an efficient solution to analyze large-scale datasets in a highly reproducible, scalable and parallelizable manner. In the last year, several workflows to analyze nanopore data have become available, which are aimed at facilitating genome assembly (e.g., Katuali),[1] genome annotation (e.g., Pinfish[2]) and single nucleotide polymorphism analyses (e.g., NanoPipe[3]). However, none of the current available pipelines cannot be used for the analysis of direct RNA sequencing datasets.

Here we provide a scalable and parallelizable workflow for the analysis of direct RNA (dRNA) sequencing datasets, termed *MasterOfPores*,[4] which uses as input raw direct RNA sequencing FAST5 reads, which is a flexible HDF5 format used by ONT to store raw sequencing data, which includes current intensity values, metadata of the sequencing run and base-called fasta sequences, among other features. The *MasterOfPores* workflow performs both data pre-processing (base-calling, quality control, demultiplexing, filtering, mapping, estimation of per-gene or per-transcript abundances) and data analysis (prediction of RNA modifications and estimation of polyA tail lengths) (**Figure 1**). Thus, the *MasterOfPores* workflow facilitates the analysis of nanopore (epi)transcriptomics sequencing data.

For each step, the workflow extracts metrics which are compiled in a final HTML report that can be easily visualized an analyzed by non-expert bioinformaticians. For each sequencing run, the pipeline produces as output a FASTQ file containing the base-called reads, a BAM file containing the mapped reads, and up to three plain text files containing gene or isoform quantifications, polyA tail length estimations and RNA modification predictions. A direct RNA sequencing run produced by MinION or GridION devices, which typically comprises 1-2M reads, takes ∼2 h to process on a CPU cluster using 100 nodes, and ∼1 h or less on a single GPU (see **Table 1** for detailed metrics). Moreover, the pipeline can also be run on the cloud (see section "Running on AWS").

MasterOfPores simplifies the analysis of direct RNA sequencing data by providing a containerized pipeline implemented in the NextFlow framework. It is important to note that this approach avoids the heavy-lifting of installing dependencies by the user, and thus, is simple and accessible to any researcher with little bioinformatics expertise. We expect that our workflow will greatly facilitate the access of Nanopore direct RNA sequencing to the community.

RESULTS

Overview of the *MasterOfPores* Workflow

Workflow management systems together with Linux containers offer a solution to efficiently analyze large scale datasets in a highly reproducible, scalable and parallelizable manner. During the last decade, an increasing interest in the field has led to the development of different programs such as Snakemake (Köster and Rahmann, 2012), NextFlow (Di Tommaso et al., 2017), Galaxy (Afgan et al., 2018), SciPipe (Lampa et al., 2019) or GenPipes (Bourgey et al., 2019), among others. These tools enable the prototyping and deployment of pipelines by abstracting computational processes and representing pipelines as directed

[1] https://github.com/nanoporetech/katuali

[2] https://github.com/nanoporetech/pipeline-pinfish-analysis

[3] https://github.com/IOB-Muenster/nanopipe2

[4] https://biocorecrg.github.io/master_of_pores/

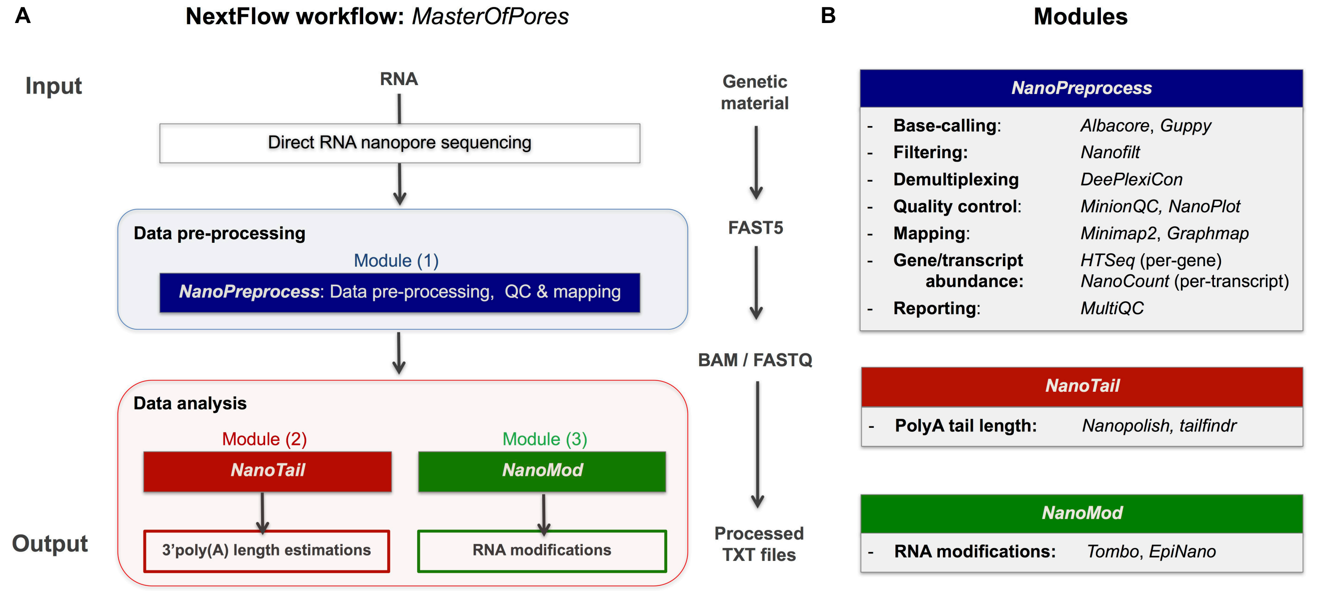

A
NextFlow workflow: *MasterOfPores*
Input
RNA
Direct RNA nanopore sequencing
Data pre-processing
Module (1)
NanoPreprocess: Data pre-processing, QC & mapping
Data analysis
Module (2)
NanoTail
Module (3)
NanoMod
Output
3'poly(A) length estimations
RNA modifications
Genetic material
FAST5
BAM / FASTQ
Processed TXT files
B
Modules
NanoPreprocess
- Base-calling: *Albacore, Guppy*
- Filtering: *Nanofilt*
- Demultiplexing *DeePlexiCon*
- Quality control: *MinionQC, NanoPlot*
- Mapping: *Minimap2, Graphmap*
- Gene/transcript abundance: *HTSeq* (per-gene) *NanoCount* (per-transcript)
- Reporting: *MultiQC*
NanoTail
- PolyA tail length: *Nanopolish, tailfindr*
NanoMod
- RNA modifications: *Tombo, EpiNano*

FIGURE 1 | Overview of the *MasterOfPores* workflow for the processing of direct RNA nanopore sequencing datasets. **(A)** Overview of the 4 modules included in the *MasterOfPores* workflow. The pre-processing module (*NanoPreprocess*) accepts both single FAST5 and multi-FAST5 reads and includes 8 main steps: (i) base-calling, (ii) demultiplexing (iii) filtering, (iv) quality control, (v) mapping and (vi) gene or transcript quantification and (vii) final report building. The outputs generated by *NanoPreprocess* (BAM, FastQ and base-called Fast5) are used as input by the subsequent *MasterOfPores* data analysis modules, to predict RNA modifications (*NanoMod*) and polyA tail length estimations (*NanoTail*). **(B)** Detailed description of the individual steps and software used for each of the 4 modules included in *MasterOfPores*.

graphs, in which nodes represent tasks to be executed and edges represent either data flow or execution dependencies between different tasks.

Here we chose the workflow framework NextFlow (Di Tommaso et al., 2017) because of its native support of different batch schedulers (SGE, LSF, SLURM, PBS, and HTCondor), cloud platforms (Kubernetes, Amazon AWS, and Google Cloud) and GPU computing, which is crucial for processing huge volumes of data produced by nanopore sequencers. NextFlow has tight integration with lightweight Linux containers, such as Docker and Singularity. Automatic organization of intermediate results produced during the NextFlow pipeline execution allows reducing the complexity of intermediary file names and the possibility of name clashing. Continuous check-pointing with the possibility of resuming failed executions, interoperability and meticulous monitoring and reporting of resource usage are among other thought-after features of NextFlow. The executables of the presented pipeline have been bundled within Docker images accessible at DockerHub that can be converted on the fly into a Singularity image, thus allowing the HPC usage.

The *MasterOfPores* workflow includes all steps needed to process raw FAST5 files produced by Nanopore direct RNA sequencing and executes the following steps, allowing users a choice among different algorithms (**Figure 1**). The pipeline consists of 3 modules:

(i) *NanoPreprocess*: this module takes as input the raw Fast5 reads and produces as output base-called sequences both in FAST5 and FASTQ formats, as well as alignments in BAM format. The pre-processing module performs base-calling, demultiplexing, filtering, quality control, mapping and gene and/or transcript quantification, generating a final report of the performance and results of each of the steps performed.

(ii) *NanoTail*: this module takes as input the output from the NanoPreprocess module and produces polyA tail length estimations using two different algorithms.

(iii) *NanoMod*: this module takes as input the files generated during the pre-processing step, and produces flat text files with the predicted RNA modifications using two different algorithms.

Pre-processing Module: *NanoPreprocess*

The *NanoPreprocess* module consists of 8 main steps (**Figure 2**):

(i) Read base-calling with the algorithm of choice, using *Albacore*[5] or *Guppy*.[5] This step can be run in parallel and the user can decide the number of files to be processed in a single job by using the command –*granularity*.

(ii) Demultiplexing of the reads using *DeePlexiCon* (Smith et al., 2019). This step is optional, and can only be used if the libraries have been barcoded using the oligonucleotides used to train the deep neural classifier[6]

(iii) Filtering of the resulting fastq files using *Nanofilt* (De Coster et al., 2018). This step is optional and can be run in parallel.

[5] https://nanoporetech.com

[6] https://github.com/Psy-Fer/deeplexicon

TABLE 1 | Comparison of computing time and RAM used to run the pipeline for the four *S. cerevisiae* polyA(+) direct RNA sequencing datasets used in this study.

		Yeast WT *rep1*	**Yeast ime△ KO *rep1***	**Yeast WT *rep2***	**Yeast ime△ KO *rep2***
Raw data	**Number of reads**	**1,197,462**	**694,907**	**629,270**	**573,404**
Module (1): NanoPreprocess					
CPU*	Total time	2 h 13 min	2 h 6 min	2 h 11 min	2 h 1 min
	Total time per 1000 reads (s)	7 s	10 s	12 s	12 s
GPU**	Total time	6 h 44 min	4 h 05 min	3 h 59 min	3 h 19 min
	Total time per 1000 reads (s)	20 s	21 s	23 s	21 s
GPU***	Total time	1 h 8 m	37 min	36 min	30 min
	Total time per 1000 reads (s)	3 s	2 s	2 s	1 s
Module (2): NanoTail					
CPU*	Total time	3 h 26 min			
	Total time per 1000 reads (s)	4 s			
Module (3): NanoMod					
CPU*	Total time	5 h 40 min			
	Total time per 1000 reads (s)	7 s			

**CPU time computed using a maximum of 100 nodes with 8 CPU per node; **GPU time computed using 1 card GIGABYTE GeForce RTX 1660 Ti; ***GPU time computing using 1 card INNO3D GeForce RTX 2080.*

(iv) Quality control of the base-called data, using *MinIONQC* (Lanfear et al., 2019) and FastQC.[7]

(v) Read mapping to the reference genome or transcriptome, using *minimap2*[8] or *graphmap2*.[9]

(vi) Quality control on the alignment, using *NanoPlot*[10] and *bam2stats*.[11]

(vii) Gene or transcript quantification, using *HTSeq* (Anders et al., 2015) or *NanoCount*.[12] The latter estimates transcript abundance using an expectation-maximization algorithm. *NanoCount* will be run if reads have been mapped to the transcriptome, using the flag *–reference_type* transcriptome, whereas *HTSeq* will be employed to quantify per-gene counts if the reads have been mapped to the genome.

(viii) Final report of the data processing using *multiQC*[13] that combines the single quality controls done previously, as well as global run statistics (**Figure 3**).

Data Analysis Modules: *NanoTail* and *NanoMod*

The *MasterOfPores* pipeline contains two additional modules for the downstream analyses of the mapped reads, namely *NanoTail* and *NanoMod*, which provide polyA tail length estimations and RNA modification predictions, respectively (**Figure 2**). The modules can be run using as input the output from the *NanoPreprocess* module.

The *NanoTail* module estimates polyA tail lengths using *Nanopolish*[14] and *TailfindR*,[15] producing a plain text file with polyA tail length estimations for each read, computed using both algorithms. The correlation between the two algorithms is also reported as a plot.

The *NanoMod* module predicts RNA modifications using *Tombo*[16] and *EpiNano*,[17] producing a plain text files with the predicted sites by each algorithm. The NanoMod module is run "paired mode," i.e., providing two conditions, as both *EpiNano* and *Tombo* identify RNA modifications by comparing two conditions.

Running *MasterOfPores:* Installation, Input, Parameters and Output

To run *MasterOfPores*, the following steps are required:

(i) Install NextFlow (version 19.10.0):

```
$ curl -s https://get.nextflow.io | bash
```

(ii) Clone the MasterOfPores repository:

```
$ git clone –depth 1 https://github.com/biocorecrg/master_of_pores.git
```

(iii) Install Docker or Singularity (for Singularity, version 2.6.1 and Docker 19.03 or later are required):

Docker: https://docs.docker.com/install/
Singularity: https://sylabs.io/guides/2.6/user-guide/quick_start.html#quick-installation-steps

(iv) Download Nanopore base-calling algorithms: *guppy* with or without GPU support and or the albacore Wheel file (a standard built-package format used for Python distributions) and install them inside the *bin* folder inside the MasterOfPores directory. The users can place their preferred version of guppy and/or albacore in the *bin* folder (in the example below, albacore version 2.1.7 and guppy 3.1.5).

```
$ cd master_of_pores/NanoPreprocess/bin
$ tar -zvxf ont-guppy_3.1.5_linux64.tar.gz
$ ln -s ont-guppy_3.1.5_linux64/ont-guppy/bin/guppy_*.
$ pip3 install –target = ./albacore ont_albacore-2.1.7-cp36-cp36m-manylinux1_x86_64.whl
$ ln -s albacore/bin/multi_to_single_fast5
$ ln -s albacore/bin/read_fast5_basecaller.py
```

[7] http://www.bioinformatics.babraham.ac.uk/projects/fastqc
[8] https://github.com/lh3/minimap2
[9] https://github.com/lbcb-sci/graphmap2
[10] https://github.com/wdecoster/NanoPlot
[11] https://github.com/lpryszcz/bin
[12] https://github.com/a-slide/NanoCount
[13] https://github.com/ewels/MultiQC
[14] https://github.com/jts/nanopolish
[15] https://github.com/adnaniazi/tailfindr
[16] https://github.com/nanoporetech/tombo
[17] https://github.com/enovoa/EpiNano

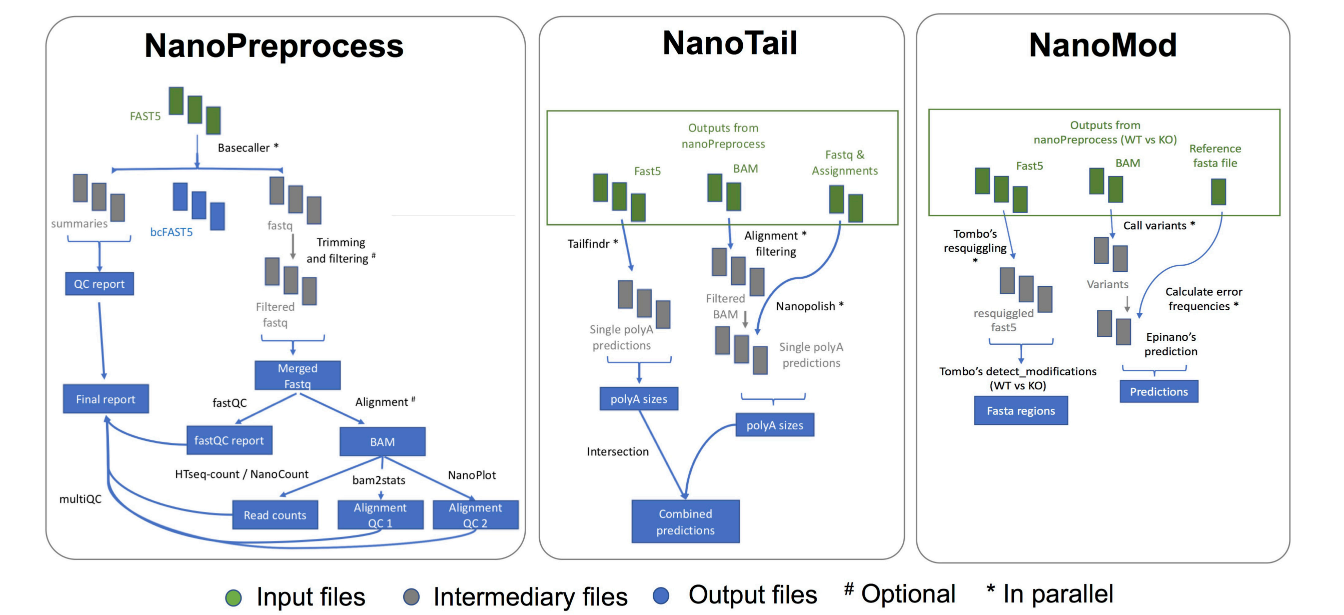


FIGURE 2 | Scheme of the individual steps performed, inputs and outputs of the three modules (NanoPreprocess, NanoTail, and NanoMod) included in *MasterOfPores* workflow. The inputs required by each module are depicted in green, whereas final outputs generated by each module are shown in blue.

(v) Optional step: install CUDA drivers (only needed for GPU support):

https://docs.nvidia.com/cuda/cuda-installation-guide-linux/index.html

(vi) Run the pre-processing step of the pipeline (using singularity *or* docker):

```
$ cd./
$ nextflow run nanopreprocess.nf -with-singularity
or
$ nextflow run nanopreprocess.nf -with-docker
```

(vii) Run polyA tail estimation module

```
$ cd./NanoTail
$ nextflow run nanotail.nf -bg -with-singularity –input_folders ".NanoPreprocess/output/RNA*"
```

(viii) Run RNA modification prediction module

```
$ cd./NanoMod
$ nextflow run nanomod.nf -with-singularity input_path ".NanoPreprocess/output/"
```

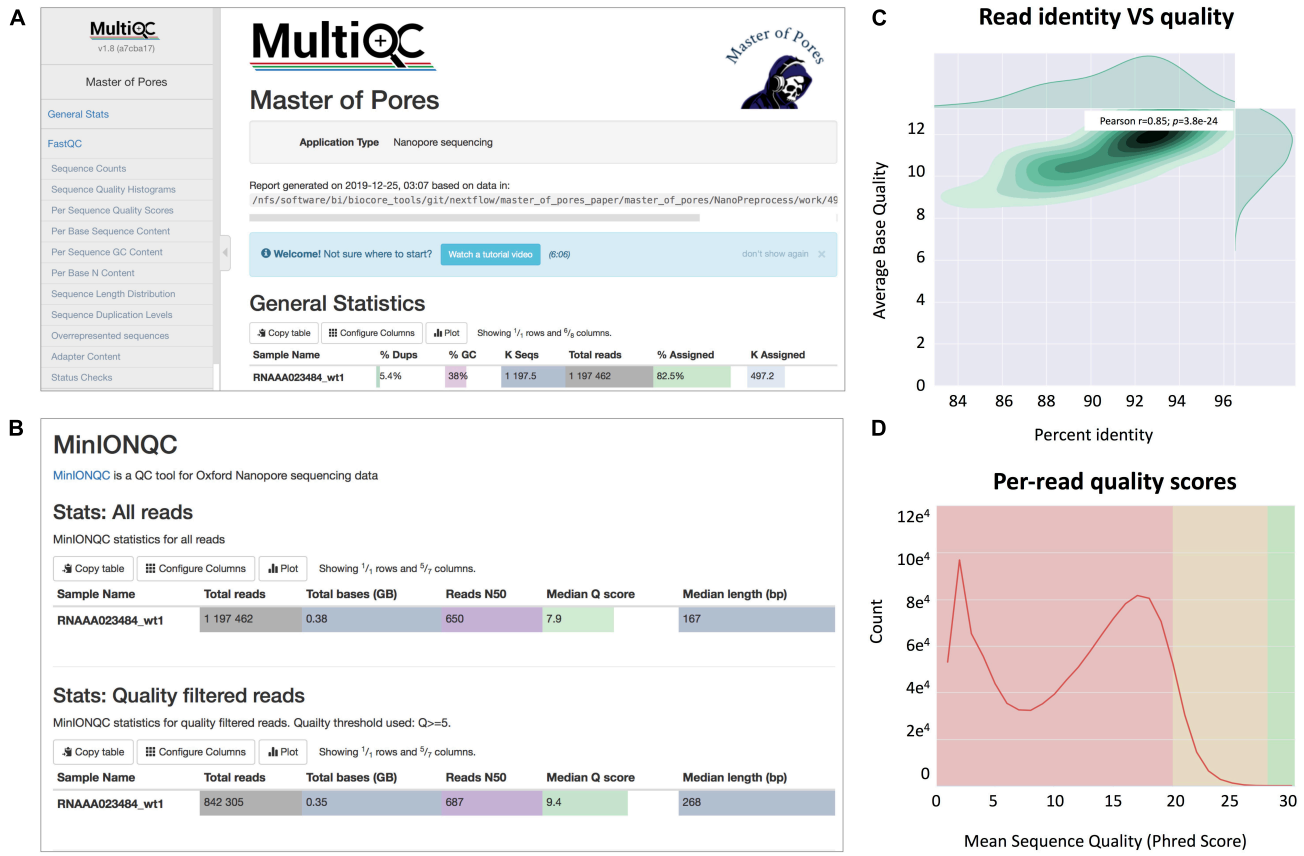
The *NanoPreprocess* module can handle both single- and multi-FAST5 reads as input. To execute the workflow, several parameters can be defined by the user, including the choice of the basecaller (albacore or guppy), mapper (minimap2 or graphmap2), as well as their command line options. If these are not specified by the user, the workflow will be run with default parameter settings specified in the params.config file (**Table 2**). The final report includes four different types of metrics: (i) *General statistics* of the input, including the total number of reads, GC content and number of identical base-called sequences; (ii) *Per-read statistics* of the input data, including scatterplots of the average read length versus sequence identity, the histogram of read lengths, and the correlation between read quality and identity; (iii) *Alignment statistics*, including the total number of mapped reads, the total number of mapped bases, the average length of mapped reads, and the mean sequence identity; (iv) *Quality filtering statistics,* including the number of filtered reads, median Q-score and read length, compared to those observed in all sequenced reads*;* and (v) *Per-read analysis of biases*, including information on duplicated reads, over-represented reads and possible adapter sequences (**Figure 3**). The final outputs of this module include:

- Basecalled fast5 files within the "fast5_files" folder.
- Filtered fastq files within "fastq_files" folder.
- QC reports within "QC" folder.
- Final report within "report" folder.
- Aligned reads in sorted BAM files within the "aln" folder.
- Read counts within the "counts" folder.

The *NanoMod* module requires two samples to detect RNA modifications, typically wild-type and knock-out (or knock-down) matched conditions. The user must provide a tab-delimited file (–comparison "comparison.tsv") indicating which input file is the wild-type condition and which one is the knock-out or knock-down condition (see, for example[18]), which is specified in the parameter file. The *NanoMod* module will output the results into two different folders:

- RNA modification results predicted using *Tombo* in the "Tombo" folder
- RNA modification results predicted using *EpiNano* in the "EpiNano" folder

[18] https://github.com/biocorecrg/master_of_pores/NanoMod/comparison.tsv



MultiQC
v1.8 (a7cba17)

Master of Pores

General Stats

FastQC

- Sequence Counts
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Status Checks

MultiQC

Master of Pores

Application Type Nanopore sequencing

Report generated on 2019-12-25, 03:07 based on data in:

```
/nfs/software/bi/biocore_tools/git/nextflow/master_of_pores_paper/master_of_pores/NanoPreprocess/work/49
```

Welcome! Not sure where to start? Watch a tutorial video (6:06) don't show again ×

General Statistics

Copy table | Configure Columns | Plot | Showing $^1/_1$ rows and $^6/_8$ columns.

Sample Name	% Dups	% GC	K Seqs	Total reads	% Assigned	K Assigned
RNAAA023484_wt1	5.4%	38%	1 197.5	1 197 462	82.5%	497.2

MinIONQC

MinIONQC is a QC tool for Oxford Nanopore sequencing data

Stats: All reads

MinIONQC statistics for all reads

Copy table | Configure Columns | Plot | Showing $^1/_1$ rows and $^5/_7$ columns.

Sample Name	Total reads	Total bases (GB)	Reads N50	Median Q score	Median length (bp)
RNAAA023484_wt1	1 197 462	0.38	650	7.9	167

Stats: Quality filtered reads

MinIONQC statistics for quality filtered reads. Quailty threshold used: Q>=5.

Copy table | Configure Columns | Plot | Showing $^1/_1$ rows and $^5/_7$ columns.

Sample Name	Total reads	Total bases (GB)	Reads N50	Median Q score	Median length (bp)
RNAAA023484_wt1	842 305	0.35	687	9.4	268

FIGURE 3 | Snapshots of the final report generated by *MasterOfPores*. **(A)** Main menu and overview of the final report generated by *MasterOfPores*. **(B)** The report includes detailed metrics on the input reads ("MinIONQC"), as well as on the mapped reads ("AlignmentQC"). **(C,D)** Example of plots that are included as part of the *MasterOfPores* final report, some of which are generated by integrating Nanoplot **(C)** and FastQC **(D)** software.

The *NanoTail* module will output the results into three different folders:

- PolyA tail length estimates predicted using *Nanopolish*, in the "Nanopolish" folder.
- PolyA tail length estimates predicted using *tailfindR*, in the "Tailfindr" folder.
- In this module, an additional "NanoMod_final" folder is generated, containing combined *Nanopolish* and *tailfindR* estimates of polyA tail lengths, as well as information regarding the geneID or transcriptID where the read is mapped to.

Running *MasterOfPores* on the Cloud (AWS Batch and AWS EC2)

Nanopore sequencing allows for real-time sequencing of samples. While GridION devices come with built-in GPUs that allows live base-calling, smaller MinION devices do not have built-in CPU or GPU. Thus, the user has to connect the MinION to a computer with sufficient CPU/GPU capabilities, or run base-calling after the sequencing. In all these contexts, the possibility of running the *MasterOfPores* pipeline on the cloud presents a useful alternative.

The Amazon Web Services (AWS) Batch is a computing service that enables users to submit jobs to a cloud-based user-defined infrastructure, which can be easily set up via either code-based definitions or a web-based interface. Computation nodes can be allocated in advance or according to resource availability. Cloud infrastructure can be also deployed or dismantled on demand using automation tools, such as CloudFormation or Terraform.

Here we show that the MasterOfPores pipeline can be successfully implemented on the cloud, and provide the Terraform script for running *MasterOfPores* on the AWS Batch CPU environments, available in the GitHub repository.[19] To run the pipeline using the AWS Batch, the users needs to change only a few parameters related to their accounts in a configuration file. The pipeline can be run from either a local workstation or an Amazon EC2 entrypoint instance initiated for this purpose (we recommend the latter). Data to be analyzed can be uploaded to an Amazon S3 storage bucket.

Similarly, we also tested whether our pipeline could be run in Amazon Web Services (AWS) Elastic Compute Cloud (EC2), which is one of the most popular cloud services

[19] https://biocorecrg.github.io/master_of_pores/

TABLE 2 | Settings and parameters that can be customized to run the NanoPreprocess module of the MasterOfPores workflow.

	Parameter	Description of the parameter	Default Values
RUN_INFO	kit	Sequencing kit used (SQK-RNA001 or SQK-RNA002)	SQK-RNA002
	flowcell	flowcell type	FLO-MIN106
	fast5	fast5 files including the path	“$baseDir/data/multifast/*.fast5”
	annotation	annotation file (GTF) including path	“”
	reference	reference genome or transcriptome sequence	“$baseDir/anno/curlcake_constructs.fasta.gz”
	ref_type	reference type (genome or transcriptome)	“genome”
RUN_SETUP	seqtype	sequence type (RNA or DNA)	“RNA”
	output	Output folder	“$baseDir/output”
	qualityqc	Quality threshold for QC	5
	granularity	Number of files analyzed per process	“”
DEMULTIPLEXING	demultiplexer	Option to run demultiplexing, in case the run is barcoded (ON or OFF)	“OFF”
	demultiplex_opt	choose between different pre-trained models	“-m pAmps-final-actrun_newdata_nanopore_UResNet20v2_model.030.h5”
BASE-CALLING	basecaller	Can be: albacore/guppy	“guppy”
	basecaller_opt	Command line options for basecalling	“”
	GPU	Whether or not using GPU (ON or OFF)	“OFF”
FILTERING	filter	Can be empty, OFF or nanofilt	“”
	filter_opt	command line options for filtering	“”
MAPPING	mapper	Can be minimap2 or graphmap2 or empty	minimap2
	mapper_opt	Command line options for mapping	“-uf -k14”
	map_type	Can spliced or unspliced	“spliced”
	reference_type	can be transcriptome, genome or both	“genome”
GENE COUNTING	counter	Option to compute per-gene or per-transcript counts from the mapped BAM file (YES or NO)	“YES”
	counter_opt	Command line options for counting. Of note, per-gene counts will be computed using HTSeq if reference_type is “genome,” or computed using NanoCount if reference_type is “transcriptome”	“”
REPORTING	email	Email (to receive the report when finished)	“”

(**Supplementary Table S1**). Compared to AWS Batch, to run any workflow in AWS EC2, the user must first create an Amazon Machine Image (AMI). The AMI can be created using the same instructions as provided in **Supplementary File S1**, starting from the official Ubuntu 18.04 LTS AMI, and including both Docker and Singularity software with NVIDIA libraries support. Here we show that the resulting image can be used to run the *MasterOfPores* workflow with NVIDIA Tesla V100 GPU cards. Automation scripts to run *MasterOfPores* in AWS EC2 can be found in the GitHub repository.[20]

Test Case: Analysis of *Saccharomyces cerevisiae* SK1 PolyA(+) RNA

Running the *MasterOfPores* Pipeline on *S. cerevisiae* PolyA(+) RNA

To benchmark the performance of the *MasterOfPores* workflow, we employed four publicly available direct RNA sequencing runs of polyA(+)-selected *S. cerevisiae* WT and ime4△ strains, in biological replicates, which had been sequenced using MinION and GridION devices, producing a total of ∼3 million reads (**Table 1**). We used up to 100 nodes with 8 CPUs for testing the base-calling in CPU mode and 1 node with 1 GPU card for testing the base-calling in GPU mode (**Table 1**).

The MasterOfPores *NanoPreprocess* module was ran using guppy version 3.1.5 as the base-caller and minimap2 version 2.17 as the mapping algorithm. Reads were filtered by running nanofilt with the options “-q 0 –headcrop 5 –tailcrop 3 –readtype 1D”. Filtered reads were mapped to the yeast SK1 fasta genome. Specifically, the command that was executed to run the pipeline with these settings was:

```
$ cd master_of_pores/NanoPreprocess
$ nextflow run nanopreprocess.nf –basecaller guppy –seqtype RNA \
–fast5 “FOLDERNAME/*.fast5” –demultiplexing “OFF” \
–map_type “spliced” –mapper_opt “-uf -k14” \
–reference genome.fa.gz –mapper minimap2 –ref_type “genome”\
–filter nanofilt –filter_opt “-q 0 –headcrop 5 –tailcrop 3 –readtype 1D”.
```

Then, the two data analysis modules were executed as follows:

```
$ nextflow run nanotail.nf –input_folders “./NanoPreprocess/output/* ” \
–nanopolish_opt “” –tailfindr_opt “” –reference “genome.fa.gz”
```

```
$ nextflow run nanomod.nf –input_path “./NanoPreprocess/output/” \
```

[20] https://biocorecrg.github.io/master_of_pores/

```
–comparison "./comparison.tsv" –reference "genome.fa.gz" \
–tombo_opt "–num-bases 5" –epinano_opt ""
```

Benchmarking the Time Used for the Analysis of *S. cerevisiae* PolyA(+) RNA

Here we have tested the pipeline using both CPU and GPU computing. Specifically, we ran the pipeline on the following configurations: (i) a single CPU node (e.g., emulating the computing time on a single laptop); (ii) a CPU cluster with 100 nodes; (iii) a single mid-range GPU card (RTX2080); and (iv) a single high-end GPU card (GTX1080 Ti). We found that the computing time required to run the pipeline on a single GPU card was significantly lower than the running time in parallel on a high performance CPU cluster with 100 nodes, 8 cores per node (**Table 1**, see also **Supplementary Table S1**). Moreover, we found that the computing time of the NanoPreprocess module can be significantly reduced depending on the GPU card (base-calling step was ∼2X faster for GTX1080 Ti than for RTX2080).

Reporting Resources Used for the Analysis of *S. cerevisiae* PolyA(+) RNA

Taking advantage of the NextFlow reporting functions, the pipeline can produce detailed reports on the time and resources consumed by each process (**Figure 4**), in addition to the output files (bam, fastq) and final report (html), if the workflow is executed with parameters *-with-report* (formatted report) or-*with-trace* (plain text report). Running the base-calling on each multi-fast5 file in parallel on our dataset showed that the most memory intensive tasks (about 5 Gbytes) were the mapping step (using minimap2) and the quality control step (using *Nanoplot*) (**Table 3**), while the most CPU-intensive and time-consuming step (∼80 min) was the base-calling (using *Guppy*) (**Table 4**).

Finally, we should note that the latest (19.10.0) version of NextFlow allows the user to control the execution of a pipeline remotely. To enable this feature, the user needs to login to the https://tower.nf/website developed by the NextFlow authors and retrieve a token for communicating with the pipeline. For doing that, the user should set this token as an environmental variable and run the pipeline as follows:

```
$ export TOWER_ACCESS_TOKEN = YOUR_TOKEN
$ cd master_of_pores/NanoPreprocess
$ nextflow run nanopreprocess.nf -with-docker -with-report -bg -
with-tower
```

DISCUSSION

The direct RNA sequencing technology developed by Oxford Nanopore technologies (ONT) offers the possibility of sequencing native RNA molecules, allowing to investigate the (epi)transcriptome at an unprecedented resolution, in full-length RNA molecules and in its native context. Although the direct RNA sequencing library preparation kit was made available in April 2017, only a modest number of researchers have started to adopt this new technology, partly due to the complexity of analyzing the resulting raw FAST5 data. Moreover, even in those cases when specific software and tools have been made available, the users typically experience many difficulties in installing dependencies and running the software. To overcome these issues and facilitate the data analysis of direct RNA sequencing to the general user, we propose the use of NextFlow workflows.

Specifically, we propose the use of *MasterOfPores* workflow for the analysis of direct RNA sequencing datasets, which is a containerized pipeline implemented in the NextFlow framework. *MasterOfPores* can handle both single- and multi-FAST5 reads as input, is highly customizable by the user (**Table 2**) and produces informative detailed reports on both the FAST5 data processing and analysis (MultiQC report, **Figure 3**) as well as on the computing resources used to perform each step (NextFlow report, see **Figure 4**). Thus, the current outputs of the *MasterOfPores* workflow include: (i) base-called FAST5 files, (ii) base-called fastq file, (iii) sorted BAM file containing mapped reads, (iv) per-gene or per-transcript counts (depending on algorithm choice), (v) MultiQC report, (vi) NextFlow report, (vii) per-read polyA tail length estimations, including the correlation of predictions using two distinct algorithms, and (viii) per-site RNA modification predictions, including a final plain text file containing the consensus sites predicted by two distinct algorithms.

The process of Nanopore read base-calling, that is, converting ion current changes into the sequence of RNA/DNA bases, has significantly improved during the last few years, mainly due to the adoption of deep learning approaches, such as the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are currently the most commonly used strategies for base-calling. The adoption of RNN and CNN-based base-calling algorithms has led to a dramatic improvement in base-calling accuracy. However, this has come at the expense of a higher computational cost: only 5–10 reads can be base-called on 1 CPU core per second using the latest versions of the base-calling algorithms. The use of graphic processing units (GPUs) can greatly accelerate certain CPU-intensive computational tasks, thus allowing to process 50–500 reads per second (**Supplementary Table S1**). We therefore developed our pipeline for both CPU and GPU computing. Moreover, we provide the GPU-enabled docker image and detailed information on how to setup the GPU computing (see section: "Running MasterOfPores"). We encourage users to adopt the GPU computing for the analysis of Nanopore sequencing data whenever possible, as this option is both more time- and cost-efficient.

MATERIALS AND METHODS

Code Availability

The pipeline is publicly available at https://github.com/biocorecrg/master_of_pores under an MIT license. The example input data as well as expected outputs are included in the GitHub repository. Detailed information on program versions used can be found in the GitHub repository. EpiNano was modified from its original version (1.0) to decrease the

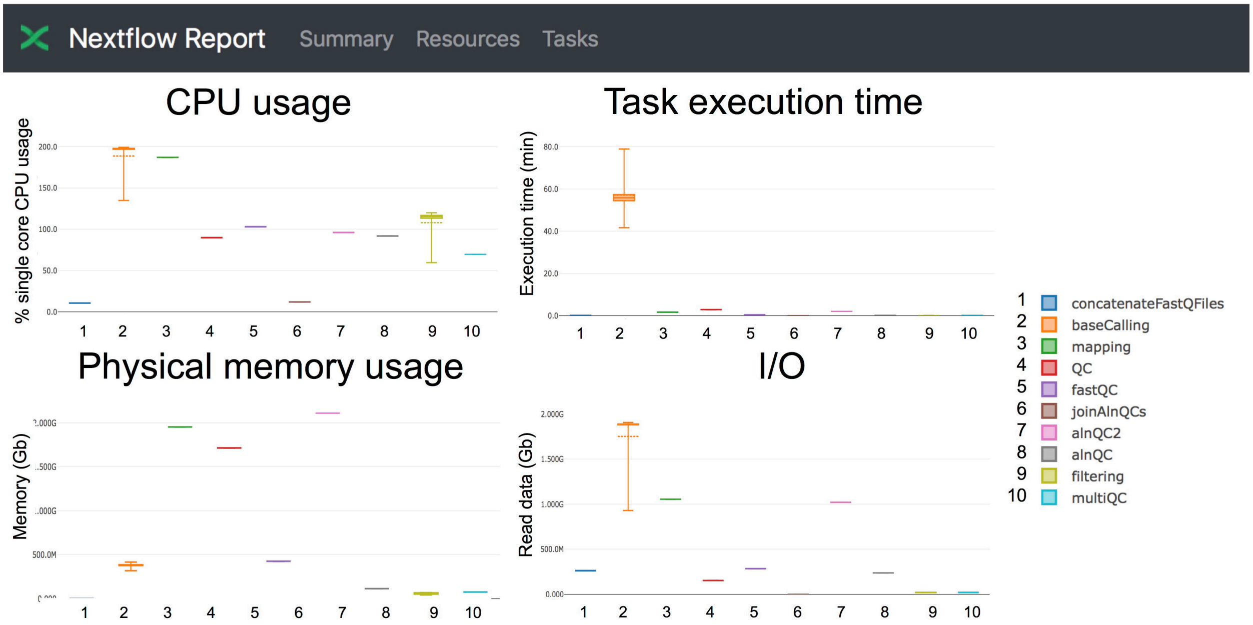


FIGURE 4 | Snapshot of the NextFlow resources report. The report includes detailed information of the computing resources and time needed to execute each of the modules of the pipeline. Base-calling and mapping are the most CPU demanding tasks. The base-calling step is the longest to run, whereas mapping and generation of alignment QC metrics are the most memory-demanding tasks.

computing time of the pipeline (EpiNano version 1.1, available at https://github.com/enovoa/EpiNano).

Documentation Availability

Detailed documentation on how to install and use the pipeline can be found at: https://biocorecrg.github.io/master_of_pores/

Availability of Docker Files and Docker Images

The pipeline uses software that is embedded within Docker containers. Docker files are available in the GitHub repository.[21]

The pipeline retrieves a specific Docker image from DockerHub. In particular, the workflow retrieves four distinct images: one for basecalling,[22] one for demultiplexing,[23] one for pre-processing[24] and one for measuring polyA tail lengths and detecting RNA modifications.[25]

[21] https://github.com/biocorecrg/master_of_pores/tree/master/docker/

[22] https://cloud.docker.com/u/biocorecrg/repository/docker/biocorecrg/mopbasecall

[23] https://hub.docker.com/repository/docker/biocorecrg/mopdem

[24] https://cloud.docker.com/u/biocorecrg/repository/docker/biocorecrg/mopprepr

[25] https://cloud.docker.com/u/biocorecrg/repository/docker/biocorecrg/mopmod

TABLE 3 | RAM peak (Mbytes) used by each of the pre-processing module.

Sample	Number of reads (M)	Base-calling	Mapping	QC	FastQC	alnQC	alnQC2	Filtering	Counting	MultiQC
wt1	1.2	578	4,517	2,751	283	109	4,891	76	34	76
wt2	0.6	458	2,129	1,651	520	39	4,751	69	34	57
ko1	0.7	417	1,954	1,715	427	115	2,111	70	34	77
ko2	0.6	480	1,771	1,400	494	49	2,266	69	34	75

TABLE 4 | CPU time peak (min) used by each of the steps of the pre-processing module.

Sample	Number of reads (M)	Base- calling	Mapping	QC	FastQC	alnQC	alnQC2	Filtering	Counting	MultiQC
wt1	1.2	33	1	4	1	1	2	1	9	1
wt2	0.6	67	1	3	1	1	1	1	4	1
ko1	0.7	79	2	3	1	1	2	1	6	1
ko2	0.6	66	1	3	1	1	1	1	4	1

Integration of Base-Calling Algorithms in the Docker Images

Due to the terms and conditions that users agree to when purchasing Nanopore products, we are not allowed to distribute Nanopore software (binaries or in packaged form like docker images). While the original version of the *MasterOfPores* pipeline includes both guppy and albacore, we are not legally allowed to distribute it with the binaries. Therefore, here we only make available a version where the binaries must be downloaded and placed into a specific folder by the user. We expect future versions of *MasterOfPores* will include these programs within the docker image once this issue is solved.

CPU and GPU Computing Time and Resources

The *MasterOfPores* workflow was tested both locally (using either CPU or GPU) as well as in the cloud (AWS). Computing times for each mode are shown in **Table 1**. CPU time was determined using a maximum of 100 nodes simultaneously with maximum 8 cores CPU per node (2.8–3.5 GHz, 80–130 Watt). GPU time was computed using either GIGABYTE GeForce RTX 1660 Ti (1536 CUDA cores @ 1770 MHz with 6GB of GDDR6 vRAM memory, 120 Watt) or INNO3D GeForce RTX 2080 (2944 CUDA cores @ 1710 MHz with 8 GB of GDDR6 vRAM memory, 225 Watt) or NVIDIA Tesla V100 (5120 CUDA cores + 640 Tensor cores @ 1462 MHz with 16 GB of HBM2 memory). For GPU computing, both system memory (RAM) and GPU memory (vRAM) are used. Base-calling with guppy typically uses 1 or 4.2 Gb of vRAM in fast and high accuracy mode, respectively. As a result, only one base-calling process can be performed on above mentioned cards in high accuracy mode at given time. The execution time in the AWS EC2 p3.2xlarge instance involves reading files already placed in a previously set-up S3 storage bucket but not writing back output results into it.

DATA AVAILABILITY STATEMENT

Direct RNA sequencing datasets for *Saccharomyces cerevisiae* SK1 PolyA(+) RNA were taken from publicly available GEO datasets (GSE126213).

AUTHOR CONTRIBUTIONS

LC wrote the pipeline. HL optimized EpiNano code to incorporate it into the pipeline and tested the pipeline. LP tested the pipeline and implemented GPU computing for containers. TP implemented and tested the workflow for AWS cloud computing. AD-T tested the pipeline with MinION, GridION, and PromethION sequencing runs and helped with the optimization of the NanoPreprocess module. JP and EN supervised the work. LC and EN made figures and tables. EN conceived the project. LC, JP, and EN wrote the manuscript, with contributions from all authors.

FUNDING

This work was partly supported by the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) (PGC2018-098152-A-100 to EN) and by the Australian Research Council (DP180103571 to EN). LP was supported by funding from the European Union's H2020 Research and Innovation Programme under Marie Skłodowska-Curie grant agreement no. 754422. We acknowledge the support of the MEIC to the EMBL partnership, Centro de Excelencia Severo Ochoa and CERCA Programme/Generalitat de Catalunya. This article is based upon work from COST Action EPITRAN supported by COST (European Cooperation in Science and Technology).

ACKNOWLEDGMENTS

We thank all the members of the Novoa lab for their valuable insights and discussion. We would also like to thank Benjamin Lang for letting us benchmark our pipeline on his hardware (GTX1080 Ti).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00211/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., et al. (2014). m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* 15, 707–719. doi: 10.1016/j.stem.2014.09.019

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Bolisetty, M. T., Rajadinakaran, G., and Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* 16:204. doi: 10.1186/s13059-015-0777-z

Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., et al. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. *Gigascience* 8:giz037. doi: 10.1093/gigascience/giz037

Brown, C. G., and Clarke, J. (2016). Nanopore development at Oxford Nanopore. *Nat. Biotechnol.* 34, 810–811. doi: 10.1038/nbt.3622

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8:16027. doi: 10.1038/ncomms16027

Carlsen, A. T., Zahid, O. K., Ruzicka, J. A., Taylor, E. W., and Hall, A. R. (2014). Selective detection and quantification of modified DNA with solid-state nanopores. *Nano Lett.* 14, 5488–5492. doi: 10.1021/nl501340d

Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* 8:1326. doi: 10.1038/s41467-017-01343-4

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Furlan, M., Galeota, E., de Pretis, S., Caselle, M., and Pelizzola, M. (2019). m6A-Dependent RNA dynamics in T Cell differentiation. *Genes* 10:28. doi: 10.3390/genes10010028

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi: 10.1038/nmeth.4577

Haussmann, I. U., Bodi, Z., Sanchez-Moran, E., Mongan, N. P., Archer, N., Fray, R. G., et al. (2016). m6A potentiates Sxl alternative pre-mRNA splicing for robust *Drosophila* sex determination. *Nature* 540, 301–304. doi: 10.1038/nature20577

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi: 10.1038/nbt.4060

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., et al. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* 7, 885–887. doi: 10.1038/nchembio.687

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., and Novoa, E. M. (2017). The RNA modification landscape in human disease. *RNA* 23, 1754–1769. doi: 10.1261/rna.063503.117

Kan, L., Grozhik, A. V., Vedanayagam, J., Patil, D. P., Pang, N., Lim, K.-S., et al. (2017). The m6A pathway facilitates sex determination in *Drosophila*. *Nat. Commun.* 8:15737. doi: 10.1038/ncomms15737

Köster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Krause, M., Niazi, A. M., Labun, K., Torres Cleuren, Y. N., Müller, F. S., and Valen, E. (2019). tailfindr: alignment-free poly(A) length measurement for oxford nanopore RNA and DNA sequencing. *RNA* 25, 1229–1241. doi: 10.1261/rna.071332.119

Križanovic, K., Echchiki, A., Roux, J., and Šikic, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* 34, 748–754. doi: 10.1093/bioinformatics/btx668

Lampa, S., Dahlö, M., Alvarsson, J., and Spjuth, O. (2019). SciPipe: a workflow library for agile development of complex and dynamic bioinformatics pipelines. *Gigascience* 8:giz044. doi: 10.1093/gigascience/giz044

Lanfear, R., Schalamun, M., Kainer, D., Wang, W., and Schwessinger, B. (2019). MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics* 35, 523–525. doi: 10.1093/bioinformatics/bty654

Lee, H., Bao, S., Qian, Y., Geula, S., Leslie, J., Zhang, C., et al. (2019). Stage-specific requirement for Mettl3-dependent m6A mRNA methylation during haematopoietic stem cell differentiation. *Nat. Cell Biol.* 21, 700–709. doi: 10.1038/s41556-019-0318-1

Leger, A., Amaral, P. P., Pandolfini, L., and Capitanchik, C. (2019). RNA modifications detection by comparative Nanopore direct RNA sequencing. *BioRxiv.* [preprint]. Available at: https://www.biorxiv.org/content/10.1101/843136v1.abstract

Lence, T., Akhtar, J., Bayer, M., Schmid, K., Spindler, L., Ho, C. H., et al. (2016). m6A modulates neuronal functions and sex determination in *Drosophila*. *Nature* 540, 242–247. doi: 10.1038/nature20568

Lin, Z., Hsu, P. J., Xing, X., Fang, J., Lu, Z., Zou, Q., et al. (2017). Mettl3-/Mettl14-mediated mRNA N6-methyladenosine modulates murine spermatogenesis. *Cell Res.* 27, 1216–1230. doi: 10.1038/cr.2017.117

Liu, H., Begik, O., Lucas, M. C., Mason, C. E., and Schwartz, S. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *bioRxiv* [preprint]. doi: 10.1038/s41467-019-11713-9 .

Novoa, E. M., Mason, C. E., and Mattick, J. S. (2017). Charting the unknown epitranscriptome. *Nat. Rev. Mol. Cell Biol.* 18, 339–340. doi: 10.1038/nrm.2017.49

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., et al. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife* 9:e49658. doi: 10.7554/elife.49658

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., et al. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci. Rep.* 9:14908. doi: 10.1038/s41598-019-51470-9

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14, 407–410. doi: 10.1038/nmeth.4184

Smith, M. A., Ersavas, T., Ferguson, J. M., Liu, H., Lucas, M. C., Begik, O., et al. (2019). Barcoding and demultiplexing Oxford Nanopore native RNA sequencing reads with deep residual learning. *bioRxiv* [pre print]. doi: 10.1101/864322

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., and Tyson, J. R. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nature* 16, 1297–1305. doi: 10.1038/s41592-019-0617-2

Conflict of Interest: EN has received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

RNA N6-Methyladenosine-Related Gene Contribute to Clinical Prognostic Impact on Patients With Liver Cancer

*Wei Wang, Bo Sun, Yang Xia, Shenghong Sun and Chiyi He**

Department of Gastroenterology, Yijishan Hospital, Wannan Medical College, Wuhu, China

Liver cancer (LC) is the fourth leading cause of cancer-related deaths worldwide. There is an urgent need to identify novel and reliable prognostic biomarkers for LC in order to improve patient outcomes. N6-methyladenosine (m6A) is the most common internal modification in eukaryotic mRNA and has been associated with various cancers, although its roles in the prognosis of LC remains to be elucidated. We analyzed the expression profiles of 15 m6A-related genes in the International Cancer Genome Consortium (ICGC) LIRI-JP dataset, and applied consensus clustering to stratify LC patients into two subgroups (Cluster 1 and Cluster 2). Cluster1 was significantly correlated to lower tumor stage and longer overall survival (OS). Gene set enrichment analysis showed that tumorigenic markers, including DNA repair, E2F targets, G2M checkpoint, and MYC targets V1, were enriched in Cluster2. We then constructed a prognostic risk model using three m6A-related genes that were identified as independent factors affecting OS. The nomogram based on the risk model score indicated good performance in predicting the 1-, 2- and 3-year survival of the LC patients. In conclusion, m6A-related genes are potential prognostic markers and therapeutic targets for LC.

Keywords: liver cancer, m6A, ICGC, epigenetic modification, prognosis

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Italian Institute of Technology (IIT), Italy

Reviewed by:
Zhuoxian Rong,
Central South University, China
Songyao Zhang,
Northwestern Polytechnical University, China
Jia Meng,
Xi'an Jiaotong-Liverpool University, China

***Correspondence:**
Chiyi He
hechiyi11@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

Received: *29 November 2019*
Accepted: *13 March 2020*
Published: *23 April 2020*

Citation:
Wang W, Sun B, Xia Y, Sun S and He C (2020) RNA N6-Methyladenosine-Related Gene Contribute to Clinical Prognostic Impact on Patients With Liver Cancer. Front. Genet. 11:306. doi: 10.3389/fgene.2020.00306

INTRODUCTION

Liver cancer (LC) is the fourth leading cause of cancer-related deaths worldwide (Villanueva, 2019). The etiology of LC differs geographically due to differences in the prevalence of risk factors (Jiang et al., 2019). For instance, chronic viral hepatitis infection is the most important risk factor in Asian countries, whereas non-viral factors are the major causative agents of LC in the Western countries (Yau et al., 2019). In East Asia, hepatitis B virus (HBV)- and hepatitis C virus (HCV)-related LC accounts for more than 80% of the cases (Liu et al., 2019a).

Liver tumorigenesis involves multiple steps with overlapping and interacting signaling pathways (Arzumanyan et al., 2013). However, the precise underlying mechanisms have not been completely elucidated so far. N6-methyladenosine (m6A), the most common internal post-transcriptional modification in eukaryotic mRNA, associates with many biological processes such as stress responses, stem cell differentiation, gametogenesis, and T Cell Homeostasis (Liu et al., 2019b; Zhou et al., 2019), and is mediated by factors that mainly include the "writers" (METTL3, METTL14, WTAP, RBM15, and ZC3H13), "readers" (YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, and HNRNPC), and "erasers" (FTO, ALKBH3, and ALKBH5) (Yang et al., 2018; Sun et al., 2019). Writers (m6A methyltransferase enzymes) and erasers (m6A demethylase enzymes) regulate the abundance, prevalence, and distribution of m6A, whereas readers (m6A binding proteins) modulate m6A modification-related mRNA processing, covering splicing, editing, localization, export, stability, translation, and decay (Zhou et al., 2019; Esteve-Puig et al., 2020). Recent reviews summarized that m6A-dependent mRNA regulation plays a crucial part in the development and progression of human cancers, such as HCC, acute myeloid leukemia (AML), glioblastoma, lung cancer, breast cancer, cervical cancer, and prostate cancers (Liu et al., 2019b; Esteve-Puig et al., 2020). Dysregulation of writers, readers, and erasers are pertinent to tumor initiation and progression, metastasis, and cancer drug resistance (Esteve-Puig et al., 2020). For example, METTL3 and FTO promote pathogenesis through stabilizing specific sets of mRNAs in breast cancer and AML, respectively (Tan et al., 2015; Vu et al., 2017). Similarly, alterations of readers such as YTHDC2 and YTHDF2 are related to colorectal cancer and hepatic cancer, respectively (Tanabe et al., 2016; Chen et al., 2018).

Hepatocellular carcinoma (HCC) is classified into different subclasses based on pathological characteristics and/or transcriptomes (Hoshida et al., 2009; Calderaro et al., 2017), and no study has so far reported prognostic subclasses of LC based on the expression of m6A-related genes. Since the prognosis of LC patients depends on the etiology and the ethnicity and/or geographical region (Hashimoto et al., 2017; Villanueva, 2019), and as East Asia has the highest incidence of LC (Bray et al., 2018), we therefore analyzed the m6A profile in an East Asian LC cohort (LIRI-JP dataset) from the International Cancer Genome Consortium (ICGC) database. The aim of this study was to determine the prognostic value of the m6A-related gene signature in LC.

MATERIALS AND METHODS

Datasets

The RNA sequencing data and corresponding clinicopathological information of LC patients were extracted from the ICGC (LIRI-JP dataset[1]) and The Cancer Genome Atlas (TCGA, LIHC dataset[2]) databases in May 2019. The gene expression data from TCGA was estimated as Transcripts Per Kilobase of the exon model per Million mapped reads (TPM). In the LIRI-JP dataset, the clinical stages of the patients were classified as per the Stage of Liver Cancer Study Group of Japan (LCSGJ) guidelines. The simple somatic mutation data was also retrieved for calculating the tumor mutation burden (TMB). The data of non-solid tissues and non-primary tumors, and of samples lacking sufficient clinical information were excluded. In case two or more samples were derived from the same patient, the mean value was used for analysis. Finally, 231 LC patients and 199 healthy controls from the LIRI-JP dataset, and 370 LC patients from the LIHC dataset were selected. The clinicopathological data of all patients are summarized in **Supplementary Table S1**.

Bioinformatics Analysis

Fifteen m6A-related genes were extracted from the LIRI-JP dataset (**Supplementary Table S2**). We analyzed the expression of 15 m6A-related genes in LC patients and normal tissue using the Limma package. LC patients were then clustered into different subgroups using the "Consensus Cluster Plus" package. In order to functionally annotate differentially expressed genes (DEGs) in different subgroups, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were conducted using the "clusterProfiler" package (Yu et al., 2012), and the Gene set variation analysis (GSVA) R package was used to analyze significant differences between the subgroups (Hanzelmann et al., 2013). Gene set enrichment analysis (GSEA) was used to identify the hallmarks of tumor sets in different LC subgroups (Liberzon et al., 2015).

The prognostic values of the m6A-related gene were determined by univariate Cox regression analysis in the LIRI-JP dataset in terms of hazard ratio (HR) and 95% confidence interval (CI). Six prognostic relevant genes ($P < 0.05$) were then used for the multivariable Cox analysis by step-wise forward and backward selection approaches as well as the smallest Akaike information criterion (AIC). Finally, a risk model was constructed using three genes, and the risk score (designated as riskScore) was calculated for each patient in the LIRI-JP and LIHC dataset using the formula: riskScore $= \mathrm{Coef}_{\mathrm{gene1}} \times \mathrm{Exp}_{\mathrm{gene1}} + \mathrm{Coef}_{\mathrm{gene2}} \times \mathrm{Exp}_{\mathrm{gene2}} + \mathrm{Coef}_{\mathrm{gene3}} \times \mathrm{Exp}_{\mathrm{gene3}}$, where Coef is the coefficient and Exp is the gene expression value. The clinicopathological factors and riskScore were used as variates in the univariate and multivariate Cox proportional hazards (PH) regression analyses to determine the independent predictive factors of overall survival (OS) in both datasets. A nomogram for 1-, 2-, and 3-year OS was then constructed based on the independent predictive factors, and its predictive performance was evaluated by C-index (Harrell et al., 1996). The calibration curve of the nomogram was used to assess the congruency between the predicted and actual survival. Bootstraps with 1,000 resamples were used to quantify model overfit, and a decision curve analysis (DCA) was made to evaluate the clinical efficacy (Vickers et al., 2008). The prediction power of the distinct parameters was determined using the area under receiver operating characteristic (ROC) curve (AUC) values.

Statistics

The expression level of 15 genes in the LC patients and controls was analyzed using the Wilcoxon rank sum test. The correlation between genes was determined by Pearson's analysis. Patients were divided into different groups by consensus analysis or riskScore (median value as the cutoff), and the distribution of clinical parameters between the subgroups was determined by Fisher's exact test. The OS of LC patients in the different subgroups was analyzed by the Kaplan–Meier method and

[1] https://icgc.org/

[2] http://cancergenome.nih.gov/

compared with the log-rank test. All statistical analyses were performed by R v3.6.0[3].

RESULTS

Differentially Expressed m6A-Related Genes Classify Liver Cancer Patients Into Distinct Clinical Clusters

Analysis of the expression patterns of m6A-related genes in the LIRI-JP dataset identified 14 DEGs in this study, including *KIAA1429*, *HNRNPC*, *METTL3*, *YTHDF3*, *YTHDF1*, *FTO*, *WTAP*, *YTHDF2*, *ALKBH5*, *ZC3H13*, *YTHDC2*, *ALKBH3*, *RBM15*, and *YTHDC1*. Among these DEGs, 13 genes were up-regulated including *KIAA1429*, *HNRNPC*, *METTL3*, *YTHDF3*, *YTHDF1*, *FTO*, *WTAP*, *YTHDF2*, *ALKBH5*, *YTHDC2*, *ALKBH3*, *RBM15*, and *YTHDC1*, while *ZC3H13* was down-regulated (**Figure 1A** and **Supplementary Table S2**). In addition, we analyzed the correlation among m6A-related genes. The KIAA1429 and YTHDF3 were highly correlated with each other, both of them were positively correlated with METTL14 and negatively correlated with ALKBH3, respectively. For "readers," YTHDF1 was positively correlated with YTHDF2, HNRNPC, and YTHDC1. For "writers," WTAP was positively correlated with RBM15, METTL3, and YTHDC1. For "erasers," FTO was positively correlated with ALKBH3 and ZC3H13, whereas ALKBH3, and ZC3H13 were negatively correlated with each other (**Figure 1B**). According to the consensus clustering analysis, the LC patients were divided into Cluster 1 (n = 138) and Cluster 2 (*n* = 93) (**Figure 2A** and **Supplementary Figure S1**). Then, we compared the clinical features of these two Clusters. Cluster 1 was significantly correlated with lower tumor stage ($P < 0.05$), but not with gender and age (**Figure 2B**). **Figure 2C** showed that prolonged overall survival (OS) in patients with Cluster 1, and the 3-year survival rates of Cluster 1 and Cluster 2 subgroups were 87.3 and 73.8%, respectively ($P < 0.05$). In addition, *YTHDF2* levels were significantly lower in stage 1 and 2 tumors compared to that in stages 3 and 4 ($P < 0.01$), while similar trends were not observed with *METTL3* and *YTHDC2* (**Supplementary Figure S2**). Then, we identified 761 DEGs between Cluster 1 and Cluster 2 with | fold change| > 1 and FDR < 0.05 as the criteria. GO and KEGG pathway analyses showed that these DEGs mainly participated in malignancy-related pathways, including PPAR signaling pathway, retinol metabolism, chemical carcinogenesis, and xenobiotics- and drug metabolism-related cytochrome P450 (**Figures 2D,E**). GSVA resulted in similar findings (**Figures 2F,G**). Furthermore, GSEA indicated that hallmarks of tumor sets were remarkably enriched in DNA repair (NES = 1.74, normalized $P < 0.05$), E2F targets (NES = 1.91, normalized $P < 0.05$), G2M checkpoint (NES = 1.91, normalized $P < 0.05$), and MYC targets V1 (NES = 1.82, normalized $P < 0.05$) in the Cluster 2 subgroup (**Figure 2H**).

Three m6A-Related Genes Form a Prognostic Risk Signature in Liver Cancer

Six m6A-related genes significantly correlated with OS by Univariate Cox analysis ($P < 0.05$), of which *METTL3*, *YTHDC2*, and *YTHDF2* were identified as independent predictors of OS and the coefficients were obtained by the multivariate analysis (**Table 1**). A risk model was constructed using these genes, and the riskScore was calculated for LC patients. Using the median riskScore as the cutoff value, we classified the LC patients into the high and low risk groups and observed poorer OS in the former ($P < 0.001$; **Figure 3A**). In addition, the

[3] https://www.r-project.org/

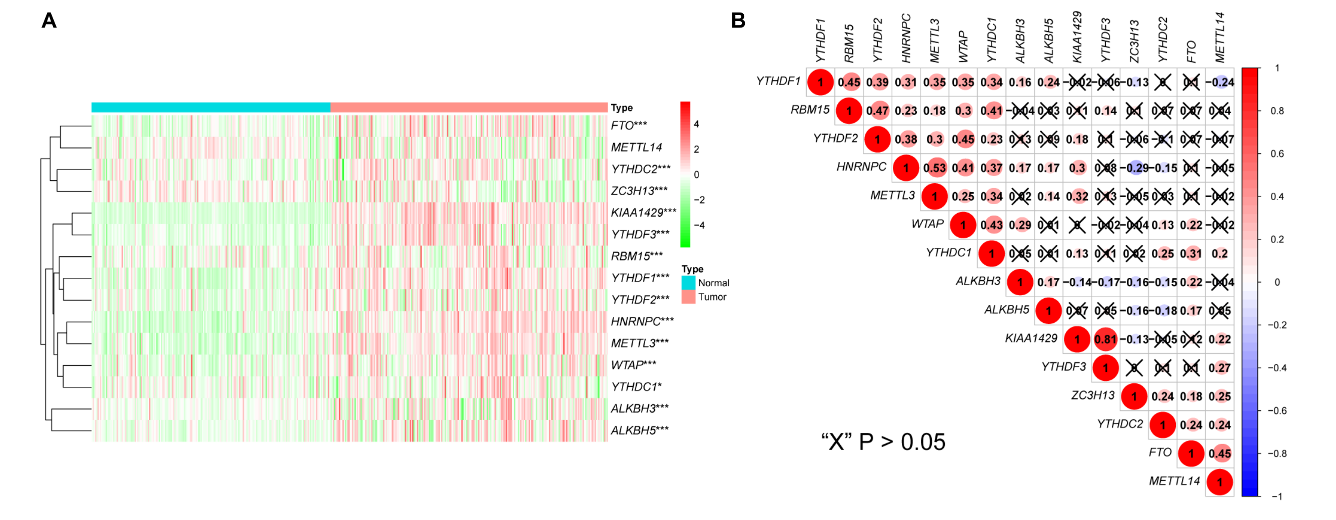


FIGURE 1 | Expression and correlation of m6A-related genes in liver cancer. **(A)** The expression levels of 15 m6A-related genes in liver cancer (Normal = 199, Tumor = 231). The heatmap shows the fold changes, with green indicates down-regulated genes, and red indicates up-regulated genes. **(B)** Pearson's correlation analysis of the 15 m6A-related genes. Blue indicates significant negative correlation and red indicates positive. **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

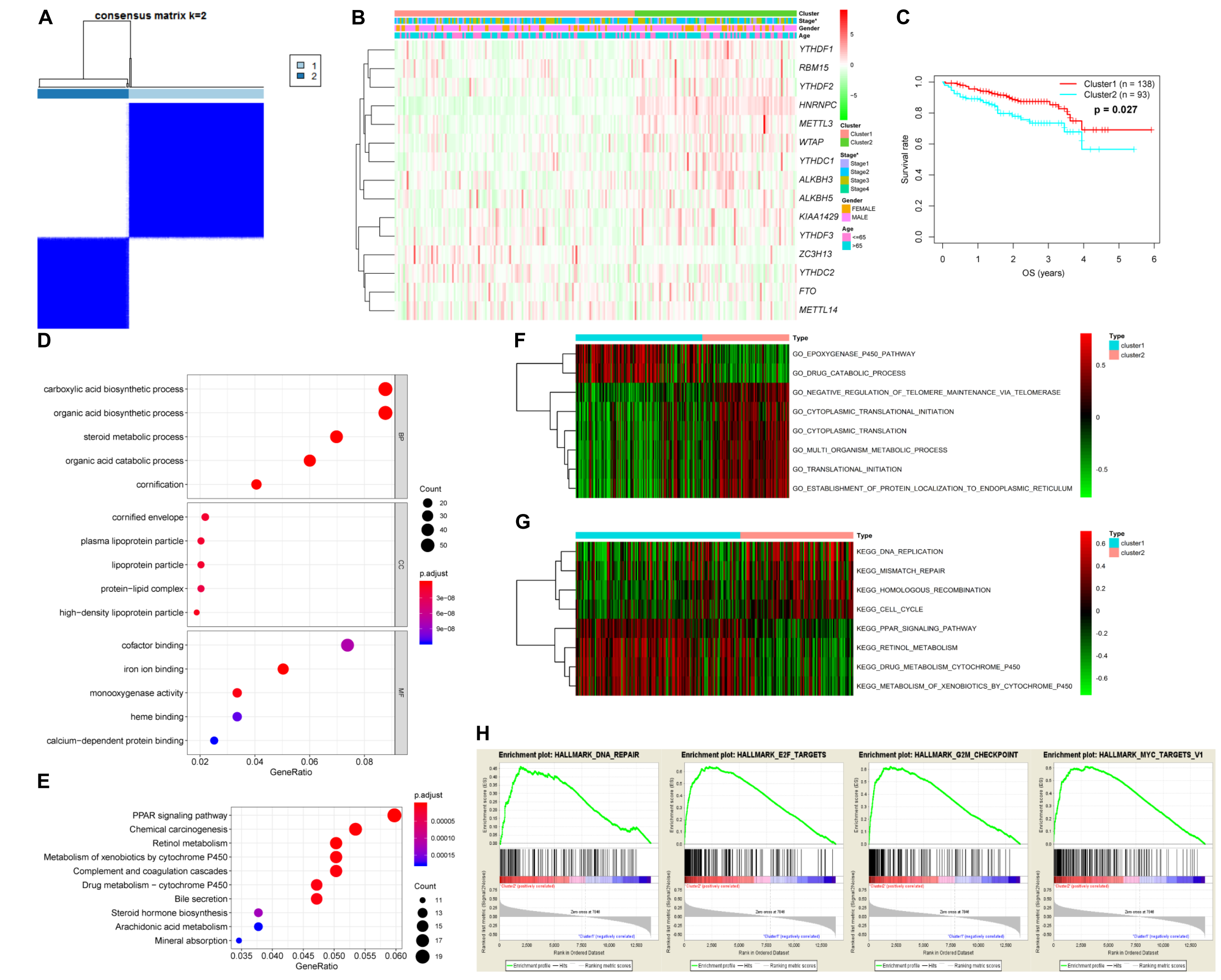


FIGURE 2 | Differential tumor stage and overall survival and functional annotation of liver cancer in Cluster 1 (n = 138) and Cluster 2 (n = 93) subgroups. **(A)** Consensus clustering matrix for k = 2. **(B)** Heatmap and clinicopathologic features of the two clusters defined by the m6A-related genes consensus expression. Green and red in the heat map indicate down-regulated and up-regulated genes, respectively. **(C)** Kaplan–Meier overall survival curves for liver cancer patients in LIRI-JP dataset. **(D,E)** Functional annotation of differentially expressed genes between Cluster 1 and Cluster 2 subgroups by GO terms **(D)** and KEGG pathway **(E)**. **(F,G)** GO terms **(F)** and KEGG pathway **(G)** significantly enriched in GSVA. **(H)** Genes with higher expression in Cluster 2 subgroup were enriched for hallmarks of tumor sets by GSEA.

risk subgroups differed significantly in terms of tumor stage (P < 0.01) and gene cluster (P < 0.001) (**Figure 3B**), but not age and gender in the LIRI-JP dataset (**Supplementary Figure S3**). The AUC values showed a better predictive ability of the riskScore for 3-year OS compared to the aforementioned parameters (**Figure 3C**). The multivariate analysis confirmed that riskScore, gender and stage were independent prognostic factors for the OS (stage and riskScore, P < 0.001; gender, P < 0.01) (**Figures 3D,E**). Furthermore, the female and Stage 4 LC patients had poorer prognosis compared to the male patients and those at other tumor stages, respectively (**Supplementary Figures S4A–C**). Interestingly, patients in the low risk subgroup stratified further by gender (female, P < 0.001; male, P < 0.05) or age (≤65, P < 0.01; >65, P < 0.05) had relatively longer OS compared to those in the high risk subgroup (**Figures 3F–I**), whereas the tumor stage was not affected by the riskScore (**Supplementary Figures S4D–G**).

Validation of the Risk Signature in TCGA Cohort

In the TCGA dataset, *METTL3*, *YTHDC2*, and *YTHDF2* were also significantly upregulated in the LC patients relative to the controls (**Supplementary Table S2**), and the riskScore was an independent prognostic factor for OS in this cohort (**Figure 4A**). We also stratified these patients into the high and low riskScore groups as with the LIRI-JP cohort, and observed significantly poorer prognosis in the former (P < 0.01) (**Figure 4B**). Furthermore, the low riskScore group had longer OS compared to the high riskScore group in the Asian cohort

TABLE 1 | Univariate and multivariate analyses of fifteen m6A-related genes in LIRI-JP dataset.

Gene	Univariate analysis			Multivariate analysis			
	HR	95% CI	*P*	Coef	HR	95% CI	*P*
ALKBH5	1.06	1.01–1.11	0.017	–	–	–	–
FTO	1.04	0.90–1.20	0.641	–	–	–	–
HNRNPC	1.02	1.01–1.03	0.005	–	–	–	–
KIAA1429	1.07	1.00–1.15	0.066	–	–	–	–
METTL14	0.92	0.81–1.05	0.211	–	–	–	–
METTL3	1.05	1.03–1.08	9.75E-05	0.039	1.04	1.01–1.07	0.004
RBM15	1.24	0.99–1.55	0.066	–	–	–	–
WTAP	1.03	1.00–1.07	0.101	–	–	–	–
YTHDC1	1.01	0.89–1.13	0.929	–	–	–	–
YTHDC2	0.81	0.70–0.95	0.009	-0.176	0.84	0.72–0.98	0.024
YTHDF1	1.05	1.00–1.10	0.034	–	–	–	–
YTHDF2	1.07	1.02–1.12	0.005	0.035	1.04	0.99–1.09	0.142
ZC3H13	0.98	0.94–1.01	0.155	–	–	–	–
ALKBH3	0.96	0.87–1.06	0.436	–	–	–	–
YTHDF3	1.00	0.96–1.05	0.939	–	–	–	–

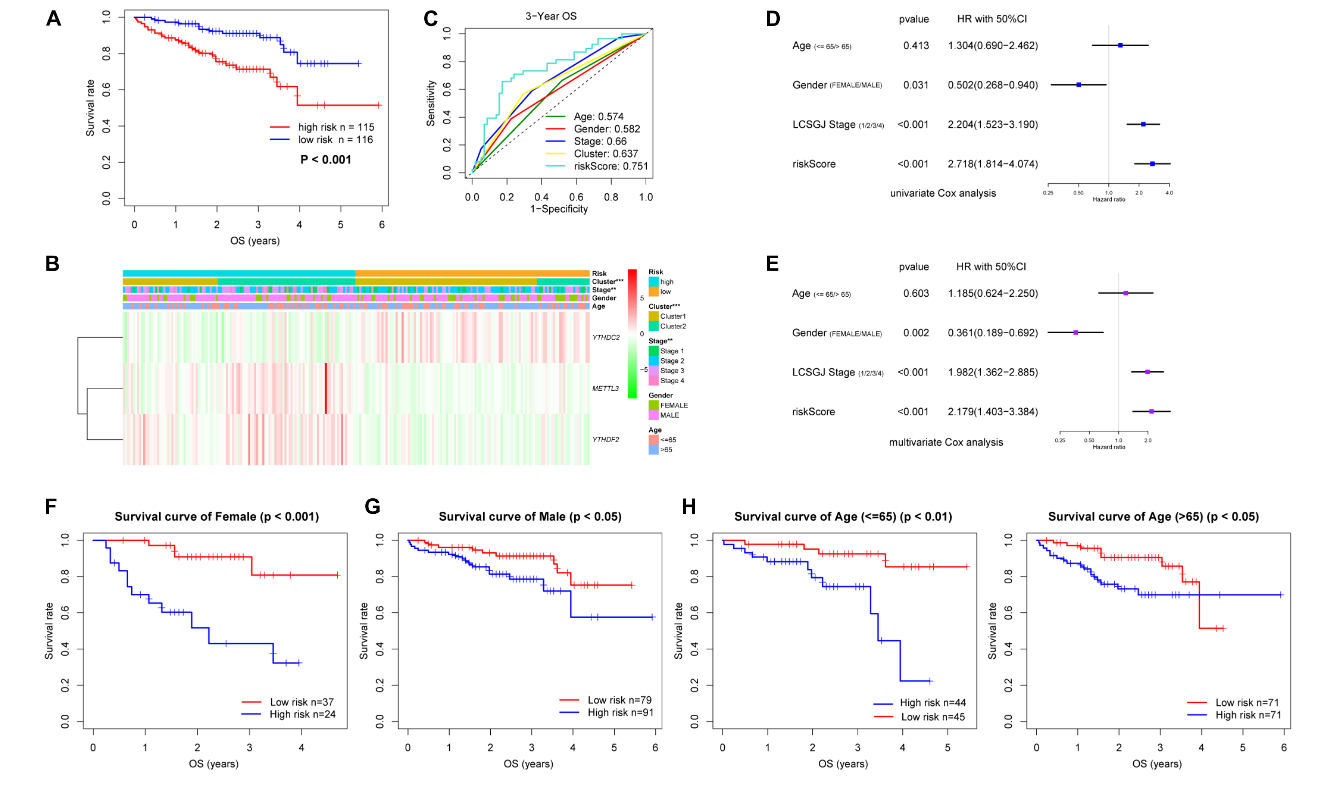


FIGURE 3 | Risk signature with three m6A-related genes in LIRI-JP dataset. **(A)** Kaplan–Meier overall survival curves for liver cancer patients classified into high and low risk groups based on the riskScore. **(B)** The differential clinicopathological features was compared between the high and low risk groups. Green and red in the heat map indicate down-regulated and up-regulated genes, respectively. **(C)** ROC curves displayed the predictive power of the riskScore, age, gender, tumor stage and cluster for the 3-year survival rate. **(D)** Univariate and **(E)** multivariate Cox regression analyses of the association between clinicopathological factors and overall survival. **(F–I)** Prognostic value of the riskScore stratified by **(F,G)** gender and **(H,I)** age. LCSGJ: Liver Cancer Study Group of Japan, $^{**}P < 0.01$, $^{***}P < 0.001$.

($P < 0.01$) (**Figure 4C**). Although there was no significant difference, a trend of better survival in the low risk group was observed in the non-Asian cohort (**Figure 4D**). The AUC values showed that the riskScore had moderate predictive ability for 1-, 2-, and 3-year OS in the TCGA dataset (**Supplementary Figure S5A**), and the AUC values in the

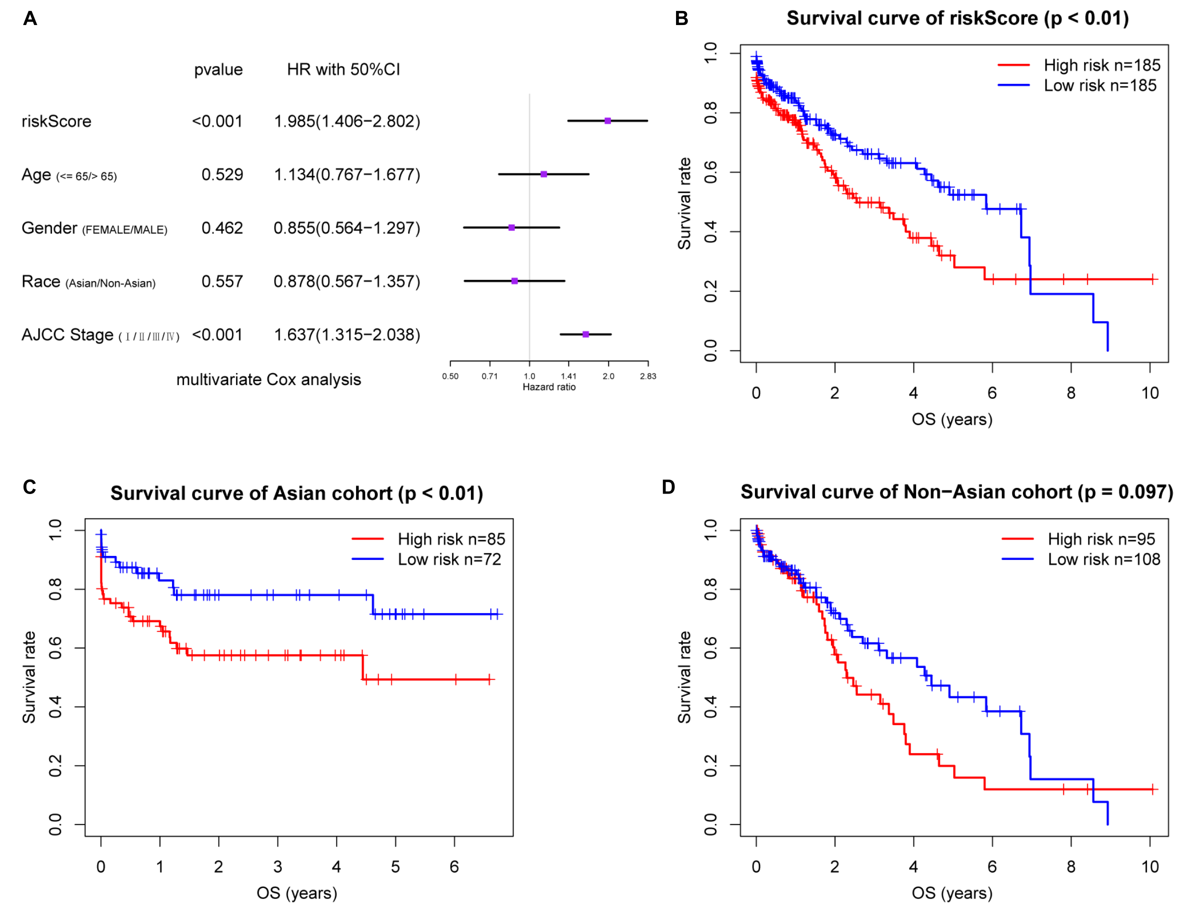


FIGURE 4 | Validation of the risk signature in TCGA cohort. **(A)** Multivariate Cox regression analysis of the association between clinicopathological factors and overall survival. Prognostic value of the riskScore in TCGA cohort **(B)** and stratified by **(C,D)** race. AJCC, American Joint Committee on Cancer.

Asian cohort were higher than those in the non-Asian cohort (**Supplementary Figures S5B,C**).

Construction and Validation of Nomogram

The 231 LC patients in the LIRI-JP dataset were arbitrarily separated into the training (n = 116) and validation cohort (n = 115) with a 5:5 split ratio (seeds = 100). In the training cohort, all independent prognostic factors were included in the predictive nomogram for OS (**Figure 5A**), and the points for each predictor are listed in **Supplementary Table S3**. The calibration curves indicated good congruency between the predicted and observed 3-year OS (**Figure 5B**). The Harrell's concordance-index (C-index) and 3-year AUC value of the nomogram were 0.797 and 0.822, respectively, which were higher compared to that of the riskScore, gender, or tumor stage (**Figure 5C** and **Table 2**). Similar outcomes were obtained in the validation as well as the entire cohort (**Figures 5D,E** and **Table 2**). In addition, DCA curves showed a greater threshold of the nomogram compared to the riskScore or tumor stage (**Figure 5F**), indicating that the nomogram has greater discriminatory capacity and accuracy for predicting survival compared to the other factors.

The m6A-Related Nomogram Has High Predictive Power

Immune checkpoint proteins including the programmed cell death protein 1 (PD-1/PDCD1), programmed death-ligand-1 (PD-L1/CD274), and cytotoxic T-lymphocyte associated antigen 4 (CTLA-4) are established prognostic markers for LC patients (Cariani and Missale, 2019; El Dika et al., 2019; Johnston and Khakoo, 2019). Recent studies showed that tumor mutation burden (TMB) is also significantly associated with the susceptibility to anti-tumor immunotherapy, and a higher TMB indicates better prognosis in many cancer types (Peng et al., 2019; Wang and Li, 2019). The m6A-related gene *YTHDF1* was also closely related to the prognosis of HCC in the TCGA dataset in a previous study (Zhao et al., 2018). We compared the AUC values of our established nomogram with these biomarkers, and found that the predictive power of the nomogram was superior for 1-, 2-, and 3-year OS in the LIRI-JP dataset (**Figure 6**). Finally, pathway enrichment analysis by Metascape[4] indicated that *METTL3*, *YTHDC2*, and *YTHDF2* and their 100 most strongly correlated co-expressed genes were enriched

[4] http://metascape.org/

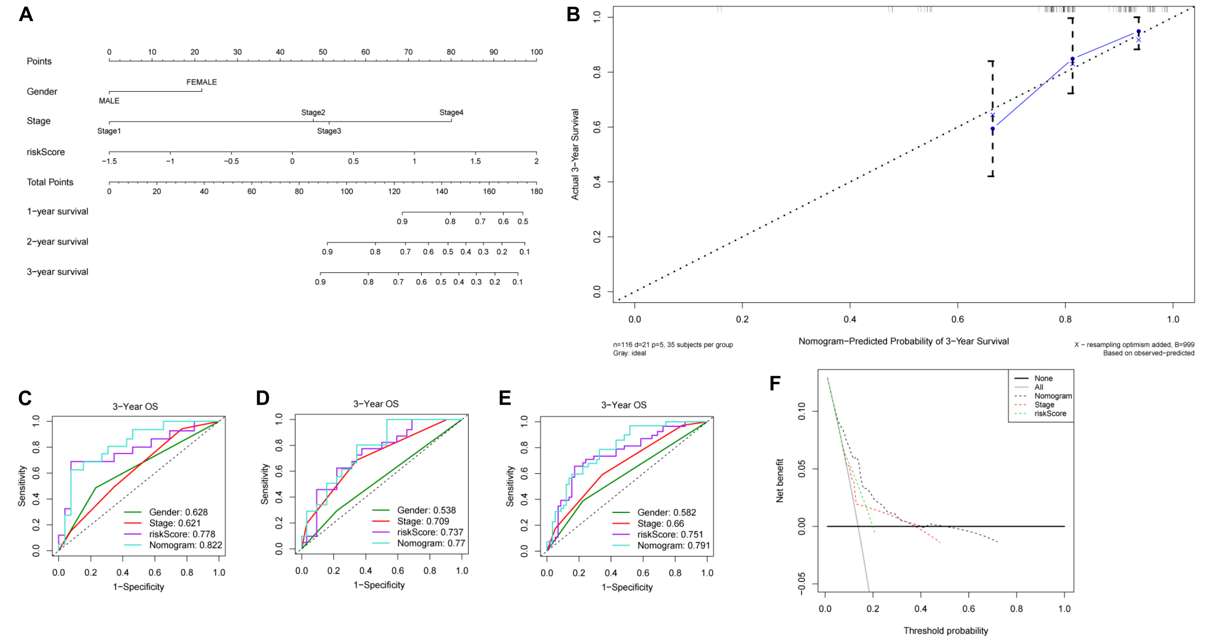


FIGURE 5 | Construction and validation of nomogram. **(A)** Nomogram predicting 1-, 2- and 3-year OS of patients with liver cancer. **(B)** Calibration plot for predicting patient OS at 3-year. ROC curves of the nomogram and clinicopathological factors for 3-year OS prediction in the **(C)** training cohort, **(D)** validation cohort and **(E)** entire cohort. **(F)** Decision curves of the nomogram, tumor stage and riskScore for predicting OS.

TABLE 2 | Comparison of C-index between the nomogram and other parameters in LIRI-JP cohort.

	C-index (95% confidence interval)		
	Training cohort	**Validation cohort**	**Entire cohort**
Nomogram	0.797 (0.715–0.879)	0.800 (0.718–0.882)	0.791 (0.732–0.850)
riskScore	0.709 (0.599–0.819)	0.714 (0.596–0.832)	0.706 (0.626–0.786)
Stage	0.667 (0.549–0.785)	0.729 (0.637–0.821)	0.699 (0.625–0.773)
Gender	0.591 (0.475–0.706)	0.566 (0.454–0.678)	0.579 (0.499–0.659)

for functions like mRNA processing, DNA repair, covalent chromatin modification, and regulation of the cell cycle, which are closely involved in tumorigenesis (**Figure 7**).

DISCUSSION

Although numerous genes and non-coding RNAs associated with LC progression have been identified in recent years (Tsuei et al., 2004; Yuan et al., 2014; Chua et al., 2015; Li et al., 2016; Mattu et al., 2016; Xiao et al., 2016; Zhang et al., 2016; Liu et al., 2017; Zhao et al., 2018; Zhou et al., 2019), the prognosis of the patients remains disappointing. Therefore, it is imperative to identify novel and reliable prognostic biomarkers or models in order to improve the clinical outcomes of LC patients. LC is a highly heterogenous cancer, and patient prognosis depends significantly on the geographical region and etiology. Chronic infection with the hepatitis virus is a major risk factor of LC in East Asia, whereas alcohol consumption and non-alcoholic fatty liver disease are the main causes in Western countries. We analyzed the gene expression data of East Asian LC patients from the ICGC LIRI-JP dataset in order to determine the prognostic potential of m6A-related genes in LC. We found that six m6A-related genes were significantly associated with the malignant progression and prognosis of LC, and a risk signature consisting of three of these genes was predictive of the prognosis.

We used consensus clustering to stratify the patients into two subgroups based on the expression of m6A-related genes, which showed significant differences in OS and the enriched pathways associated with tumor development and progression. The prognostic risk model also stratified the patients in the LIRI-JP cohort into two groups based on the 3-gene riskScore, which showed greater predictive performance compared to single clinical indicators. Multivariate Cox analysis revealed that the riskScore was an independent prognostic factor for LC in the LIRI-JP and LIHC datasets. The nomogram, constructed using the riskScore and clinicopathological features, further increased the predictive power for OS compared to the riskScore, immunotherapy-related genes, or TMB alone. Interestingly, in the TCGA dataset, the riskScore was able to make a distinction for the OS in the Asian cohort, but not in the non-Asian cohort. This difference may be due to the fact that risk factors for LC differ across ethnicities.

The three genes (*YTHDC2*, *YTHDF2*, and *METTL3*) incorporated in the prognosis risk model were upregulated in the LC patients in both LIRI-JP and LIHC datasets, which are similar to those of previous studies (Yuan et al., 2014; Chen et al., 2018). *YTHDF2* and *METTL3* have previously been reported as tumor suppressors in HCC, and as oncogenes in pancreatic cancer and acute myelocytic leukemia (Cui et al., 2017; Wang et al., 2017;

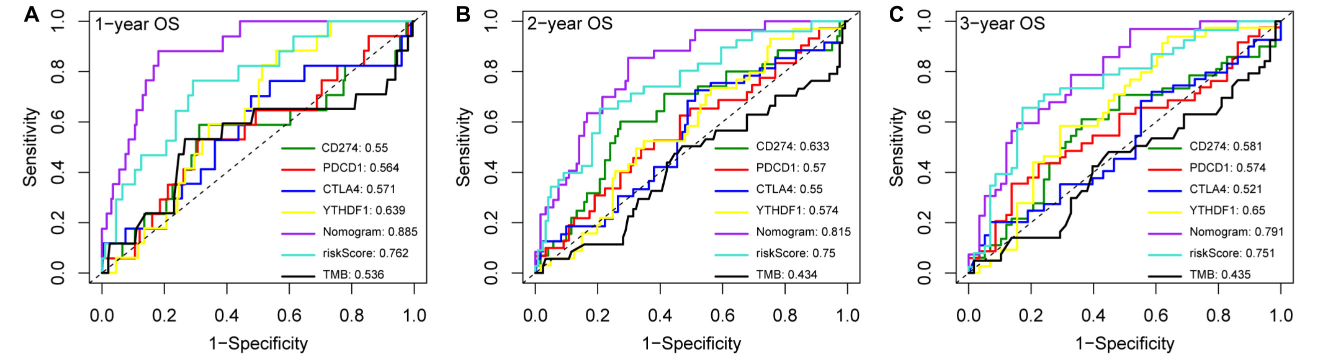


FIGURE 6 | Compare the AUC values of the nomogram with different biomarkers. ROC curves of the nomogram and different biomarkers for **(A)** 1-, **(B)** 2- and **(C)** 3-year overall survival prediction in LIRI-JP cohort.

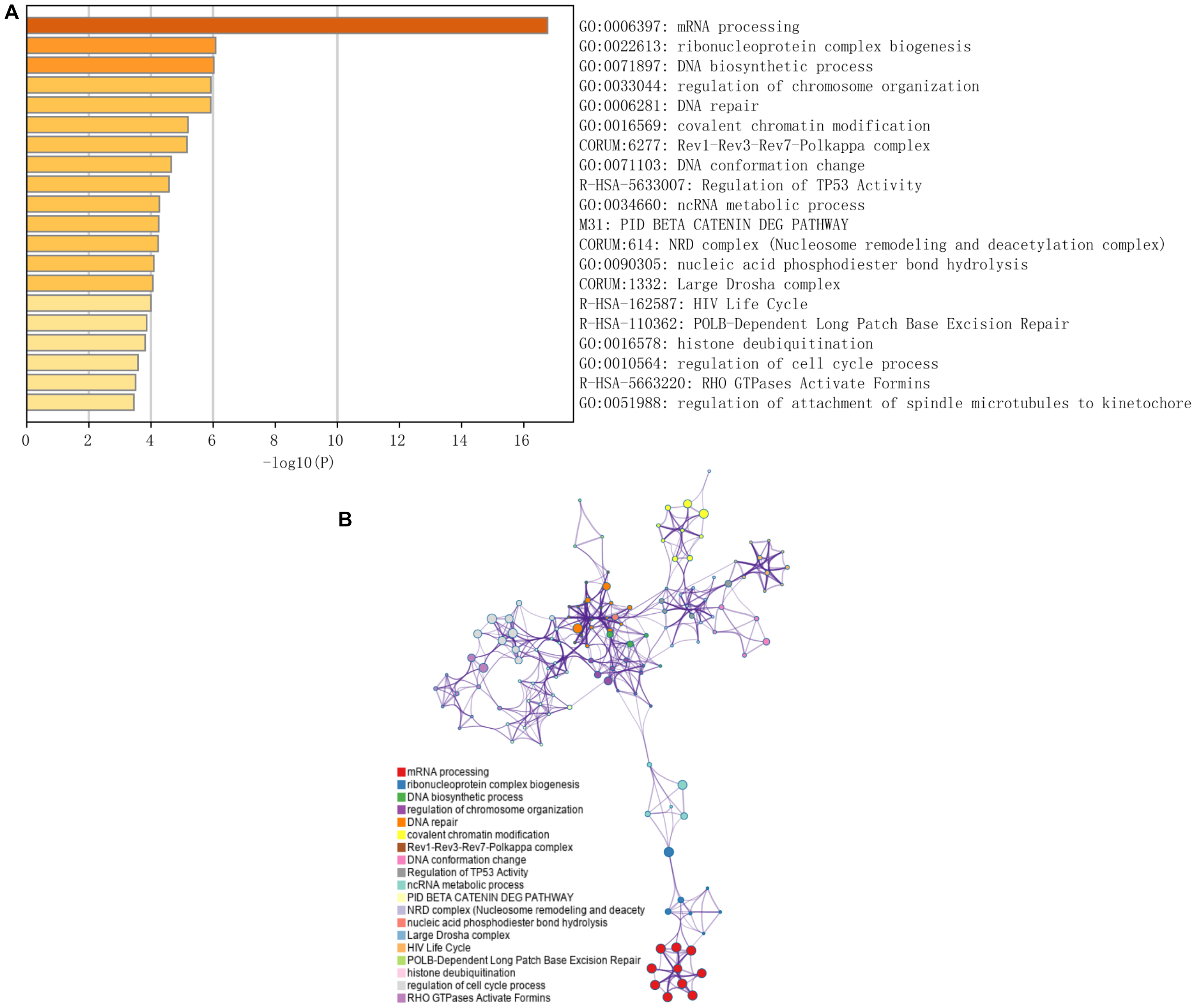


FIGURE 7 | Functional prediction of three m6A-related genes involved in the risk signature. **(A)** Significantly enriched pathways of the three genes and their coexpressed genes. **(B)** The map of functional enriched pathways. Each node represents a GO term. Node size represents the number of gene in the pathway. Different colors represent different pathways.

Zhong et al., 2019). Chen et al. (2018) demonstrated that overexpression of *METTL3* in HCC patients have poor prognosis. Further, knockout of *METTL3* suppresses HCC tumorigenicity and lung metastasis by modulation of cytokine signaling 2 through a *YTHDF2*-dependent mechanism (Chen et al., 2018). Zhong et al. (2019) demonstrated that *YTHDF2* acts as an HCC suppressor *via* promoting the degradation of epidermal growth factor receptor mRNA. Hou et al. (2019) reported that a high expression of *YTHDF2* gives rise to a better prognosis of HCC patients, and represses tumor growth and angiogenesis by degradation of interleukin 11 and serpin family E member 2 mRNAs. Tanabe et al. (2014) reported that *YTHDC2* acts as a tumor suppressor in the LC cell line by perhaps recruiting c-Jun and activating transcription factor 2 to the *YTHDC2* promoter. The above three m6A-related genes may affect HCC growth and metastasis by regulating the stability of multiple target genes.

Recent studies showed that m6A-related genes could be potential prognostic markers for predicting patient survival in a variety of cancers. These genes significantly correlated and interacted with each other which indicated that the cross-talk exists among the m6A-related genes (Li et al., 2019). Because of a complex reciprocal regulatory relationship among the m6A-related genes, it seems necessary to analyze prognostic and predictive values using a signature comprised of multiple m6A-related genes in patients with distinct tumor types. Kandimalla et al. (2019) reported that a gene expression signature consisted of seven m6A-related regulators characterized as a robust prognostic and predictive signature in 13 human cancers including HCC (relapse-free survival). This study offered a landscape of the biological and clinical characteristics pertaining to m6A machinery in tumor patients (Kandimalla et al., 2019). However, the external validation cohort was applied to colorectal cancer, gastric cancer, breast cancer, and ovarian cancer, but not to HCC for survival analysis. In our study, we successfully established a prognostic signature comprised of three m6A-related genes for predicting survival of HCC patient, using an additional RNA-seq dataset as external validation avoiding biased results to some extent.

There were some limitations in this study. First, an additional LC patient cohort for a prognostic study was needed to validate the predictive power of our prognostic signature in the future. Second, experimental studies that focus on the molecular mechanisms remain necessary to investigate the functions of these m6A-related genes in LC.

In summary, m6A-related genes have a prognostic value in LC, and the constructed riskScore can identify patients who are high risk and can enable individualized therapy. Our findings have to be validated in larger cohorts, and further studies are also needed to elucidate the mechanism of these m6A-related genes in LC.

DATA AVAILABILITY STATEMENT

The data used to support the results of this study are from the public databases ICGC (International Cancer Genome Consortium, https://icgc.org/) and TCGA (The Cancer Genome Atlas, https://cancergenome.nih.gov/).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

CH designed the research study. WW, BS, YX, and SS analyzed the data and performed the data analysis. WW wrote the manuscript and interpreted the data. CH helped to revise the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Wannan Medical College (WK2017F20), Yijishan Hospital of Wannan Medical College (YR201608), and Key Research and Development Plan Projects in Anhui Province (No. 201904a07020028).

ACKNOWLEDGMENTS

The authors acknowledge contributions from ICGC and TCGA databases.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00306/full#supplementary-material

REFERENCES

Arzumanyan, A., Reis, H. M., and Feitelson, M. A. (2013). Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. *Nat. Rev. Cancer* 13, 123–135. doi: 10.1038/nrc3449

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Calderaro, J., Couchy, G., Imbeaud, S., Amaddeo, G., Letouze, E., Blanc, J. F., et al. (2017). Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. *J. Hepatol.* 67, 727–738. doi: 10.1016/j.jhep.2017.05.014

Cariani, E., and Missale, G. (2019). Immune landscape of hepatocellular carcinoma microenvironment: implications for prognosis and therapeutic applications. *Liver Int.* 39, 1608–1621. doi: 10.1111/liv.14192

Chen, M., Wei, L., Law, C. T., Tsang, F. H., Shen, J., Cheng, C. L., et al. (2018). RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2. *Hepatology* 67, 2254–2270. doi: 10.1002/hep.29683

Chua, H. H., Tsuei, D. J., Lee, P. H., Jeng, Y. M., Lu, J., Wu, J. F., et al. (2015). RBMY, a novel inhibitor of glycogen synthase kinase 3beta, increases tumor stemness and predicts poor prognosis of hepatocellular carcinoma. *Hepatology* 62, 1480–1496. doi: 10.1002/hep.27996

Cui, Q., Shi, H., Ye, P., Li, L., Qu, Q., Sun, G., et al. (2017). m(6)A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep.* 18, 2622–2634. doi: 10.1016/j.celrep.2017.02.059

El Dika, I., Khalil, D. N., and Abou-Alfa, G. K. (2019). Immune checkpoint inhibitors for hepatocellular carcinoma. *Cancer* 125, 3312–3319. doi: 10.1002/cncr.32076

Esteve-Puig, R., Bueno-Costa, A., and Esteller, M. (2020). Writers, readers and erasers of RNA modifications in cancer. *Cancer Lett.* 474, 127–137. doi: 10.1016/j.canlet.2020.01.021

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Harrell, F. E. Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15, 361–387. doi: 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4

Hashimoto, M., Tashiro, H., Kobayashi, T., Kuroda, S., Hamaoka, M., and Ohdan, H. (2017). Clinical characteristics and prognosis of non-B, non-C hepatocellular carcinoma: the impact of patient sex on disease-free survival - A retrospective cohort study. *Int. J. Surg.* 39, 206–213. doi: 10.1016/j.ijsu.2017.01.110

Hoshida, Y., Nijman, S. M., Kobayashi, M., Chan, J. A., Brunet, J. P., Chiang, D. Y., et al. (2009). Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* 69, 7385–7392. doi: 10.1158/0008-5472.can-09-1089

Hou, J., Zhang, H., Liu, J., Zhao, Z., Wang, J., Lu, Z., et al. (2019). YTHDF2 reduction fuels inflammation and vascular abnormalization in hepatocellular carcinoma. *Mol. Cancer* 18, 163. doi: 10.1186/s12943-019-1082-3

Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., et al. (2019). Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 567, 257–261. doi: 10.1038/s41586-019-0987-8

Johnston, M. P., and Khakoo, S. I. (2019). Immunotherapy for hepatocellular carcinoma: current and future. *World J. Gastroenterol.* 25, 2977–2989. doi: 10.3748/wjg.v25.i24.2977

Kandimalla, R., Gao, F., Li, Y., Huang, H., Ke, J., Deng, X., et al. (2019). RNAMethyPro: a biologically conserved signature of N6-methyladenosine regulators for predicting survival at pancancer level. *NPJ. Precis Oncol.* 3:13. doi: 10.1038/s41698-019-0085-2

Li, T., Xie, J., Shen, C., Cheng, D., Shi, Y., Wu, Z., et al. (2016). Upregulation of long noncoding RNA ZEB1-AS1 promotes tumor metastasis and predicts poor prognosis in hepatocellular carcinoma. *Oncogene* 35, 1575–1584. doi: 10.1038/onc.2015.223

Li, Y., Xiao, J., Bai, J., Tian, Y., Qu, Y., Chen, X., et al. (2019). Molecular characterization and clinical relevance of m6A regulators across 33cancer types. *Mol. Cancer* 18:137. doi: 10.1186/s12943-019-1066-3

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., Tamayo, P., et al. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 23, 417–425. doi: 10.1016/j.cels.2015.12.004

Liu, J., Chen, S., Zou, Z., Tan, D., Liu, X., and Wang, X. (2019a). Pathological Pattern of Intrahepatic HBV in HCC is phenocopied by PDX-derived mice: a novel model for antiviral treatment. *Transl. Oncol.* 12, 1138–1146. doi: 10.1016/j.tranon.2019.05.006

Liu, J., Harada, B. T., and He, C. (2019b). Regulation of gene expression by N6-methyladenosine in Cancer. *Trends Cell Biol.* 29, 487–499. doi: 10.1016/j.tcb.2019.02.008

Liu, J., Lu, C., Xiao, M., Jiang, F., Qu, L., and Ni, R. (2017). Long non-coding RNA SNHG20 predicts a poor prognosis for HCC and promotes cell invasion by regulating the epithelial-to-mesenchymal transition. *Biomed. Pharmacother.* 89, 857–863. doi: 10.1016/j.biopha.2017.01.011

Mattu, S., Fornari, F., Quagliata, L., Perra, A., Angioni, M. M., Petrelli, A., et al. (2016). The metabolic gene HAO2 is downregulated in hepatocellular carcinoma and predicts metastasis and poor survival. *J. Hepatol.* 64, 891–898. doi: 10.1016/j.jhep.2015.11.029

Peng, H., Zhang, Y., Zhou, Z., Guo, Y., Huang, X., Westover, K. D., et al. (2019). Intergrated analysis of ELMO1, serves as a link between tumour mutation burden and epithelial-mesenchymal transition in hepatocellular carcinoma. *EBiomedicine* 46, 105–118. doi: 10.1016/j.ebiom.2019.07.002

Sun, T., Wu, R., and Ming, L. (2019). The role of m6A RNA methylation in cancer. *Biomed. Pharmacother.* 112:108613. doi: 10.1016/j.biopha.2019.108613

Tan, A., Dang, Y., Chen, G., and Mo, Z. (2015). Overexpression of the fat mass and obesity associated gene (FTO) in breast cancer and its clinical implications. *Int. J. Clin. Exp. Pathol.* 8, 13405–13410. doi: 10.1142/S0218127409025353

Tanabe, A., Konno, J., Tanikawa, K., and Sahara, H. (2014). Transcriptional machinery of TNF-alpha-inducible YTH domain containing 2 (YTHDC2) gene. *Gene* 535, 24–32. doi: 10.1016/j.gene.2013.11.005

Tanabe, A., Tanikawa, K., Tsunetomi, M., Takai, K., Ikeda, H., Konno, J., et al. (2016). RNA helicase YTHDC2 promotes cancer metastasis via the enhancement of the efficiency by which HIF-1α mRNA is translated. *Cancer Lett.* 376, 34–42. doi: 10.1016/j.canlet.2016.02.022

Tsuei, D. J., Hsu, H. C., Lee, P. H., Jeng, Y. M., Pu, Y. S., Chen, C. N., et al. (2004). RBMY, a male germ cell-specific RNA-binding protein, activated in human liver cancers and transforms rodent fibroblasts. *Oncogene* 23, 5815–5822. doi: 10.1038/sj.onc.1207773

Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med. Inform. Decis. Maker.* 8:53. doi: 10.1186/1472-6947-8-53

Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380, 1450–1462. doi: 10.1056/NEJMra1713263

Vu, L. P., Pickering, B. F., Cheng, Y., Zaccara, S., Nguyen, D., Minuesa, G., et al. (2017). The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat. Med.* 23, 1369–1376. doi: 10.1038/nm.4416

Wang, S., Sun, C., Li, J., Zhang, E., Ma, Z., Xu, W., et al. (2017). Roles of RNA methylation by means of N(6)-methyladenosine (m(6)A) in human cancers. *Cancer Lett.* 408, 112–120. doi: 10.1016/j.canlet.2017.08.030

Wang, X., and Li, M. (2019). Correlate tumor mutation burden with immune signatures in human cancers. *BMC Immunol.* 20:4. doi: 10.1186/s12865-018-0285-5

Xiao, S., Chang, R. M., Yang, M. Y., Lei, X., Liu, X., Gao, W. B., et al. (2016). Actin-like 6A predicts poor prognosis of hepatocellular carcinoma and promotes metastasis and epithelial-mesenchymal transition. *Hepatology* 63, 1256–1271. doi: 10.1002/hep.28417

Yang, Y., Hsu, P. J., Chen, Y. S., and Yang, Y. G. (2018). Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* 28, 616–624. doi: 10.1038/s41422-018-0040-8

Yau, T., Hsu, C., Kim, T. Y., Choo, S. P., Kang, Y. K., Hou, M. M., et al. (2019). Nivolumab in advanced hepatocellular carcinoma: Sorafenib-experienced Asian cohort analysis. *J. Hepatol.* 71, 543–552. doi: 10.1016/j.jhep.2019.05.014

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yuan, J. H., Yang, F., Wang, F., Ma, J. Z., Guo, Y. J., Tao, Q. F., et al. (2014). A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25, 666–681. doi: 10.1016/j.ccr.2014.03.010

Zhang, D., Cao, C., Liu, L., and Wu, D. (2016). Up-regulation of LncRNA SNHG20 predicts poor prognosis in hepatocellular carcinoma. *J. Cancer* 7, 608–617. doi: 10.7150/jca.13822

Zhao, X., Chen, Y., Mao, Q., Jiang, X., Jiang, W., Chen, J., et al. (2018). Overexpression of YTHDF1 is associated with poor prognosis in patients with hepatocellular carcinoma. *Cancer Biomark.* 21, 859–868. doi: 10.3233/cbm-170791

Zhong, L., Liao, D., Zhang, M., Zeng, C., Li, X., Zhang, R., et al. (2019). YTHDF2 suppresses cell proliferation and growth via destabilizing the EGFR mRNA in hepatocellular carcinoma. *Cancer Lett.* 442, 252–261. doi: 10.1016/j.canlet.2018.11.006

Zhou, Y., Yin, Z., Hou, B., Yu, M., Chen, R., Jin, H., et al. (2019). Expression profiles and prognostic significance of RNA N6-methyladenosine-related genes in patients with hepatocellular carcinoma: evidence from independent datasets. *Cancer Manag. Res.* 11, 3921–3931. doi: 10.2147/CMAR.S191565

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 28 April 2020
doi: 10.3389/fgene.2020.00394

Direct RNA Sequencing for the Study of Synthesis, Processing, and Degradation of Modified Transcripts

*Mattia Furlan[1,2], Iris Tanaka[1], Tommaso Leonardi[1], Stefano de Pretis[1] and Mattia Pelizzola[1]**

[1] Center for Genomic Science, Istituto Italiano di Tecnologia, Milan, Italy, [2] Department of Physics, National Institute of Nuclear Physics, University of Turin, Turin, Italy

It has been known for a few decades that transcripts can be marked by dozens of different modifications. Yet, we are just at the beginning of charting these marks and understanding their functional impact. High-quality methods were developed for the profiling of some of these marks, and approaches to finely study their impact on specific phases of the RNA life-cycle are available, including RNA metabolic labeling. Thanks to these improvements, the most abundant marks, including N^6-methyladenosine, are emerging as important determinants of the fate of marked RNAs. However, we still lack approaches to directly study how the set of marks for a given RNA molecule shape its fate. In this perspective, we first review current leading approaches in the field. Then, we propose an experimental and computational setup, based on direct RNA sequencing and mathematical modeling, to decipher the functional consequences of RNA modifications on the fate of individual RNA molecules and isoforms.

Keywords: RNA modification, m^{6}A, direct RNA sequencing, metabolic labeling, nascent RNA, RNA metabolism, long reads sequencing, nanopore

OPEN ACCESS

Edited by:
Zhengqing Ouyang,
University of Massachusetts Amherst,
United States

Reviewed by:
Jianzhen Xu,
Shantou University, China
Chao Cheng,
ABLife Inc., China

Correspondence:
Mattia Pelizzola
mattia.pelizzola@iit.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 November 2019*
Accepted: *30 March 2020*
Published: *28 April 2020*

Citation:
Furlan M, Tanaka I, Leonardi T, de Pretis S and Pelizzola M (2020) Direct RNA Sequencing for the Study of Synthesis, Processing, and Degradation of Modified Transcripts. Front. Genet. 11:394. doi: 10.3389/fgene.2020.00394

INTRODUCTION

More than a 100 RNA modifications have been identified since the 1950s (Boccaletto et al., 2018). They were first observed in abundant populations of non-coding transcripts (e.g., tRNAs) and in a second moment, due to the improvement of profiling techniques, their pervasive presence was confirmed in coding transcripts (Roundtree et al., 2017). Different modifications were found to co-occur on the same RNA molecule (Jackman and Alfonzo, 2013). In some cases, rather than a mere stochastic effect due to the modification frequency, their co-occurrence suggested reciprocal regulation mechanisms (Xiang et al., 2018).

The N^6-methyladenosine (m^{6}A) emerged as one of the most abundant modifications of coding transcripts (Roundtree et al., 2017), and it was shown to be involved in the regulation of various biological processes, including cellular differentiation (Lin and Gregory, 2014; Wang Y. et al., 2014; Chen et al., 2015; Geula et al., 2015; Zhang et al., 2017a), meiosis (Bushkin et al., 2019), heat stress response (Zhou et al., 2015), gametogenesis (Wojtas et al., 2017), and neurons activity (Engel et al., 2018). Furthermore, aberrant m^{6}A patterning was shown to be associated with diseases insurgence and progression (Tong et al., 2018; Ianniello et al., 2019; Yang et al., 2019). A number of effectors were identified that are responsible for m^{6}A deposition (e.g., METTL3 and METTL14) (Liu et al., 2014; Ping et al., 2014; Schwartz et al., 2014), recognition (e.g., members of the YTH domain family) (Luo and Tong, 2014; Xu et al., 2014; Zhu et al., 2014; Xiao et al., 2016), and removal (FTO and ALKBH5) (Jia et al., 2011; Zheng et al., 2013), suggesting that this mark could be dynamically regulated. Genome-wide m^{6}A profiling, through immunoprecipitation with m^{6}A-specific antibodies followed by short-reads RNA sequencing (srRNA-seq), revealed the preferential, while not exclusive, association of the mark with the central adenosine in the RRACH sequence context around the stop codon of messenger RNAs (R = G or A and H = A, C, or U) (Dominissini et al., 2012; Meyer et al., 2012). Notably, m^{6}A marks have been linked to different biological processes depending on their relative position within a transcript, suggesting a context-specific role for this mark (Shi et al., 2019). However, we have only started revealing the rules that determine the preference of the mark for specific bases, and their impact on specific downstream biological processes (Yue et al., 2018). Altogether, m^{6}A was identified as a key determinant of RNA decay (Wang X. et al., 2014) and translation (Wang et al., 2015), while discordant reports were published about its involvement in splicing regulation (Haussmann et al., 2016; Xiao et al., 2016; Bartosovic et al., 2017; Ke et al., 2017; Darnell et al., 2018; Kasowitz et al., 2018; Louloupi et al., 2018).

RNA metabolic labeling (Dolken et al., 2008) emerged as a powerful approach that not only allows to characterize the association of m^{6}A, or other RNA modifications, with nascent transcripts, but also allows to quantify the impact of these marks on the dynamics of all key steps of the RNA life cycle, and specifically on the kinetic rates of RNA synthesis, processing, and degradation. The application of this technique confirmed the role of m^{6}A on the regulation of RNA stability, and suggested its influence on the dynamics of RNA synthesis and processing (Furlan et al., 2019b).

The application of the current leading approaches for profiling RNA modifications, such as m^{6}A, generated important findings about the functional role of these marks (Roundtree et al., 2017). However, these approaches are heavily based on srRNA-seq, and are afflicted by a number of downsides: different methods were developed for various modifications, they only allow to indirectly map the targeted mark, they are poorly suitable for analyses at the level of single molecules and isoforms, they cannot be readily used to profile co-occurring modifications, and they are difficult to be paired with RNA metabolic labeling. In this perspective, we discuss how direct RNA sequencing (such as nanopore-based sequencing of native RNAs) is rapidly emerging as a powerful alternative approach, which has the potential to overcome these issues, bursting the field of epitranscriptomics.

EXPERIMENTAL AND COMPUTATIONAL APPROACHES FOR THE QUANTIFICATION OF RNA KINETIC RATES

The state of the art approach to infer the kinetic rates governing the RNA life cycle – synthesis of premature RNA, its processing into mature RNA, and the degradation of the latter – is based on the joint quantitative analysis of total and nascent RNA (**Figure 1**). While the former is simply obtained through RNA-seq, the latter can be profiled through RNA metabolic labeling. In this technique, a nucleotide carrying an exogenous modification (e.g., 4-thiouridine, 4sU) is provided in the cells' medium, and is incorporated into nascent transcripts during the labeling time. Thus, the presence of the exogenous modification can be used for the physical (Dolken et al., 2008) or *in silico* (Baptista and Dölken, 2018) separation of newly synthetized transcripts from pre-existing ones.

Mathematical modeling is then used for the gene-level quantification of RNA kinetic rates, for example as implemented and documented in the INSPEcT R/Bioconductor library (de Pretis et al., 2015; Furlan et al., 2019a). Briefly, when short labeling times are adopted (<1 h), the quantification of nascent RNA for each gene provides a proxy for the rate of synthesis of premature RNA. Then, total RNA-seq reads are used to measure the abundance of premature and mature transcripts: reads that entirely map to one or more exons are used to quantify mature RNA species, and the remaining mapped reads (entirely, or partially, covering introns) are used for the quantification of premature species. Finally, the combination of synthesis rate and premature RNA abundance is used to quantify the rate of processing, while the combination of synthesis rate and mature RNA abundance allows the quantification of degradation rates (Furlan et al., 2019a).

The joint analysis of the information gained from RNA metabolic labeling experiments, together with the profiling of specific RNA modifications, would be extremely powerful for the study of the functional consequences of these marks on specific RNA life cycle steps. However, while the application of metabolic labeling for the profiling of nascent RNA (Dolken et al., 2008) and for the quantification of the RNA kinetic rates (Dolken et al., 2008; Miller et al., 2011; Rabani et al., 2011, 2014; de Pretis et al., 2015; Furlan et al., 2019a) is an established approach, its combination with the profiling of RNA modifications is more problematic. In fact, the joint profiling of nascent and modified RNA requires the identification of at least two RNA modifications: the endogenous mark (e.g., m^{6}A), and the exogenous modification used for the labeling (e.g., 4sU). As we discuss in the following sections, this is a complex task that can be only indirectly implemented through current approaches.

DETECTION OF RNA MODIFICATIONS THROUGH SHORT-READS RNA SEQUENCING

Numerous protocols based on srRNA-seq were developed for the identification of either endogenous (e.g., m^{6}A) or exogenous (e.g., 4sU) RNA modifications. A first class of methods is based on the enrichment of modified RNAs before the sequencing. This relies either on the use of specific antibodies [e.g., MeRIP-seq for m^{6}A detection (Dominissini et al., 2012; Meyer et al., 2012)], or the use of enzymes involved in the metabolism of the modification [e.g., tRNA methyltransferase DnmA (Muller et al., 2013)], or on the availability of tags such

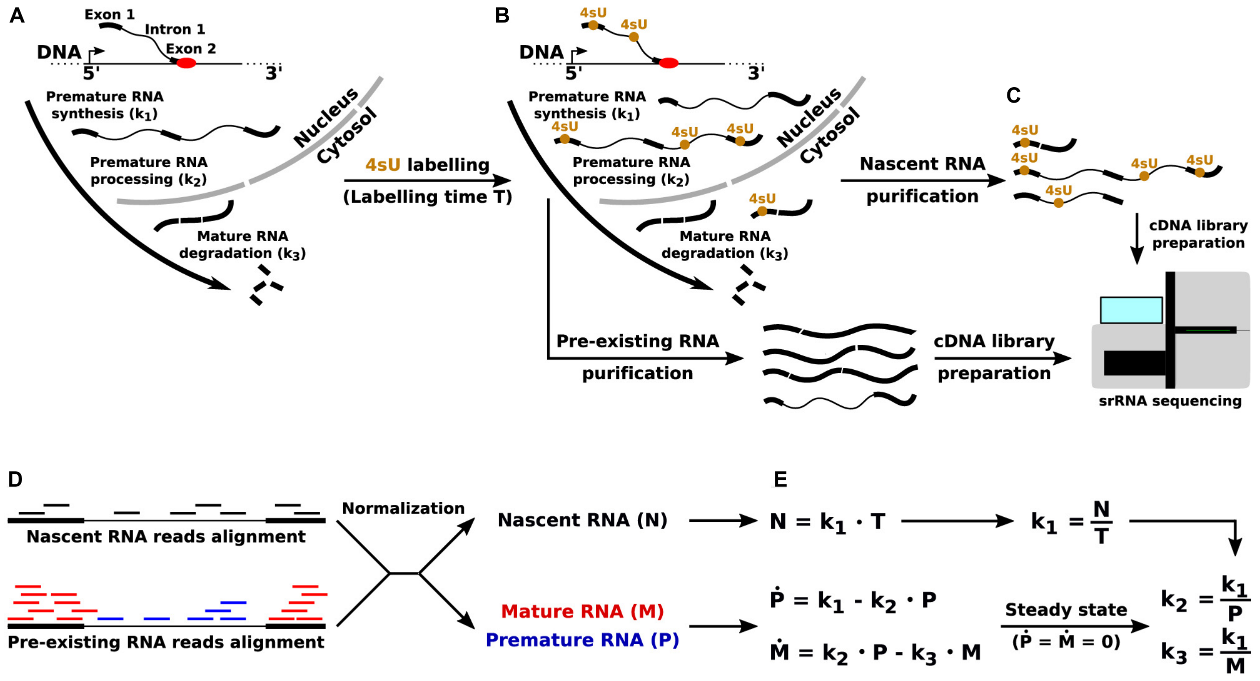

A
Exon 1
Intron 1
Exon 2
DNA
5'
3'
Premature RNA synthesis (k_1)
Nucleus
Cytosol
Premature RNA processing (k_2)
Mature RNA degradation (k_3)
4sU labelling
(Labelling time T)
B
4sU
DNA
5'
3'
Premature RNA synthesis (k_1)
Premature RNA processing (k_2)
Mature RNA degradation (k_3)
Nucleus
Cytosol
C
Nascent RNA purification
cDNA library preparation
Pre-existing RNA purification
cDNA library preparation
srRNA sequencing
D
Nascent RNA reads alignment
Pre-existing RNA reads alignment
Normalization
Nascent RNA (N)
Mature RNA (M)
Premature RNA (P)
E
$N = k_1 \cdot T$
$k_1 = \frac{N}{T}$
$\dot{P} = k_1 - k_2 \cdot P$
$\dot{M} = k_2 \cdot P - k_3 \cdot M$
Steady state
$(\dot{P} = \dot{M} = 0)$
$k_2 = \frac{k_1}{P}$
$k_3 = \frac{k_1}{M}$

FIGURE 1 | Quantification of the RNA kinetic rates through RNA metabolic labeling coupled with srRNA-seq. **(A)** The key steps of the RNA life cycle, and the corresponding RNA kinetic rates: synthesis (k_1) of premature RNA, processing (k_2) of premature into mature RNA, and degradation (k_3) of mature transcripts. **(B)** Incorporation of the uridine analog 4sU into newly synthetized transcripts. **(C)** Pre-existing and nascent RNA purification and sequencing through srRNA-seq. **(D)** Quantification of premature (P), mature (M), and nascent (N) RNA from srRNA-seq reads. **(E)** RNA life cycle mathematical modeling and quantification of the RNA kinetic rates in the steady-state limit.

as biotin on the modified residues [e.g., 4sU-based RNA metabolic labeling (Dolken et al., 2008)]. These techniques do not provide neither the exact modification site (they are limited to 100–200 bp resolution), nor a precise quantification of the proportion of modified transcripts (Molinie et al., 2016), despite the development of *ad hoc* experimental (Sun et al., 2012) and computational (de Pretis et al., 2015) normalization techniques. Indeed, an alternative approach, m^{6}A-LAIC-seq (Molinie et al., 2016) has been developed that relies on spike-ins to provide a precise quantification of the m^{6}A abundance, at the cost of skipping the RNA fragmentation step and losing positional information on the mark. A second class of methodologies is based on the identification of RNA modifications signatures in the retro-transcribed cDNA. One approach belonging to this class exploits the early interruption of retrotranscription at the modification site to produce specific truncation signatures [e.g., ICE-seq for inosine detection (Sakurai et al., 2010)]. Alternative approaches were developed to retro-transcribe the modified bases and their native counterparts to different nucleotides, thus inferring the site of the modification based on specific mismatches in the reads alignment (Baptista and Dölken, 2018). For example, SLAM-seq allows the *in silico* identification of reads derived from nascent RNAs by inducing the pairing of alkylated 4sU to guanines (Herzog et al., 2017). These methods markedly increase the resolution, but are typically semi-quantitative, suffering from low sensitivity (Neumann et al., 2019). Hybrid techniques were also developed. For example, methylation induced cross-linking and immunoprecipitation (miCLIP) combines m^{6}A-immunoprecipitation with the antibody cross-linking, leading to conversion and truncation events. Their identification in the sequencing results allows the mapping of m^{6}A at base-resolution (Linder et al., 2015). However, this method is affected by low crosslink efficiency, reducing the sensitivity. Recently, two novel approaches were developed that do not rely on immunoprecipitation. MAZTER-seq (Garcia-Campos et al., 2019) allows the quantitative and base-resolution identification of m^{6}A marks, relying on the use of a restriction enzyme that cuts only when the target site is not methylated. As a downside, the mapping is limited to the identification of m^{6}A marks in specific context sites (16% of all expected m^{6}A sites in mammals). DART-seq (Meyer, 2019) recruits APOBEC1 proteins at m^{6}A sites through readers of the YTH family, allowing the identification of the marks by the detection of adjacent C to U mutations. It was used in combination with srRNA-seq, with as little as 10 ng of total RNA, and with long-reads RNA sequencing (lrRNA-seq), leading to single transcript m^{6}A detection. The key downside of this method is the required cells transfection with APOBEC1-YTH fusion protein. Finally, the ability to quantify the abundance of m^{6}A marks remains to be established.

A number of computational tools were developed that are useful for calling RNA modifications on srRNA-seq data, especially tailored toward the analysis of m^{6}A marks in MeRIP-seq datasets. exomePeak, while not originally developed for this task, is one of the most frequently adopted tools for the identification of m^{6}A peaks (Meng et al., 2013). Indeed, a detailed protocol was described for its application on MeRIP-seq datasets (Meng et al., 2014). This tool adopts a sliding window approach with a conditional test relying on Poisson distributions. HEPeak is an HMM-based tool dedicated to the identification of m^{6}A marks, claiming improved sensitivity and specificity compared to exomePeak (Cui et al., 2015). From the same authors, MeTPeak was later proposed that is able to take advantage of the variance across replicates, and models the reads dependency across a region (Cui et al., 2016). A number of tools were developed that are dedicated to differential RNA methylation analysis, including MeTDiff (Cui et al., 2018), FunDMDeep (Zhang S. Y. et al., 2019), and RADAR (Zhang Z. et al., 2019). Finally, m^{6}A viewer is a Java stand alone application that supports detection, analysis, and visualization of m^{6}A marks, the former relying on the previously described tools (Antanaviciute et al., 2017).

Besides the specific limitations of each technique, all available protocols for the profiling of RNA modifications through srRNA-seq share some key limitations. *First*, they require specific reagents for each modification of interest, which currently limits the profiling to a handful of modifications (Helm and Motorin, 2017). *Second*, the library preparations, and the sequencing procedure, remove the RNA marks. As a consequence, most available approaches for the modifications profiling are indirect, reducing specificity and sensitivity (Helm and Motorin, 2017). *Third*, the reduced length of srRNA-seq reads (50–300 bp) is a major obstacle for the analysis of individual RNA molecules, despite the development of methods to infer isoforms expression from these data (Zhang et al., 2017b). As a consequence, the assignment of individual or co-occurring modifications to a given RNA molecule, or even to a given isoform, is not feasible. *Fourth*, srRNA-seq protocols are not readily applicable to detect two (or more) RNA modifications simultaneously.

Although recent interesting technical advances are starting to appear [e.g., simultaneous detection of N^1-methyladenosine, 5-methylcytosine, and pseudouridine (Khoddami et al., 2019)], these methods highly depend on the specific combination of marks. The reasons for this limitation are manifolds. Likely, the methods for the profiling of different modifications should be consecutively applied, and the output of one method could be poorly suitable for the subsequent. For the same reason, a high amount of starting material is likely to be necessary, to avoid capturing only highly expressed transcripts. Alternatively, numerous rounds of PCR would be necessary, introducing amplification biases (Aird et al., 2011; Kebschull and Zador, 2015). The limitations in specificity and sensitivity of each method would combine. Moreover, it would be crucial and cumbersome to develop normalization procedures for the comparison of the results from each approach, possibly based on spike-ins. Finally, it would be hard to keep track of the positional information of each modification.

Things would get even more complicated when, in addition to the mark of interest, the dynamics of RNA metabolism are also of interest, which require the identification of an exogenous modification as second mark. In this case, to quantify the RNA kinetic rates of modified and unmodified RNAs, it would be necessary to quantify all four possible combinations: nascent/modified, nascent/unmodified, pre-existing/modified, and pre-existing/unmodified transcripts (**Figure 2**). Currently, the best approach to jointly identify 4sU and m^{6}A would be to start by separating nascent and pre-existing RNA using 4sU metabolic labeling and purification (Dolken et al., 2008). Then, for each of these, the m^{6}A-LAIC-seq protocol could be applied to separate m^{6}A methylated RNAs from unmethylated transcripts. At the end, four samples per condition should be prepared and sequenced. This approach is evidently very complex and onerous, it would require a lot of starting material and complicated downstream analyses, including spike-ins based normalization of the datasets. For all these reasons, the most common compromise is to profile m^{6}A, and to perform metabolic labeling through independent experiments (Li et al., 2017; Furlan et al., 2019b). However, this type of approach completely compromises the possibility of a direct quantification of the dynamics of modified and unmodified transcripts, since it only allows to quantify the dynamics of the pool of transcripts for each gene, and then combine this information with the expected degree of modification for that population. Altogether, approaches based on srRNA-seq are increasingly inadequate and could hamper the progress in the field of epitranscriptomics.

LONG-READS DIRECT RNA SEQUENCING FOR THE IDENTIFICATION OF MODIFICATIONS IN NATIVE RNAs

In the last few years remarkable efforts were dedicated to overcoming the limitations of srRNA-seq based approaches (Stark et al., 2019) for the identification of RNA modifications within individual RNA molecules and isoforms. As a result, few novel sequencing approaches that emerged recently allow rRNA-seq. One platform, PacBio (developed by Pacific Biosciences), exploits a sequencing by synthesis approach mediated by an immobilized polymerase (Eid et al., 2009). Another one, which will be the main focus in the next sections of this perspective, was developed by Oxford Nanopore Technologies (ONT), and consists of an array of thousands of nanopores which allow a flow of ions across a dielectric membrane, thus generating a measurable current. The active translocation of a molecule of nucleic acids (either DNA, cDNA, or RNA) through each pore, mediated by an engineered motor protein, results in a sequence-specific perturbation of the measured current. In turn, this signal can be exploited to infer the corresponding sequence of nucleotides (Kasianowicz et al., 1996; Smith et al., 2015). lrRNA-seq approaches were successfully used to study transcriptional and post-transcriptional regulation in various physiological and disease conditions (De Roeck et al., 2017; Aneichyk et al., 2018; Anvar et al., 2018; Nattestad et al., 2018), including single-cells (Byrne et al., 2017). Focusing on RNAs, these techniques can produce single reads of up to 10^4 bases, with an average length of almost 1 Kb for ONT (Workman et al., 2018). Hence, in a number

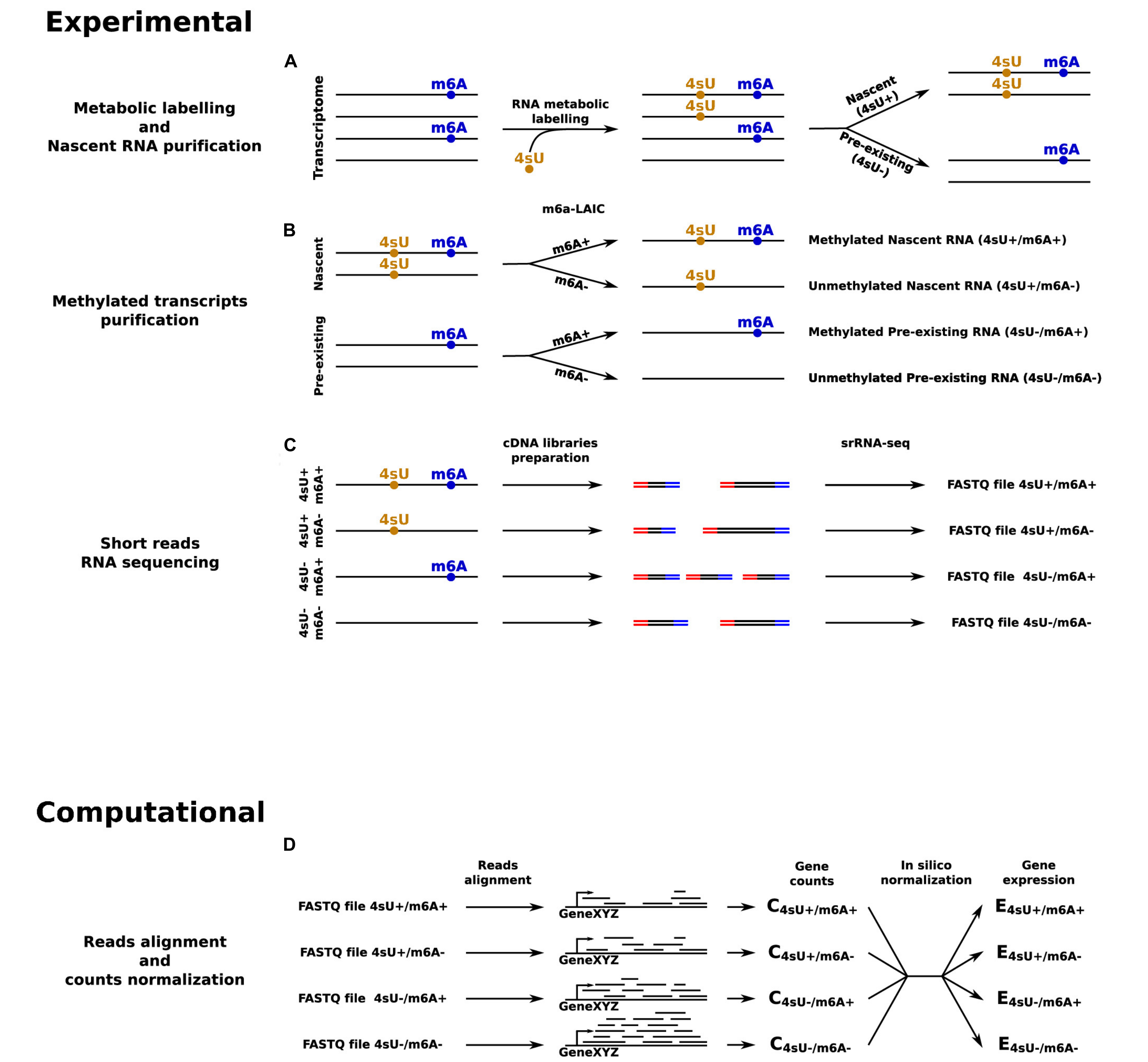


FIGURE 2 | srRNA-seq based approach to quantify transcripts' expression levels in all the four possible combinations given by the presence or absence of 4sU and m^{6}A RNA modifications. **(A)** RNA metabolic labeling, based on the incorporation of 4sU, is applied to separate the nascent portion of the transcriptome from the pre-existing counterpart. **(B)** m^{6}A-LAIC-seq is applied for both nascent and pre-existing RNAs to separate methylated from unmethylated transcripts. **(C)** cDNA library preparation and sequencing for: pre-existing unmethylated RNAs, pre-existing methylated RNAs, nascent unmethylated RNAs, and nascent methylated RNAs. **(D)** *In silico* reads alignment, counts quantification, and normalization to estimate transcripts' expression levels across all the four conditions.

of cases, this allows the profiling of full-length RNA molecules, and the fine characterization of their alternative isoforms. This is especially true for mature transcripts, whose median length for human and mouse mRNAs is around 2 Kb [based on the hg19 and mm10 UCSC genome releases (Haeussler et al., 2019)]. Instead, the likelihood of sequencing full-length premature transcripts is lower. Indeed, their median open reading frame length is in the 13–18 Kb range, although co-transcriptional splicing could significantly reduce this figure (it is likely that some intron was already excised before the completion of RNA synthesis).

The direct RNA sequencing approach developed by ONT does not go through the conversion of RNA into cDNA, and does not rely on amplification steps. For these reasons, the RNA modifications are preserved and can induce specific alterations in the current registered by the sequencer (Garalde et al., 2018). Altogether, this approach represents a potential solution to most of the limitations of srRNA-seq discussed above, due to its ability to directly identify any, and possibly multiple, RNA modification in single, full-length molecules. dRNA-seq was recently applied to study the transcriptome of viruses (Moldován et al., 2018; Tombácz et al., 2018; Boldogkõi et al., 2019; Depledge et al., 2019), yeast (Garalde et al., 2018), animals (Jiang et al., 2019; Roach et al., 2019; Smith et al., 2019), and plants (Zhao et al., 2019).

However, a number of limitations characterize the young field of dRNA-seq. *First*, current dRNA-seq protocols are available only for the sequencing of targeted, non-polyadenylated RNAs (Keller et al., 2018; Smith et al., 2019) or polyadenylated RNAs. This is due to the library preparation protocolos, which typically targets polyA tails or specific 3′ sequences for ligating sequencing adapters anchoring the motor protein. This limitation could be addressed using adapters with random 3′ sequences, with the risk of introducing a bias for recurrent RNA motifs, or through *in vitro* polyadenylation of transcripts devoid of a polyA-tail (Wongsurawat et al., 2018). *Second*, while the throughput of dRNA-seq is rapidly growing, it currently compares to the low- or mid-end coverage of srRNA-seq experiments. This could limit the number of detectable transcripts, although, importantly, the abundance of those that can be detected is well correlated with high-coverage srRNA-seq data (Garalde et al., 2018). This issue could be solved in the future by improving the speed of translocation of RNAs across the nanopore, and/or extending the sequencing time by prolonging the pores' lifetime. Noteworthy, given the same throughput in terms of sequenced bases, lrRNA-seq vs srRNA-seq data have a substantial difference: while the former allows detecting entire transcripts, the latter offers a more unbiased sampling of any RNA fragment, thus also covering a larger portion of the transcriptome (Soneson et al., 2019). This could in part be obviated by a coarse RNA fragmentation before the library preparation, and would also reduce the 3′ coverage bias of dRNA-seq data, whose reads start from a transcript's 3′ end. A drawback of this approach is that it would compromise the one-to-one correspondence between reads and RNA molecules. *Third*, the accuracy of base calling on dRNA-seq data is currently significantly lower than srRNA-seq. When base calling errors occur at sites of RNA modification, they are likely due to the inability of the base caller's to deal with changes in the signal originated by those marks. However, these errors represent a small fraction of incorrect base calls, due to the low number of marks per transcripts (e.g., 2–3 m^6A marks per RNA). Hence, reduced base calling accuracy is not considered a major issue in the field of RNA modifications but, on the contrary, represents an opportunity for aiding the identification of modified bases (Liu et al., 2019). *Fourth*, there could be limitations on the detectability of specific RNA modifications. For example, in the context of RNA metabolic labeling, the ability of dRNA-seq to identify various (exogenous) modified nucleotides was tested (Maier et al., 2019). This revealed that 4sU modified nucleotides, commonly used in metabolic labeling through srRNA-seq, were not compatible with the nanopores, leading to blockages during the sequencing, although this issue was not confirmed in a more recent report (Drexler et al., 2019). Instead, other marks, such as 5-ethynyluridine (5eU), were found to be suitable for these experiments.

In conclusion, this is a young and rapidly evolving research field, based on a highly collaborative research community. Hence, numerous labs are actively involved to find solutions or improvements to all these limitations, which are likely to be fully or partially overcome in the next few years (Rang et al., 2018).

COMPUTATIONAL TOOLS FOR THE DETECTION OF MODIFICATIONS IN LONG-READS DIRECT RNA SEQUENCING

Recent and growing literature is available about the footprints left by RNA modifications on dRNA-seq data, and how to exploit them to detect RNA marks (Xu and Seki, 2019). Differences in current levels between native bases and their modified counterparts were reported for m^6A, m^5C, m^7G, and pseudouridine (Garalde et al., 2018; Workman et al., 2018; Smith et al., 2019). Moreover, the increase of base miscalls frequency in concomitance to modified sites were observed next to "A-to-I," 7-methylguanosine and pseudouridine sites (Workman et al., 2018; Smith et al., 2019). These observations led to the development of specific computational tools for the detection of RNA modifications.

Tombo, an official tool provided by ONT, requires a model of the signal generated by the modification in all possible sequence contexts, to be used as a baseline for the identification of the same mark at single molecule resolution within a new dRNA-seq dataset (Stoiber et al., 2016). Notably, baseline data for 5-methylcytosine marks are included in the tool (Viehweger et al., 2019). Alternatively, data for a condition devoid of modifications can be provided. With a similar approach, Tombo was recently used to identify m^6A in yeast with an accuracy of 69% and a recovery of 59%, compared with m^6A peaks identified with MeRIP-seq (Liu et al., 2019). Obviating for the need of these positive or negative baseline data, Tombo can be used to compare the signal observed for each k-mer with that of any possible unmodified k-mer, although this approach is affected by high false positive rates.

EpiNano relies on a support vector machine, and exploits the increased frequency of alignment errors and the low base quality caused by the presence of the modification of interest (Liu et al., 2019). The tool is first trained and tested on two sets of *in vitro* transcribed synthetic RNAs that contain either m^6A only or unmodified adenosine only. Its classification performance in the context of the expected m^6A RRACH motif was excellent (area under the curve up to 0.944). Rather, the performance decreased when the tool was applied on *in vivo* yeast data and benchmarked with MeRIP-seq m^6A calls for the same conditions (accuracy: 87% and recovery: 32%). In terms of downsides, EpiNano requires prior knowledge on the sequence motif for the mark of interest, and it cannot achieve single molecule resolution, since it aggregates the information derived from multiple reads alignments.

ELIGOS aims at the unbiased identification of any RNA modification that would impact bases errors frequencies. It relies on the comparison between dRNA-seq of native and cDNA-converted transcripts, the latter used as a reference that is devoid of any mark due to the retro-transcription to cDNA (Wongsurawat et al., 2018). ELIGOS was tested on *in vitro* fully modified transcripts, rRNAs from various species, and a human lymphoblastoid cell line. Like Tombo, the main downside of ELIGOS is in terms of false positive rates.

A further method for m^6A identification that was recently released is called MINES (Lorenz et al., 2019). This software implements a random forest classifier trained on a set of high confidence, experimentally defined, m^6A sites within canonical DRACH motifs. This method showed high accuracy and precision, and also has single-isoform, single-base resolution. However, MINES can only predict m^6A sites within DRACH motifs, which only comprise a portion of all m^6A sites. A further potential limitation is due to the fact that the classifier was trained on m^6A sites defined with CLIP and – as such – might suffer of biases similar to those caused by antibody-based methods.

Nano-ID was recently developed for detecting the incorporation of the exogenous mark 5eU into nascent RNA (Maier et al., 2019), implementing the analysis of RNA metabolic labeling on the ONT platform. This tool relies on a neural network trained to distinguish dRNA-seq signal of fully unlabeled from fully labeled RNAs (24 h 5eU labeling time), to classify reads from nascent transcripts, while no positional information on 5eU marks is returned. The results achieved by nano-ID on this test set were very encouraging (area under the curve 0.95), and the tool was applied to infer the isoform-level rates of synthesis and degradation in K562 cells, and how they were affected by heat shock.

Nanocompore is a novel tool recently released, which is based on the comparison of a condition of interest with a condition where the writer for a specific mark was depleted or removed (Leger et al., 2019). The idea is that the removal of the mark leads to a change in the ONT signal, which could be identified through statistical tests by comparing the two conditions. As a result, Nanocompore can provide near base-resolution and single molecule calls for the mark of interest. Alternatively, analogously to ELIGOS, if the baseline condition is depleted of multiple or possibly all marks (e.g., via *in vitro* transcription), the tool returns the corresponding changes in the signal to identify all marks occurrence, while mark-specific calls are not possible. Advantages and disadvantages of the tools discussed above are reported in **Table 1**.

APPLYING DIRECT RNA SEQUENCING TO QUANTIFY THE DYNAMICS OF MODIFIED RNAs

The recent surge in the number of tools for the identification of specific modifications indicates that the field is quickly progressing. However, a number of improvements are required for the joint analysis of the patterning of an endogenous modification, such as m^6A, with the quantification of the corresponding RNA dynamics, via metabolic labeling and profiling of exogenous modifications such as 4sU or 5eU (**Figure 3**).

First, the modifications have to be profiled at single molecule resolution, a prerequisite for the direct matching of the RNA dynamics with the modification status. This would allow understanding how the RNA kinetic rates are impacted by the presence of a modification, and, potentially, by its patterning (numerosity and position). Notably, the frequency and the specific position of occurrence of the marks is increasingly recognized as an important factor. For example, the fate of RNAs carrying multiple m^6A marks was shown to be influenced by a liquid–liquid phase separation processes driven by the binding of readers of the YTH family. Eventually, those transcripts were shown to be targeted to specific cellular compartments, including stress-granules and P-bodies, with important consequences for their translation and stability (Ries et al., 2019).

Second, tools based on supervised machine learning could be preferable in the field, compared to methods for the unsupervised identification of the marks. In fact, various confounding factors could potentially affect direct RNA sequencing data, which could be easier to address in a supervised framework. However, supervised methods require training on sets of modified transcripts, which should be built so that they closely reflect the characteristics of *in vivo* datasets. For example, for endogenous modifications, rather than producing *in vitro* fully modified transcripts, the level of modification could be tuned by mixing unmodified and modified nucleotides to match the expected frequency of the mark. For exogenous marks, the approach described in Maier et al. (2019) could be followed, where physiological high-level of incorporation of a modified nucleotides are obtained by its prolonged availability in the cells medium.

Third, the current ONT signal (amplitude and dwell time) is the most direct data type for the identification of the marks, compared to more indirect measurements, such as the error rate. While tools, such as EpiNano, showed a good performance by only using the latter, we would recommend trying to incorporate information from the former. Indeed, indirect measurements could be completely or partially originated by unexpected causes, which could lead to high false positive rates with *in vivo* datasets.

Fourth, the quantification of RNA dynamics should include the step of premature RNA processing. This is often neglected, by assuming the corresponding rate being constant. However, RNA synthesis and processing are tightly coupled, then when the former is modulated, which often occurs, the latter is also expected to be altered (Neugebauer, 2019). Moreover, recent reports start unveiling the frequency and importance of changes in splicing dynamics (Rabani et al., 2014; de Pretis et al., 2015, 2017; Louloupi et al., 2018; Furlan et al., 2019a; Wachutka et al., 2019). The cost of considering the processing step is two fold: it markedly increases the complexity of the underlying mathematical models, and implies the quantification of the abundance of premature RNA species. The latter is specifically problematic for the ONT platform. Indeed, the library preparation procedure expects transcripts with the polyA tail, which are lacking in premature RNAs. *In vitro* polyadenylation

TABLE 1 | Comparing strengths and pitfalls of four software packages for m^{6}A detection from Nanopore dRNA-seq data.

	EpiNano	ELIGOS	MINES	Nanocompore
Requires training dataset	Yes	No	Yes	No
Requires comparison condition	No	Yes (cDNA)	No	Yes
Limited to RACH motifs	Yes	No	Yes	No
Single nucleotide resolution	Yes	Yes	Yes	No
Isoform resolution	Yes	Yes	Yes	Yes
Single molecule resolution	No	No	No	Yes
Able to distinguish different modifications	Yes	No	Yes	Yes

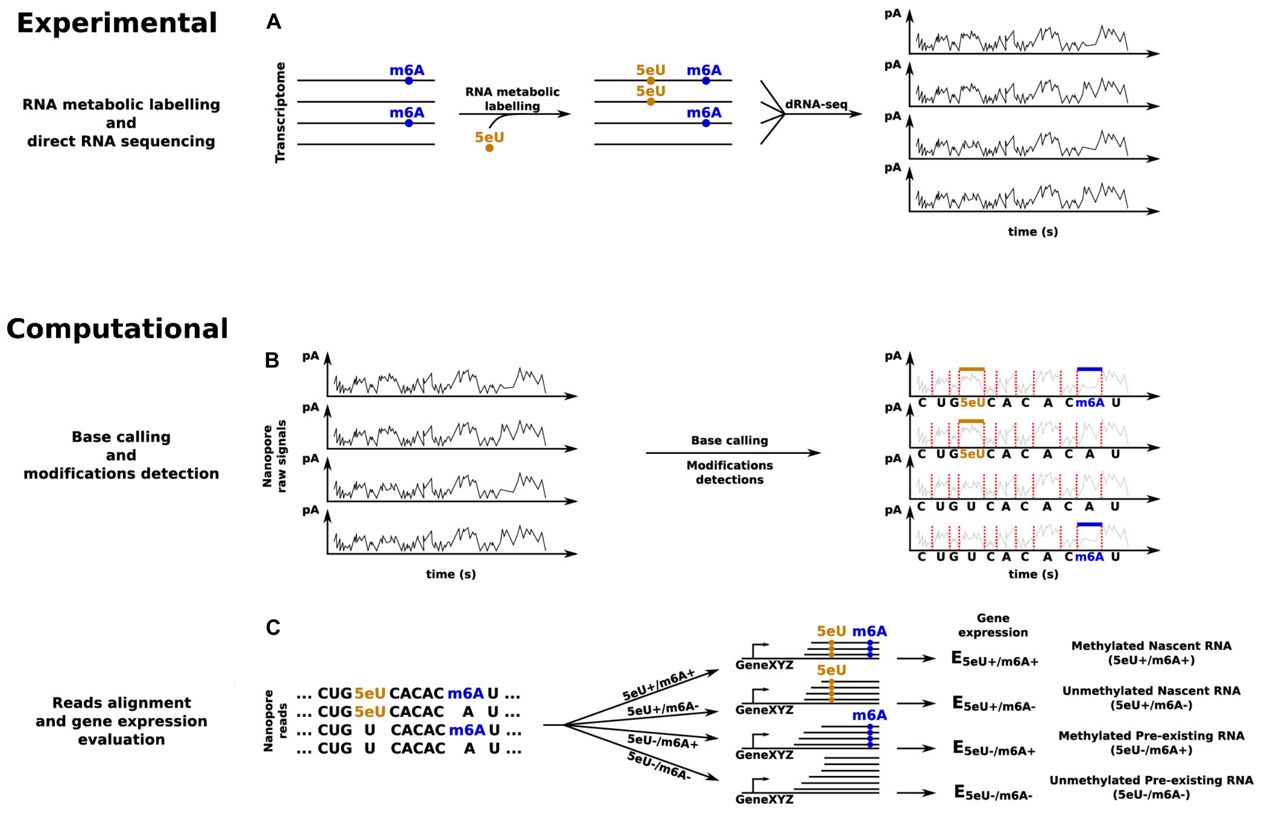


FIGURE 3 | dRNA-seq based approach to quantify transcripts' expression levels in all the four possible combinations given by the presence or absence of 5eU and m^{6}A RNA modifications. **(A)** RNA metabolic labeling, based on the incorporation of 5eU, is applied to mark nascent transcripts, before direct RNA sequencing. **(B)** Base calling and identification of the two RNA modifications. **(C)** Reads alignment and *in silico* separation, according to the presence or absence of each RNA modification, to estimate transcripts' expression levels across all the four conditions.

with m^{6}A could be used for adding m^{6}A-tails to premature transcripts. This would allow the sequencing of premature RNAs, and would preserve the sequencing information about the endogenous tails of mature transcripts, for studies on their functional impact on RNA dynamics.

Fifth, reads from premature RNAs would have to be distinguished from those from mature species. The presence of an endogenous polyA tail would provide a way to computationally identifying reads from mature species. However, this approach would fail for those mRNAs that are not polyadenylated in their endogenous mature form. An alternative criterion is to consider the reads containing introns as premature RNA. This could be problematic in case of intron retention, which in many organisms, including humans, is not infrequent (Chaudhary et al., 2019; Monteuuis et al., 2019). The request of more than one intron in order to classify a read as premature RNA would probably eliminate this issue. Of course, such a strict condition would cause the exclusion of those genes that have less than two introns, which often occurs in some organisms (e.g., yeast or plants). The best criterion could eventually be a mix of the proposed approaches, selected according to the biological system under analysis and the transcripts of interest. For instance, to study mRNA kinetics in mammalian cells, mature RNA could be estimated considering fully spliced, polyadenylated transcripts, while premature RNA could be quantified from the remaining reads, possibly requiring the presence of one or more introns.

Once proficient algorithms for the detection of the endogenous (e.g., m^6A) and exogenous (e.g., 5eU) marks at single molecule resolution are in place, they could be used, in series, for the identification of the four possible classes defined by the presence or absence of each modification. The performance of such an approach should be tested on a dataset generated *ad hoc*. The genesis of reads with both the RNA modifications, or missing only the exogenous mark, is feasible by using or avoiding long-time metabolic labeling, respectively. Instead, reads devoid of both the base analogs can be produced sequencing the corresponding cDNA. It is more difficult to generate transcripts that lack only the endogenous modification, which could be obtained by knocking-out the corresponding writer (for those marks for which this is known). However, genetic compensation (El-Brolosy and Stainier, 2017) or writer's redundancy could lead to the incomplete depletion of the RNA modification.

ADDITIONAL REMARKS

The study of the impact of RNA modifications on the RNA life cycle dynamics would largely benefit from the development of a unified computational framework. This, starting from long reads dRNA-seq data, should manage the RNA kinetic rates inference, according to their modification status, at the level of individual transcriptional units or specific isoforms.

A convenient starting point could be INSPEcT (de Pretis et al., 2015), a tool developed in our lab for the inference of all RNA kinetic rates (synthesis, processing, and degradation) from srRNA-seq data. The user should only pay attention to quantify premature and mature RNA in both nascent and pre-existing fractions according to the guidelines presented above. Additionally, if the quantification of dynamics at the level of specific isoforms is desired, the analysis should be conducted considering the reads associated with each isoform, rather than those associated with the whole transcriptional unit. Finally, if this analysis is applied independently on the set of modified and unmodified reads, it would allow comparing the kinetic rates among them, as illustrated in **Figure 3B**.

INSPEcT has been recently extended by implementing a novel approach that allows the inference of synthesis, processing and degradation kinetic rates without nascent RNA profiling (Furlan et al., 2019a). This approach could be an interesting alternative to study the relation between RNA modifications and RNA life cycle dynamics without requiring metabolic labeling and the consequent identification of the exogenous modification. This would also allow studying the impact on RNA dynamics of those modifications that mark the same base targeted by metabolic labeling, such as pseudouridine and 5eU.

In conclusion, a number of recent and on-going technology advancements are significantly facilitating the study of the functional consequences of RNA modifications on the fate of marked transcripts. In particular, the combined application of RNA metabolic labeling, for the profiling of nascent transcripts and the quantification of the kinetic rates governing the RNA life cycle dynamics, and of long-reads direct RNA sequencing, is particularly promising. Indeed, they promise to deliver data of unprecedented quality and resolution, and should allow studying the impact of RNA modifications at the level of individual molecules and isoforms.

AUTHOR CONTRIBUTIONS

MF and MP conceived the study. MF led the writing and produced the figures. MP supervised the study and the writing of the manuscript. All authors contributed discussing and writing the manuscript.

FUNDING

This article is based upon work from COST Action EPITRAN (CA16120), supported by COST (European Cooperation in Science and Technology).

REFERENCES

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12:R18. doi: 10.1186/gb-2011-12-2-r18

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., et al. (2018). Dissecting the causal mechanism of X-linked dystonia-parkinsonism by integrating genome and transcriptome assembly. *Cell* 172, 897–909.e21. doi: 10.1016/j.cell.2018.02.011

Antanaviciute, A., Baquero-Perez, B., Watson, C. M., Harrison, S. M., Lascelles, C., Crinnion, L., et al. (2017). m^6aViewer: software for the detection, analysis, and visualization of N^6 -methyladenosine peaks from m^6A-seq/ME-RIP sequencing data. *RNA* 23, 1493–1501. doi: 10.1261/rna.058206.116

Anvar, S. Y., Allard, G., Tseng, E., Sheynkman, G. M., de Klerk, E., Vermaat, M., et al. (2018). Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* 19:46. doi: 10.1186/s13059-018-1418-0

Baptista, M. A. P., and Dölken, L. (2018). RNA dynamics revealed by metabolic RNA labeling and biochemical nucleoside conversions. *Nat. Methods* 15, 171–172. doi: 10.1038/nmeth.4608

Bartosovic, M., Molares, H. C., Gregorova, P., Hrossova, D., Kudla, G., and Vanacova, S. (2017). N^6-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3′-end processing. *Nucleic Acids Res.* 45, 11356–11370. doi: 10.1093/nar/gkx778

Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

Boldogkõi, Z., Moldován, N., Balázs, Z., Snyder, M., and Tombácz, D. (2019). Long-read sequencing – a powerful tool in viral transcriptome research. *Trends Microbiol.* 27, 578–592. doi: 10.1016/j.tim.2019.01.010

Bushkin, G. G., Pincus, D., Morgan, J. T., Richardson, K., Lewis, C., Chan, S. H., et al. (2019). m^6A modification of a 3′ UTR site reduces RME1 mRNA levels to promote meiosis. *Nat. Commun.* 10:3414. doi: 10.1038/s41467-019-11232-7

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8:16027. doi: 10.1038/ncomms16027

Chaudhary, S., Khokhar, W., Jabre, I., Reddy, A. S. N., Byrne, L. J., Wilson, C. M., et al. (2019). Alternative splicing and protein diversity: plants versus animals. *Front. Plant Sci.* 10:708. doi: 10.3389/fpls.2019.00708

Chen, T., Hao, Y.-J., Zhang, Y., Li, M.-M., Wang, M., Han, W., et al. (2015). m^{6}A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* 16, 289–301. doi: 10.1016/j.stem.2015.01.016

Cui, X., Meng, J., Rao, M. K., Chen, Y., and Huang, Y. (2015). HEPeak: an HMM-based exome peak-finding package for RNA epigenome sequencing data. *BMC Genomics* 16:S2. doi: 10.1186/1471-2164-16-S4-S2

Cui, X., Meng, J., Zhang, S., Chen, Y., and Huang, Y. (2016). A novel algorithm for calling mRNA m 6 A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics* 32, i378–i385. doi: 10.1093/bioinformatics/btw281

Cui, X., Zhang, L., Meng, J., Rao, M. K., Chen, Y., and Huang, Y. (2018). MeTDiff: a novel differential RNA methylation analysis for MeRIP-Seq data. *IEEE/ACM Trans. Comput. Biol. and Bioinform.* 15, 526–534. doi: 10.1109/TCBB.2015.2403355

Darnell, R. B., Ke, S., and Darnell, J. E. (2018). Pre-mRNA processing includes *N* 6 methylation of adenosine residues that are retained in mRNA exons and the fallacy of "RNA epigenetics.". *RNA* 24, 262–267. doi: 10.1261/rna.065219.117

de Pretis, S., Kress, T., Morelli, M. J., Melloni, G. E. M., Riva, L., Amati, B., et al. (2015). INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* 31, 2829–2835. doi: 10.1093/bioinformatics/btv288

de Pretis, S., Kress, T. R., Morelli, M. J., Sabò, A., Locarno, C., Verrecchia, A., et al. (2017). Integrative analysis of RNA polymerase II and transcriptional dynamics upon MYC activation. *Genome Res.* 27, 1658–1664. doi: 10.1101/gr.226035.117

De Roeck, A., Van den Bossche, T., van der Zee, J., Verheijen, J., De Coster, W., Van Dongen, J., et al. (2017). Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. *Acta Neuropathol.* 134, 475–487. doi: 10.1007/s00401-017-1714-x

Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., et al. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat. Commun.* 10:754. doi: 10.1038/s41467-019-08734-9

Dolken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* 14, 1959–1972. doi: 10.1261/rna.1136108

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Drexler, H. L., Choquet, K., and Churchman, L. S. (2019). Splicing kinetics and coordination revealed by direct nascent RNA sequencing through Nanopores. *Mol. Cell* 77, 985–998.e8. doi: 10.1016/j.molcel.2019.11.017

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

El-Brolosy, M. A., and Stainier, D. Y. R. (2017). Genetic compensation: a phenomenon in search of mechanisms. *PLoS Genet.* 13:e1006780. doi: 10.1371/journal.pgen.1006780

Engel, M., Eggert, C., Kaplick, P. M., Röh, S., Tietze, L., Namendorf, C., et al. (2018). The role of m^{6}A/m-RNA methylation in stress response regulation. *Neuron* 99, 389–403.e9. doi: 10.1016/j.neuron.2018.07.009

Furlan, M., Galeota, E., Del Gaudio, N., Dassi, E., Caselle, M., de Pretis, S., et al. (2019a). Genome-wide dynamics of RNA synthesis, processing and degradation without RNA metabolic labeling. *bioRxiv* [Preprint]. doi: 10.1101/520155

Furlan, M., Galeota, E., De Pretis, S., Caselle, M., and Pelizzola, M. (2019b). m^{6}A-dependent RNA dynamics in T cell differentiation. *Genes* 10:28. doi: 10.3390/genes10010028

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi: 10.1038/nmeth.4577

Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., et al. (2019). Deciphering the "m^{6}A Code" via antibody-independent quantitative profiling. *Cell* 178, 731–747.e16. doi: 10.1016/j.cell.2019.06.013

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., et al. (2015). m 6 A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* 347, 1002–1006. doi: 10.1126/science.1261417

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., et al. (2019). The UCSC genome browser database: 2019 update. *Nucleic Acids Res.* 47, D853–D858. doi: 10.1093/nar/gky1095

Haussmann, I. U., Bodi, Z., Sanchez-Moran, E., Mongan, N. P., Archer, N., Fray, R. G., et al. (2016). m^{6}A potentiates Sxl alternative pre-mRNA splicing for robust *Drosophila* sex determination. *Nature* 540, 301–304. doi: 10.1038/nature20577

Helm, M., and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat. Rev. Genet.* 18, 275–291. doi: 10.1038/nrg.2016.169

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., et al. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* 14, 1198–1204. doi: 10.1038/nmeth.4435

Ianniello, Z., Paiardini, A., and Fatica, A. (2019). N^6-methyladenosine (m^{6}A): a promising new molecular target in acute myeloid leukemia. *Front. Oncol.* 9:251. doi: 10.3389/fonc.2019.00251

Jackman, J. E., and Alfonzo, J. D. (2013). Transfer RNA modifications: nature's combinatorial chemistry playground: transfer RNA modifications. *Wiley Interdiscip. Rev. RNA* 4, 35–48. doi: 10.1002/wrna.1144

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., et al. (2011). N^6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* 7, 885–887. doi: 10.1038/nchembio.687

Jiang, F., Zhang, J., Liu, Q., Liu, X., Wang, H., He, J., et al. (2019). Long-read direct RNA sequencing by 5′-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. *RNA Biol.* 16, 950–959. doi: 10.1080/15476286.2019.1602437

Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13770–13773. doi: 10.1073/pnas.93.24.13770

Kasowitz, S. D., Ma, J., Anderson, S. J., Leu, N. A., Xu, Y., Gregory, B. D., et al. (2018). Nuclear m^{6}A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. *PLoS Genet.* 14:e1007412. doi: 10.1371/journal.pgen.1007412

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., et al. (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. doi: 10.1101/gad.301036.117

Kebschull, J. M., and Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 43:e143. doi: 10.1093/nar/gkv717

Keller, M. W., Rambo-Martin, B. L., Wilson, M. M., Ridenour, C. A., Shepard, S. S., Stark, T. J., et al. (2018). Direct RNA sequencing of the coding complete influenza A virus genome. *Sci. Rep.* 8:14408. doi: 10.1038/s41598-018-32615-8

Khoddami, V., Yerra, A., Mosbruger, T. L., Fleming, A. M., Burrows, C. J., and Cairns, B. R. (2019). Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U.S.A.* 116, 6784–6789. doi: 10.1073/pnas.1817334116

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Barbieri, I., et al. (2019). RNA modifications detection by comparative Nanopore direct RNA sequencing. *bioRxiv* [Preprint]. doi: 10.1101/843136

Li, H.-B., Tong, J., Zhu, S., Batista, P. J., Duffy, E. E., Zhao, J., et al. (2017). m^{6}A mRNA methylation controls T cell homeostasis by targeting the IL-7/STAT5/SOCS pathways. *Nature* 548, 338–342. doi: 10.1038/nature23450

Lin, S., and Gregory, R. I. (2014). Methyltransferases modulate RNA stability in embryonic stem cells. *Nat. Cell Biol.* 16, 129–131. doi: 10.1038/ncb2914

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m^{6}A and m^{6}Am throughout the transcriptome. *Nat. Methods* 12, 767–772. doi: 10.1038/nmeth.3453

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., et al. (2019). Accurate detection of m^{6}A RNA modifications in native RNA sequences. *Nat. Commun.* 10:4079. doi: 10.1038/s41467-019-11713-9

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3–METTL14 complex mediates mammalian nuclear RNA N^6-adenosine methylation. *Nat. Chem. Biol.* 10, 93–95. doi: 10.1038/nchembio.1432

Lorenz, D. A., Sathe, S., Einstein, J. M., and Yeo, G. W. (2019). Direct RNA sequencing enables m^6A detection in endogenous transcript isoforms at base specific resolution. *RNA* 26, 19–28. doi: 10.1261/rna.072785.119

Louloupi, A., Ntini, E., Conrad, T., and Ørom, U. A. V. (2018). Transient N-6-methyladenosine transcriptome sequencing reveals a regulatory role of m^6A in splicing efficiency. *Cell Rep.* 23, 3429–3437. doi: 10.1016/j.celrep.2018.05.077

Luo, S., and Tong, L. (2014). Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13834–13839. doi: 10.1073/pnas.1412742111

Maier, K. C., Gressel, S., Cramer, P., and Schwalb, B. (2019). Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *bioRxiv* [Preprint]. doi: 10.1101/601856

Meng, J., Cui, X., Rao, M. K., Chen, Y., and Huang, Y. (2013). Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics* 29, 1565–1567. doi: 10.1093/bioinformatics/btt171

Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., et al. (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods* 69, 274–281. doi: 10.1016/j.ymeth.2014.06.008

Meyer, K. D. (2019). DART-seq: an antibody-free method for global m^6A detection. *Nat. Methods.* 16, 1275–1280. doi: 10.1038/s41592-019-0570-0

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near Stop Codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* 7:458. doi: 10.1038/msb.2010.112

Moldován, N., Tombácz, D., Szûcs, A., Csabai, Z., Snyder, M., and Boldogkõi, Z. (2018). Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* 8:2708. doi: 10.3389/fmicb.2017.02708

Molinie, B., Wang, J., Lim, K. S., Hillebrand, R., Lu, Z., Van Wittenberghe, N., et al. (2016). m^6A-LAIC-seq reveals the census and complexity of the m^6A epitranscriptome. *Nat. Methods* 13, 692–698. doi: 10.1038/nmeth.3898

Monteuuis, G., Wong, J. J. L., Bailey, C. G., Schmitz, U., and Rasko, J. E. J. (2019). The changing paradigm of intron retention: regulation, ramifications and recipes. *Nucleic Acids Res.* 47, 11497–11513. doi: 10.1093/nar/gkz1068

Muller, S., Windhof, I. M., Maximov, V., Jurkowski, T., Jeltsch, A., Forstner, K. U., et al. (2013). Target recognition, RNA methylation activity and transcriptional regulation of the *Dictyostelium discoideum* Dnmt2-homologue (DnmA). *Nucleic Acids Res.* 41, 8615–8627. doi: 10.1093/nar/gkt634

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F. J., Rescheneder, P., et al. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* 28, 1126–1135. doi: 10.1101/gr.231100.117

Neugebauer, K. M. (2019). Nascent RNA and the coordination of splicing with transcription. *Cold Spring Harb. Perspect. Biol.* 11, a032227. doi: 10.1101/cshperspect.a032227

Neumann, T., Herzog, V. A., Muhar, M., von Haeseler, A., Zuber, J., Ameres, S. L., et al. (2019). Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics* 20:258. doi: 10.1186/s12859-019-2849-7

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., et al. (2014). Mammalian WTAP is a regulatory subunit of the RNA N^6-methyladenosine methyltransferase. *Cell Res.* 24, 177–189. doi: 10.1038/cr.2014.3

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442. doi: 10.1038/nbt.1861

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* 159, 1698–1710. doi: 10.1016/j.cell.2014.11.015

Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19:90. doi: 10.1186/s13059-018-1462-9

Ries, R. J., Zaccara, S., Klein, P., Olarerin-George, A., Namkoong, S., Pickering, B. F., et al. (2019). m^6A enhances the phase separation potential of mRNA. *Nature* 571, 424–428. doi: 10.1038/s41586-019-1374-1

Roach, N. P., Sadowski, N., Alessi, A. F., Timp, W., Taylor, J., and Kim, J. K. (2019). The full-length transcriptome of *C. elegans* using direct RNA sequencing. *bioRxiv* [Preprint]. doi: 10.1101/598763

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Sakurai, M., Yano, T., Kawabata, H., Ueda, H., and Suzuki, T. (2010). Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat. Chem. Biol.* 6, 733–740. doi: 10.1038/nchembio.434

Schwartz, S., Mumbach, M. R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G. G., et al. (2014). Perturbation of m^6A writers reveals two distinct classes of mRNA methylation at internal and 5′ Sites. *Cell Rep.* 8, 284–296. doi: 10.1016/j.celrep.2014.05.048

Shi, H., Wei, J., and He, C. (2019). Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers. *Mol. Cell* 74, 640–650. doi: 10.1016/j.molcel.2019.04.025

Smith, A. M., Abu-Shumays, R., Akeson, M., and Bernick, D. L. (2015). Capture, unfolding, and detection of individual tRNA molecules using a nanopore device. *Front. Bioeng. Biotechnol.* 3:91. doi: 10.3389/fbioe.2015.00091

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* 14:e0216709. doi: 10.1371/journal.pone.0216709

Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M. D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10:3359. doi: 10.1038/s41467-019-11272-z

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2

Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celniker, S., Neely, R. K., et al. (2016). *De novo* Identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* [Preprint]. doi: 10.1101/094672

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Lariviere, L., et al. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.* 22, 1350–1359. doi: 10.1101/gr.130161.111

Tombácz, D., Prazsák, I., Szûcs, A., Dénes, B., Snyder, M., and Boldogkõi, Z. (2018). Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* 7:giy139. doi: 10.1093/gigascience/giy139

Tong, J., Flavell, R. A., and Li, H.-B. (2018). RNA m^6A modification and its function in diseases. *Front. Med.* 12, 481–489. doi: 10.1007/s11684-018-0654-8

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., et al. (2019). Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* 29, 1545–1554. doi: 10.1101/gr.247064.118

Wachutka, L., Caizzi, L., Gagneur, J., and Cramer, P. (2019). Global donor and acceptor splicing site kinetics in human cells. *eLife* 8:e45056. doi: 10.7554/eLife.45056

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., et al. (2015). N^6-methyladenosine modulates messenger RNA translation efficiency. *Cell* 161, 1388–1399. doi: 10.1016/j.cell.2015.05.014

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., and Zhao, J. C. (2014). N^6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* 16, 191–198. doi: 10.1038/ncb2902

Wojtas, M. N., Pandey, R. R., Mendel, M., Homolka, D., Sachidanandam, R., and Pillai, R. S. (2017). Regulation of m^6A Transcripts by the 3′→5′ RNA Helicase

YTHDC2 Is essential for a successful meiotic program in the mammalian germline. *Mol. Cell* 68, 374–387.e12. doi: 10.1016/j.molcel.2017.09.021

Wongsurawat, T., Jenjaroenpun, P., Wassenaar, T. M., Wadley, T. D., Wanchai, V., Akel, N. S., et al. (2018). Decoding the epitranscriptional landscape from native RNA sequences. *bioRxiv* [Preprint]. doi: 10.1101/487819

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., et al. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv* [Preprint]. doi: 10.1101/459529

Xiang, J.-F., Yang, Q., Liu, C.-X., Wu, M., Chen, L.-L., and Yang, L. (2018). N^6-methyladenosines modulate A-to-I RNA editing. *Mol. Cell* 69, 126–135.e6. doi: 10.1016/j.molcel.2017.12.006

Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., et al. (2016). Nuclear m^{6}A reader YTHDC1 regulates mRNA splicing. *Mol. Cell* 61, 507–519. doi: 10.1016/j.molcel.2016.01.012

Xu, C., Wang, X., Liu, K., Roundtree, I. A., Tempel, W., Li, Y., et al. (2014). Structural basis for selective binding of m^{6}A RNA by the YTHDC1 YTH domain. *Nat. Chem. Biol.* 10, 927–929. doi: 10.1038/nchembio.1654

Xu, L., and Seki, M. (2019). Recent advances in the detection of base modifications using the Nanopore sequencer. *J. Hum. Genet* 65, 25–33. doi: 10.1038/s10038-019-0679-0

Yang, S., Wei, J., Cui, Y.-H., Park, G., Shah, P., Deng, Y., et al. (2019). m^{6}A mRNA demethylase FTO regulates melanoma tumorigenicity and response to anti-PD-1 blockade. *Nat. Commun.* 10:2782. doi: 10.1038/s41467-019-10669-0

Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., et al. (2018). VIRMA mediates preferential m^{6}A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov.* 4:10. doi: 10.1038/s41421-018-0019-0

Zhang, C., Chen, Y., Sun, B., Wang, L., Yang, Y., Ma, D., et al. (2017a). m^{6}A modulates haematopoietic stem and progenitor cell specification. *Nature* 549, 273–276. doi: 10.1038/nature23883

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. (2017b). Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* 18:583. doi: 10.1186/s12864-017-4002-1

Zhang, S.-Y., Zhang, S.-W., Fan, X.-N., Zhang, T., Meng, J., and Huang, Y. (2019). FunDMDeep-m^{6}A: identification and prioritization of functional differential m^{6}A methylation genes. *Bioinformatics* 35, i90–i98. doi: 10.1093/bioinformatics/btz316

Zhang, Z., Zhan, Q., Eckert, M., Zhu, A., Chryplewicz, A., De Jesus, D. F., et al. (2019). RADAR: differential analysis of MeRIP-seq data with a random effect model. *Genome Biol.* 20:294. doi: 10.1186/s13059-019-1915-9

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K. V. S. K., Gu, L., and Reddy, A. S. N. (2019). Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and nanopore-based direct RNA sequencing. *Front. Genet.* 10:253. doi: 10.3389/fgene.2019.00253

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C.-M., Li, C. J., et al. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell* 49, 18–29. doi: 10.1016/j.molcel.2012.10.015

Zhou, J., Wan, J., Gao, X., Zhang, X., Jaffrey, S. R., and Qian, S.-B. (2015). Dynamic m^{6}A mRNA methylation directs translational control of heat shock response. *Nature* 526, 591–594. doi: 10.1038/nature15377

Zhu, T., Roundtree, I. A., Wang, P., Wang, X., Wang, L., Sun, C., et al. (2014). Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N^6-methyladenosine. *Cell Res.* 24, 1493–1496. doi: 10.1038/cr.2014.152

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 20 May 2020
doi: 10.3389/fgene.2020.00398

How Do You Identify m^6A Methylation in Transcriptomes at High Resolution? A Comparison of Recent Datasets

Charlotte Capitanchik[1]*, *Patrick Toolan-Kerr*[1,2], *Nicholas M. Luscombe*[1,3,4†] and *Jernej Ule*[1,2†]

[1] The Francis Crick Institute, London, United Kingdom, [2] Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, United Kingdom, [3] Department of Genetics, Environment and Evolution, UCL Genetics Institute, London, United Kingdom, [4] Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

A flurry of methods has been developed in recent years to identify N6-methyladenosine (m^6A) sites across transcriptomes at high resolution. This raises the need to understand both the common features and those that are unique to each method. Here, we complement the analyses presented in the original papers by reviewing their various technical aspects and comparing the overlap between m^6A-methylated messenger RNAs (mRNAs) identified by each. Specifically, we examine eight different methods that identify m^6A sites in human cells with high resolution: two antibody-based crosslinking and immunoprecipitation (CLIP) approaches, two using endoribonuclease MazF, one based on deamination, two using Nanopore direct RNA sequencing, and finally, one based on computational predictions. We contrast the respective datasets and discuss the challenges in interpreting the overlap between them, including a prominent expression bias in detected genes. This overview will help guide researchers in making informed choices about using the available data and assist with the design of future experiments to expand our understanding of m^6A and its regulation.

Keywords: RNA, N6-methyladenosine, m^6A, epitranscriptomics, bioinformatics

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Italian Institute of Technology (IIT), Italy

Reviewed by:
Miguel Angel García-Campos,
Weizmann Institute of Science, Israel
Kate Meyer,
Duke University, United States

****Correspondence:***
Charlotte Capitanchik
charlotte.capitanchik@crick.ac.uk

† These authors have contributed equally to this work

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *28 January 2020*
Accepted: *30 March 2020*
Published: *20 May 2020*

Citation:
Capitanchik C, Toolan-Kerr P, Luscombe NM and Ule J (2020) How Do You Identify m^6A Methylation in Transcriptomes at High Resolution? A Comparison of Recent Datasets. Front. Genet. 11:398. doi: 10.3389/fgene.2020.00398

INTRODUCTION

N6-methyladenosine (m^6A) is the most abundant internal modification of messenger RNA (mRNA), occurring ubiquitously across the tree of life. In mammals, m^6A is thought to be deposited cotranscriptionally by the METTL3–METTL14–WTAP complex, with METTL3 being the catalytically active methyltransferase (Ke et al., 2017; Bertero et al., 2018). There is a strong enrichment for this modification within a degenerate DRACH sequence context (D = A, G, or U; R = A or G; H = A, C, or U), with early chromatographic studies suggesting a core RAC motif (Wei and Moss, 1977). The knockout of *METTL3* is embryonic lethal in mice, indicating its critical role in regulating mammalian development (Geula et al., 2015): the modification is implicated in diverse cellular processes such as differentiation, meiosis, circadian rhythms, and proliferation in cancer (Fustin et al., 2013; Schwartz et al., 2013; Batista et al., 2014; Geula et al., 2015; Cui et al., 2017). As a posttranscriptional regulator, m^6A is especially interesting in the context of neurons, where it can potentially regulate localized translation (Merkurjev et al., 2018; Shi et al., 2018). The best understood mechanism of m^6A

function is via the direct binding of YTH domain proteins, which target m^{6}A-containing transcripts for nuclear export, translation, and decay (reviewed in Patil et al., 2018).

To develop a detailed understanding of how m^{6}A dictates mRNA fate, we need to determine exactly which mRNA sites are m^{6}A modified in a given biological system. To this end, high-throughput approaches have been developed to map m^{6}A transcriptome-wide (**Table 1**). However, the modification presents significant challenges, as reverse transcription of native m^{6}A nucleotides using common reverse transcriptases does not yield a specific mutational or truncation-based signature, unlike other RNA modifications.

Here, we provide a brief technical overview of the major methods to identify m^{6}A transcriptome-wide at single nucleotide, or near single nucleotide, resolution highlighting the respective advantages and drawbacks of each method. Furthermore, by comparing genes identified by each method, we begin to explore their resulting datasets.

Antibody-Based Methods

The first described methods for transcriptome-wide profiling of m^{6}A were m^{6}A-seq and MeRIP-seq. These methods use an antibody for m^{6}A to perform RNA immunoprecipitation, followed by next generation sequencing (NGS) (Dominissini et al., 2012; Meyer et al., 2012). However, the resolution of m^{6}A-seq is limited to the size of RNA fragments, with no objective way of determining where in the fragment the modification occurred. Greater resolution was achieved by UV crosslinking the antibody to RNA, following the principles of the crosslinking and immunoprecipitation (CLIP) protocol (König et al., 2010). Such approaches were simultaneously developed in the laboratories of Samie Jaffrey and Robert Darnell, named miCLIP and m^{6}A-CLIP, respectively (**Figure 1A**; Ke et al., 2015; Linder et al., 2015). Here, purified RNA is incubated *in vitro* with an m^{6}A antibody. Following immunoprecipitation, the antibody is digested with proteinase K, leaving an amino acid adduct attached to the RNA base. During preparation of the complementary DNA (cDNA) library, the reverse transcriptase either reads through this crosslinked adduct, causing a substitution or deletion mutation, or is stopped, resulting in cDNA truncation. These signals can be analyzed computationally to identify the modification site at single nucleotide resolution (Haberman et al., 2017). The Jaffrey group found that antibodies differed in their propensities to introduce a mutation or truncation and in the positions of these signals in relation to the modified adenosine. The authors concluded that the polyclonal Abcam and Synaptic Systems antibodies were most efficient at immunoprecipitating and gave the most predictable mapping signatures; as a result, they remain the most commonly used antibodies in subsequent miCLIP publications.

N6-methyladenosine-crosslinking and immunoprecipitation is conceptually similar to miCLIP but requires preparation of multiple libraries and has so far exclusively used the Synaptic Systems antibody. Two sequencing libraries are prepared from the same sample: one using the MeRIP-seq approach to identify m^{6}A-modified oligonucleotides and one using the miCLIP approach, which is then analyzed to identify both reverse transcription read-through and truncation events. These signals are then filtered to retain only those that overlap with peaks from the MeRIP-seq library. In this way, the authors claimed greater specificity in identifying true modification sites. The protocol differs from the miCLIP protocol in several additional ways; for example, size selection of RNA fragments prior to immunoprecipitation and a bromodeoxyuridine (BrdU) cDNA-purification approach. There are also differences in the starting RNA/antibody ratios—miCLIP uses an excess of RNA, whereas m^{6}A-CLIP uses an excess of antibody.

A major drawback with these approaches is the promiscuity of m^{6}A antibodies; for example, some interact with m^{6}Am, which is found as the first nucleotide after the cap in certain mRNAs (Schwartz et al., 2013; Linder et al., 2015). Devising appropriate methods to eliminate false positives is challenging. Studies generally tackle this issue by only reporting sites found within the consensus DRACH motif or by perturbing methyltransferase activity. Neither is optimal: DRACH-only reporting prevents discovery of m^{6}A in RAC or noncanonical motifs, whereas knockout or knockdown controls exclude sites that can be modified by another methyltransferase. Furthermore, disrupting the m^{6}A machinery may introduce global changes in RNA abundance that are difficult to account for, except with the careful use of input libraries and spike-ins (Liu et al., 2020).

Finally, methods that depend on crosslink-induced mutations as the readout—as opposed to truncations—may be more susceptible to gene expression changes because higher read coverage is required to call sites. Additionally, for all strategies, the necessary integration of multiple control datasets (methyltransferase depletion, RNA input, etc.) increases the variance in the experimental design, reducing the statistical power to call sites. In summary, although antibody-based methods have been fundamental to paving the way for transcriptomic analysis of m^{6}A and remain the most common way to survey the modification, issues with antibody specificity make orthogonal approaches desirable.

Enzyme-Based Methods

In 2017, the MazF endoribonuclease was described, which cuts RNA within an ACA sequence motif, but with greater preference for ACA over m^{6}A-CA sites (Imanishi et al., 2017). Thus, m^{6}A-modified sites, usually present within a DRACH motif, can be detected as a reduction in MazF cleavage efficiency. Two new methods, MAZTER-seq and m^{6}A-REF-seq (**Figure 1B**) developed by the laboratories of Schraga Schwartz and Guan-Zheng Luo, respectively, showed how this enzyme can be used to map m^{6}A at single-nucleotide resolution (Garcia-Campos et al., 2019; Zhang Z. et al., 2019).

In both approaches, purified mRNA is treated with the MazF enzyme, leaving RNA fragments containing an ACA site at the 5′ end and finishing just before the next ACA motif within the transcript. After sequencing, any ACA sequences present within a read indicate an uncut and, therefore, modified site. The main advantage of this approach is that it can provide stoichiometric information on the m^{6}A modification, based on the cut/uncut ratio of reads for every ACA site, something the antibody-based methods currently lack.

TABLE 1 | Single nucleotide resolution, transcriptome-wide methods for detecting m^6A.

Method type	Method	Cell lines (human)	Strengths	Weaknesses	Motif restriction?	Diagnostic signature	UMI	RNA selection	References and (data access)
Antibody based	miCLIP	HEK293 MOLM13	• High throughput, can be used to assess multiple conditions • RNA can be taken from any source as crosslinking occurs *in vitro* • Reproducible data	• Difficult to correct for nonspecific antibody binding • Requires UV crosslinker • Complex library preparation • Requires high amounts of input material	DRACH	Truncations and C → T mutations	Yes	Total RNA and poly(A) selected available	Linder et al., 2015; Vu et al., 2017 (GSE98623)
	m^6A-CLIP	A549 CD8+ T cells HeLa			RRACU/RAC	Truncations and mutations (substitutions and deletions)	Yes	poly(A) HeLa—ribo0, poly(A), nucleoplasm, chromatin	Ke et al., 2015 (GSE71154); Ke et al., 2017 (GSE86336)
MazF enzyme based	MAZTER-seq	HEK293T	• Generates stoichiometric data • Semiquantitative output	• Can only detect sites in ACA sequence context • Sequence-specific biases in enzyme cutting efficiency • Complex bioinformatics analysis	ACA	Enzymatic cleavage efficiency, measured as truncations vs. read-through	No	poly(A)	Garcia-Campos et al., 2019
	m^6A-REF-seq	HEK293T			ACA		No	poly(A)	Zhang Z. et al., 2019
Fusion domain based	DART-seq	HEK293T	• Low RNA input • Simple library preparation	• Biases in background APOBEC1 targeting • Mapping is limited to YTH-recognized sites • Resolution is low compared to CLIP methods • Must express fusion construct *in vivo* for maximum efficiency	Mutation site must be C → U	C → U mutations	No	None	Meyer, 2019
In silico prediction	WHISTLE	Any	• Can predict m^6A sites in any gene, regardless of expression	• Trains based on CLIP datasets, so will learn CLIP biases	RRACH	Truncations and mutations	Yes	poly(A)	Chen et al., 2019 (http://180.208.58.19/whistle/download.html)
Direct RNA sequencing by Nanopore	MINES	HEK293	• Potential for measuring stoichiometry of sites and combinatorial modification dynamics (although currently not systematically implemented)	• Trains based on CLIP datasets, so will learn CLIP biases	RGACH	Tombo's fraction modified values and coverage files	NA	poly(A)	Lorenz et al., 2019
	NanoCompore	MOLM13	• Can detect other modifications as well as m^6A • Potential for measuring stoichiometry of sites and combinatorial modification dynamics (although currently not systematically implemented)	• Currently low throughput • High input requirements • Requires a low or no methylation control, which might be difficult to obtain	No	Difference in k-mer current intensity and dwell time in pore between WT and METTL3 KD control	NA	poly(A)	Leger et al., 2019

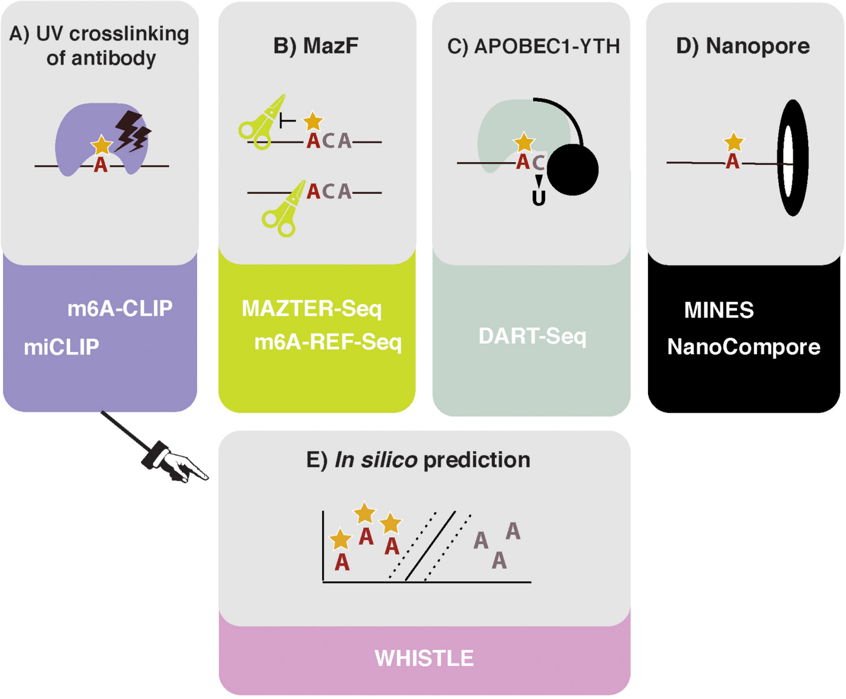


FIGURE 1 | High throughput methods to detect or predict m^{6}A in transcriptomes. **(A)** Crosslinking and immunoprecipitation (CLIP) methods involve UV crosslinking of the m^{6}A antibody to purified RNA. m^{6}A-CLIP and miCLIP differ in the antibodies used, complementary DNA (cDNA) library preparation, and computational processing, among other differences. **(B)** MazF *Escherichia coli* endoribonuclease preferentially cuts at nonmethylated ACA sites. This forms the basis of MAZTER-seq and m^{6}A-REF-seq. **(C)** DART-seq expresses an APOBEC1-YTH fusion protein. The YTH domain targets APOBEC1 to m^{6}A sites, where it deaminates surrounding cytosines to uracil. **(D)** Direct RNA sequencing with Nanopore technologies facilitates detection of m^{6}A due to differences in ionic current intensities between A- and m^{6}A-containing sequences and dwell time in the pore. Methods differ by how these signals are deconvolved. m^{6}A identification using nanopore sequencing (MINES) is a combination of four random forest models, pretrained using CLIP m^{6}A sites as true positives. NanoCompore relies on a comparison in signal between two conditions, for example wild type (WT) and METTL3 knockdown, or *in vivo* RNA vs. nonmodified *in vitro* transcribed RNA. **(E)** *In silico* prediction of m^{6}A sites is performed by WHISTLE, a support vector machine algorithm that uses miCLIP and m^{6}A-CLIP sites as training data.

Nevertheless, due to the specific attributes of the MazF enzyme, careful quality control in calculating m^{6}A stoichiometry is required. In MAZTER-seq, potential m^{6}A sites are prefiltered to remove any ACA sequences that are too close to each other to be accurately measured. Furthermore, reads that do not begin and end within a cleaved ACA sequence are removed, as they could occur through random RNA fragmentation or nonspecific cutting. Finally, for a subset of analyses, ACA sites containing a G at the +3 position are removed, as this impairs MazF cleavage efficiency. The authors calculate that, theoretically, 25% of DRACH sites in yeast and 16% in mammals can be quantified using MAZTER-seq. In contrast, m^{6}A-REF-seq does not apply filters based on incorrect read endings or calculations of the minimal ACA proximity; instead, ACA sites predicted to be in double-stranded RNA regions are discarded, as they are considered to alter cutting efficiency. Furthermore, for a site to be called, the authors require a decrease in the modification ratio >10% when the RNA is treated with the demethylase enzyme FTO.

In addition to calculating stoichiometric ratios of CLIP-annotated m^{6}A sites, MAZTER-seq was used to identify previously unknown m^{6}A sites. This was achieved by comparing cleavage efficiencies within DRACH motifs in three different control scenarios. The first was between WT and m^{6}A methyltransferase deletion input libraries, the second was m^{6}A-IP with the same strains, and the third, a comparison between input and m^{6}A IP WT conditions. In this way, the authors classified all published sites into confidence groups and found a number of previously unannotated sites within the high-confidence groups. Crucially, this suggests that probable m^{6}A sites have been missed by antibody-based methods.

MazF clearly enables valuable approaches to calculate m^{6}A stoichiometry at a focused set of sites, validate previously identified m^{6}A sites, and identify a number of novel sites. The limitation of the MazF enzyme to ACA sites and the extensive filtering requirements do mean, however, that these methods alone cannot provide a full transcriptome-wide map of m^{6}A. Nonetheless, the careful work to identify and quantify the biases inherent in this system is of great value in developing high-confidence m^{6}A maps and offers an important orthogonal method to other transcriptome-wide mapping approaches.

Fusion Domain-Based Methods

DART-seq employs the *in vivo* expression of a YTH protein domain fused to the APOBEC1 enzyme (**Figure 1C**;

Meyer, 2019). The YTH domain was identified in numerous studies as the major "reader" of the m^{6}A modification (Zaccara et al., 2019), whereas the APOBEC1 enzyme deaminates cytosine to uracil, which can be detected as a mutation compared with a reference sequence. Thus, this construct allows deamination of cytosine residues in the vicinity of m^{6}A sites recognized by YTH. Previous studies suggest that m^{6}A is invariably followed by cytosine (Wei et al., 1976), raising the possibility of single-nucleotide resolution mapping, although in practice, more distant cytosines are also modified.

The most notable benefit is the low input requirements: libraries can be made with as little as 10 ng of total RNA as starting material. Additionally, as the YTH-APOBEC1 construct can be transiently expressed in cells, library preparation is much more straightforward than either the antibody- or enzyme-based methods, since no treatment of the RNA is required to identify the m^{6}A signal following extraction. Owing to targeting by the major m^{6}A reader, it is also possible that DART-seq will identify more functionally relevant m^{6}A sites than other methods. One possible drawback is that the APOBEC1 enzyme displays sequence preferences: expressed alone, it modifies cytosine residues in the 3′ untranslated region (UTR), making it difficult to detect confidently in this region, while ∼70% of APOBEC1-only deaminated sites are preceded by an adenosine (**Supplementary Figure 6C** from Meyer, 2019), meaning that using APOBEC1 and APOBEC1-YTH mutant as a control is likely to result in false negatives.

Direct Sequencing-Based Methods

Ideally, it would be possible to detect m^{6}A via direct RNA sequencing. Pore-based sequencers measure changes in an ionic current as nucleic acids pass through a nanopore: information about changes in current and dwell time in the pore is used to identify the nucleotide in question. Several publications demonstrated that RNA modifications produce specific current and dwell time signals, suggesting nanopore-based methods could identify modified nucleotides in a high throughput manner (**Figure 1D**; Garalde et al., 2018; Workman et al., 2018; Smith et al., 2019). The potential benefits of this approach for mapping RNA modifications are huge, as stoichiometric and positional information of multiple modifications could be interpreted simultaneously. The reality of deconvolving the raw signal to infer m^{6}A sites, however, is not straightforward.

The first application of the Oxford Nanopore technology (Nanopore) to detect m^{6}A in a whole transcriptome examined yeast mRNA (Liu et al., 2019). The authors trained a support vector machine (SVM), called EpiNano, on Nanopore sequencing data of synthetic transcripts containing m^{6}A residues in every possible 5-mer combination to identify the most informative signals that distinguish m^{6}A from other nucleotides. Surprisingly, the raw current intensities alone were found to be poor predictors of methylation status; instead, the selected training features included mean per-base quality, mismatch frequency, and deletion frequency. The model achieved ∼90% prediction accuracy for the training dataset. It was then used to recover 363 previously identified, high-confidence m^{6}A sites, previously identified using m^{6}A-seq, which it was able to do with 87% accuracy.

An alternative approach, m^{6}A identification using nanopore sequencing (MINES), was used to create the first Nanopore-based m^{6}A transcriptome for humans (Lorenz et al., 2019). This method applied Tombo, a program that was previously developed to detect *de novo* modifications in Nanopore DNA-sequencing data based on base-calling errors (Oxford Nanopore Technologies, 2018). The authors trained random forest models using the Tombo modification values to classify the m^{6}A status of four RGACH motifs. Those RGACH sites overlapping with HEK293 miCLIP and HeLa m^{6}A-CLIP sites (Linder et al., 2015; Ke et al., 2017) were labeled as true positives in the training data, and the models achieved an average accuracy of 79%, representing 35% of m^{6}A sites identified with CLIP-based methods (in part due to the motif restriction). The authors then predicted 13,034 novel RGACH m^{6}A sites, which were validated by METTL3 knockdown.

A further approach is NanoCompore (Leger et al., 2019), which compares Nanopore signals between two datasets and therefore does not require a training dataset. Specifically, this is achieved by contrasting the median current intensities and dwell times of k-mers between the experiment and a control with perturbed modifications (e.g., wild type vs. knockdown, or *in vitro* modified vs. unmodified controls). To identify METTL3-dependent m^{6}A sites, the authors processed polyA+-selected RNA sequencing data from wild-type and METTL3 short-hairpin RNA (shRNA) knockdown MOLM13 cells. NanoCompore is not restricted to m^{6}A and can be readily extended to other modifications that have a reliable control. A major advantage is that it avoids being biased by the accuracy of previous mapping methods to train the models, as site identification is instead determined by the sensitivity to a specific modification enzyme. Of course, the dependence on a comparison between samples is a limitation, as reliable controls are currently unavailable for many modifications and biological systems, and specific sites or RNA species are often modified by distinct enzymes. As a result, there is probably a reduced risk of false-positive site assignment at the cost of sensitivity.

Finally, a simplified approach was recently published for the *Arabidopsis thaliana* transcriptome (Parker et al., 2020), in which the base-calling error rate was used as the sole parameter for identifying m^{6}A sites. The authors compared the transcriptomes for a *vir-1* mutant, an Arabidopsis m^{6}A methyltransferase, with a *vir-1* restored line, identifying ∼17,000 sites with an error rate twofold greater in the control line compared to mutant. Taking this approach 66% of identified m^{6}A sites fell within five nucleotides of a miCLIP peak.

The above methods demonstrate that direct RNA sequencing can be used to detect m^{6}A. A common limitation pertains to the resolution and accuracy of modification assignment for transcripts with low sequencing depth. However, with third-generation sequencing technologies developing rapidly, the benefits of using direct sequencing to map RNA modifications—such as the possibility of correlating modifications with other transcriptomic features within a single RNA molecule, and

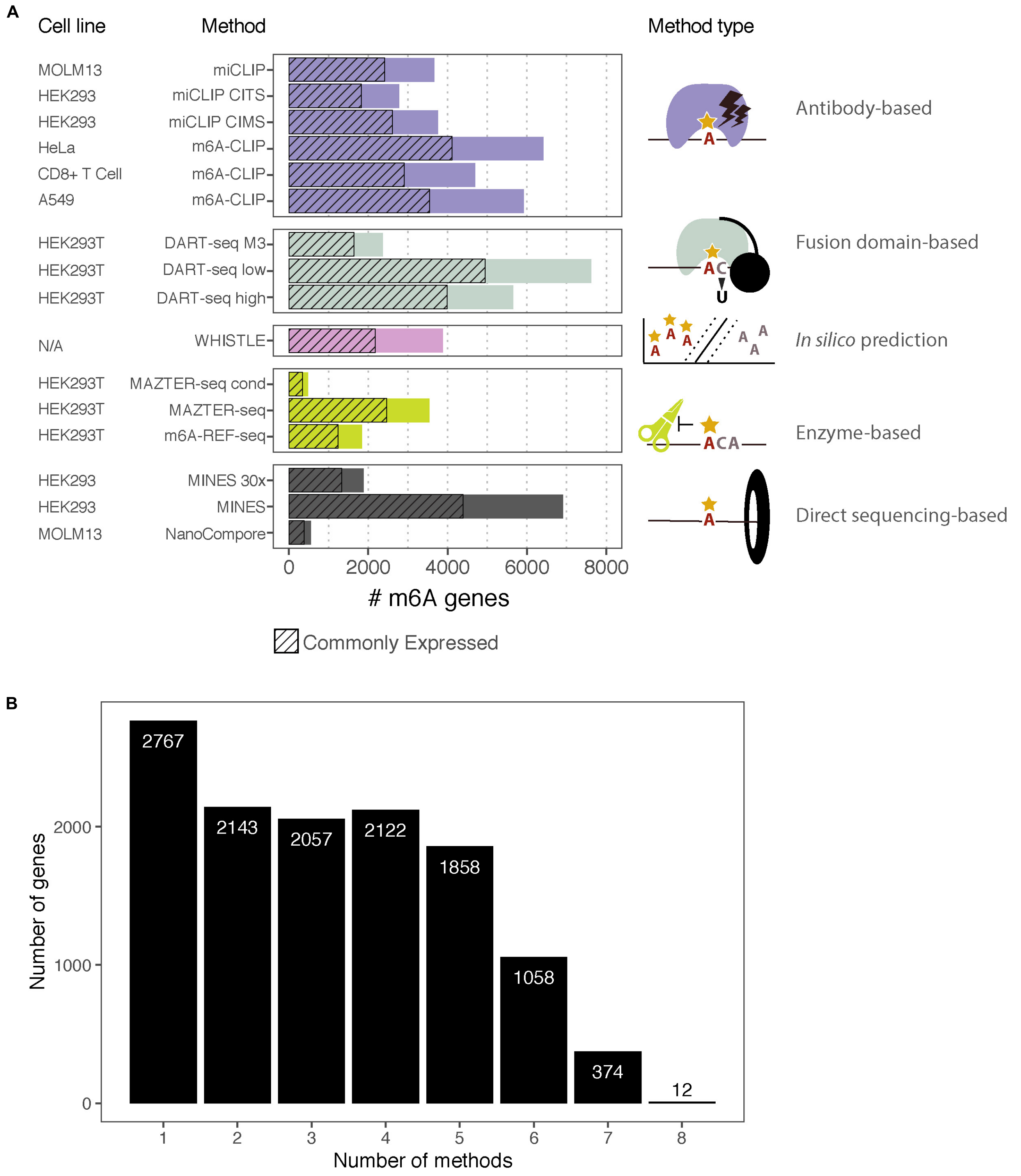


FIGURE 2 | m^{6}A-containing genes identified by eight methods. **(A)** Bar chart showing the number of m^{6}A-containing transcripts identified by each method. Some methods have data from multiple cell lines or apply several possible thresholds, which are shown separately. The cell lines for each dataset are indicated along with the type of method. The hashed bars denote genes that are commonly expressed between all the cell lines considered here. For DART-Seq, MAZTER-Seq, and MINES, several thresholds were possible: "DART-Seq M3" refers to sites identified by comparison with METTL3 knockdown. "Low" and "high" refer to two stringency thresholds applied by the authors. "MAZTER-Seq" refers to all sites with a cleavage efficiency $<50\%$, and "MAZTER-Seq cond" refers to FTO overexpression, WT $\geq 20\%$, and/or Alkbh5 overexpression, WT $\geq 20\%$. "MINES" refers to all sites identified by MINES, and "MINES 30×" refers to MINES sites with $\geq = 30\times$ coverage. **(B)** Bar chart showing the numbers of overlapping target genes between the eight methods, considering all the reported genes.

accurately calculating m^{6}A stoichiometry genome-wide—are likely to push the boundaries of the field.

In silico Prediction

Even in the best circumstances, experiments are still costly and time consuming to run and can only identify m^{6}A sites that are present in the prepared sample. *In silico* prediction offers the potential of identifying all possible m^{6}A sites (**Figure 1E**). However, algorithms rely on two critical factors: (i) the reliability of the training data and (ii) the ability to identify and encode relevant features indicating m^{6}A presence into the model. Existing approaches either use SVMs (methyRNA—Chen et al., 2017; RNAMethPre—Xiang et al., 2016; WHISTLE—Chen et al., 2019) or random forest models (RF; SRAMP—Zhou et al., 2016) to classify whether or not an adenosine is modified. The benefits of a machine-learning model, over other modeling approaches, is that predictive features do not have to be selected *a priori*. Indeed, the learned weighting of features in a model can aid our mechanistic understanding of methylation. The authors of WHISTLE (whole-transcriptome m^{6}A site prediction from multiple genomic features) showed that nucleotide sequence was the most important predictor of m^{6}A but that 14 other genomic features also contributed. Among the top features was the site being in a long exon, which was previously found to be a defining characteristic of sites measured using m^{6}A-CLIP (Dominissini et al., 2012; Ke et al., 2017). WHISTLE achieved an area under the curve of 0.948 when tested against previously unseen CLIP data.

Currently, all *in silico* m^{6}A models use antibody-based methods as training data and so will also learn the biases present in them. To continue improving predictions, it will be important to generalize models by training on orthogonal datasets.

ASSEMBLING A DATASET TO COMPARE DETECTED AND PREDICTED m^{6}A TRANSCRIPTS

The rapid expansion in orthogonal methods for transcriptomic m^{6}A detection offers an opportunity to compare the published datasets. We assembled the processed data produced by eight high-resolution methods using human cells: two antibody-based CLIP approaches (miCLIP, m^{6}A-CLIP); two endoribonuclease MazF-based (MAZTER-seq, m^{6}A-REF-seq); one deamination approach (DART-seq); two using Nanopore direct RNA sequencing (MINES, NanoCompore); and finally, one based on computational predictions (WHISTLE). Here, we examine the overlap between these methods at the level of transcripts, focusing on a single representative transcript per gene. We include only sites with a matching DRACH motif, although some datasets have additional restrictions (such as MazF “ACA,” WHISTLE “RRACH,” and MINES “RGACH”). In total, we consider 134,470 unique sites in 12,391 mRNAs (**Figures 2A,B**; sites per gene are summarised in **Supplementary Data Sheet S1**).

Filtering for Commonly Expressed Genes

Since there is not a single cell line that is used across all of the methods, we focused on commonly expressed mRNAs. For studies with no accompanying gene expression data, we accessed published RNA-seq measurements for equivalent cells lines from the EBI Expression Atlas (HEK293, HEK293T) and the Gene Expression Omnibus (MOLM13) (accession numbers listed in **Table 2**) (Edgar et al., 2002; Papatheodorou et al., 2018). For HEK293 and HEK293T, raw counts were assigned to the longest annotated transcript obtained from Ensembl BioMart v98 for GRCh38.p13, and transcripts per million (TPM) were calculated as expression measurements (Kinsella et al., 2011). For MOLM13 and HeLa, processed expression measurements were available as fragments per kilobase of transcript per million (FPKM) values. For A549 and CD8+ T cell, we used the matched poly-A sequencing data from the m^{6}A-CLIP study. BedGraph files were downloaded, and coordinates were lifted over to hg19 using UCSC liftOver (Kuhn et al., 2013). Poly(A) sites were assigned to genes using `bedtools closest -s -id -a stdin -b../hg19_mRNA_annotation.gtf -D a` (Quinlan and Hall, 2010) with a threshold of 2,000 nt from the end of the annotated 3′ UTR. Expression was quantified as read counts per transcript. Expression values were visualized in histograms, with most cell lines displaying bimodal distributions allowing a straightforward separation of expressed and unexpressed genes. For A549 and CD8+ T cells, which displayed unimodal distributions, we applied an arbitrary threshold of five counts. Finally, for each cell type, we assigned expressed genes into deciles according to their expression values.

The procedure yielded between 8,235 and 12,968 expressed genes for each cell line (**Table 2**). Transcripts that were detected by the m^{6}A measurement, but not RNA-seq, were assigned *post hoc* to the lowest expression decile of the cell line in question. In total, we considered 6,585 genes with commonly expressed transcripts across six cell lines.

Comparison of the Top-Ranking Transcripts Between Methods

The eight m^{6}A studies applied very different, and in some cases arbitrary, thresholds leading to large differences in the numbers

TABLE 2 | Number of expressed genes per cell line and origin of the expression dataset.

Cell line	Number of genes expressed	Accession	References
HEK293	11,018	E-GEOD-44384 (EBI Expression Atlas)	Hussain et al., 2013
HEK293T	11,703	E-MTAB-7029 (EBI Expression Atlas)	Doumpas et al., 2019
MOLM13	12,968	GSE114111 (GEO)	Pei et al., 2018
HeLa	12,839	GSM2300445 (GEO)	Ke et al., 2017—m^{6}A-CLIP paper
A549	9,963	GSM1828600 (GEO)	Ke et al., 2015—m^{6}A-CLIP paper
CD8T+	8,235	GSM1828598 (GEO)	Ke et al., 2015—m^{6}A-CLIP paper

TABLE 3 | Number of m^{6}A modified transcripts for each method following thresholding.

Method	Sample	Thresholding	Number of transcripts	Number of total transcripts for method	Number transcripts (**6,585 commonly expressed genes subset**)
miCLIP	CIMs HEK293	As from paper	3,755	6,282	4,000
	CITs HEK293	As from paper	2,779		
	MOLM13	As from paper	3,662		
m^{6}A-CLIP	A549	As from paper	5,915	8,560	4,694
	CD8+ T cell	As from paper	4,697		
	HeLa	As from paper	6,415		
DART-seq	High stringency HEK293T	C > U events from paper filtered for DRACH motif	5,648	8,331	5,445
	Low stringency HEK293T	C > U events from paper filtered for DRACH motif	7,614		
	WT vs. METTL3 depleted HEK239T	C > U events from paper filtered for DRACH motif	2,370		
m^{6}A-REF-seq	HEK293T	As from paper	1,843	1,843	1,243
MAZTER-seq	HEK293T	MazF cleavage efficiency < 50%	3,545	3,705	2,568
	HEK293T	FTO overexpression, WT ≥ 20%, and/or Alkbh5 overexpression, WT ≥ 20%	482		
WHISTLE	Trained on miCLIP and m^{6}A-CLIP	Posterior probability of being m^{6}A ≥ 0.95	3,877	3,877	2,177
MINES	Nanopore	As from paper	6,910	6,910	4,390
	Nanopore	Filtered for 30× coverage (threshold for NanoCompore)	1,883		
NanoCompore	WT vs. METTL3 KO Nanopore	DRACHs within clustered 5-mers with contextual $p < 0.001$	556	556	387

of reported targets. In comparing the results, we found that studies reporting greater numbers of m^{6}A targets tended to have better overlaps with other studies (data not shown), making them appear ostensibly more reliable; however, it is also possible that those methods suffer from higher false-positive rates.

To facilitate comparisons, we focused on the top ∼1,000 m^{6}A modified transcripts for each method (**Table 4**). We wished to use "modification scores" for each study to identify thresholds that produce similar numbers of top-ranking targets; however, scores are not available for all methods, so instead, we ordered genes according to the number of detected m^{6}A sites per transcript. NanoCompore reported only 387 transcripts that met our expression criteria, due to the lower sequencing throughput, the stringent requirement for 30× coverage over sites, and restriction to sites that change between wild type and METTL3 knockdown cells. In total, we considered 3,875 top-ranking transcripts among genes that are commonly expressed across all cell lines, with a total of 73,914 unique m^{6}A sites.

Of the 3,875 transcripts across all methods, 55% (2,121) are identified as m^{6}A modified by at least two, 31% (1,213) by at least three, and 16% (619) by four or more methods (**Figure 3A**). Hierarchical clustering shows that methods of the same type cluster together, indicating that they are more likely to detect similar targets (**Figure 3B**); however, the shallowness of the dendrogram highlights that despite this, distinct methods tend to differ greatly in their outputs. WHISTLE and MINES cluster with the CLIP-based methods, reflecting the underlying training datasets. MAZTER-seq and m^{6}A-REF-seq also cluster but share little overlap (40% of MAZTER-seq sites and 33% of m^{6}A-REF-seq sites overlapped with each other). The method with the highest proportion of unique genes is NanoCompore (48%), followed by m^{6}A-REF-seq (26%). The method with the lowest proportion of unique genes is m^{6}A-CLIP (10%), which suggests its sites could be the most reliable (**Figure 3C**).

In general, the higher the expression, the more likely a transcript is to be identified by multiple methods (**Figure 3D**); this is expected as most of the experimental methods described here are biased toward highly expressed genes. In this regard, NanoCompore displays the largest expression dependence (**Figure 3E**). Interestingly, miCLIP shows a greater preference for highly expressed genes compared with m^{6}A-CLIP, perhaps due to differences in starting RNA/antibody ratios in the immunoprecipitation step. In conclusion, the low overlap

TABLE 4 | Number of top-ranking targets selected per method.

Method	Number of transcripts
DART-seq	1,019
m^{6}A-CLIP	1,072
m^{6}A-REF-seq	1,243
miCLIP	1,233
NanoCompore	387
WHISTLE	1,198
MINES	1,104
MAZTER-seq	944

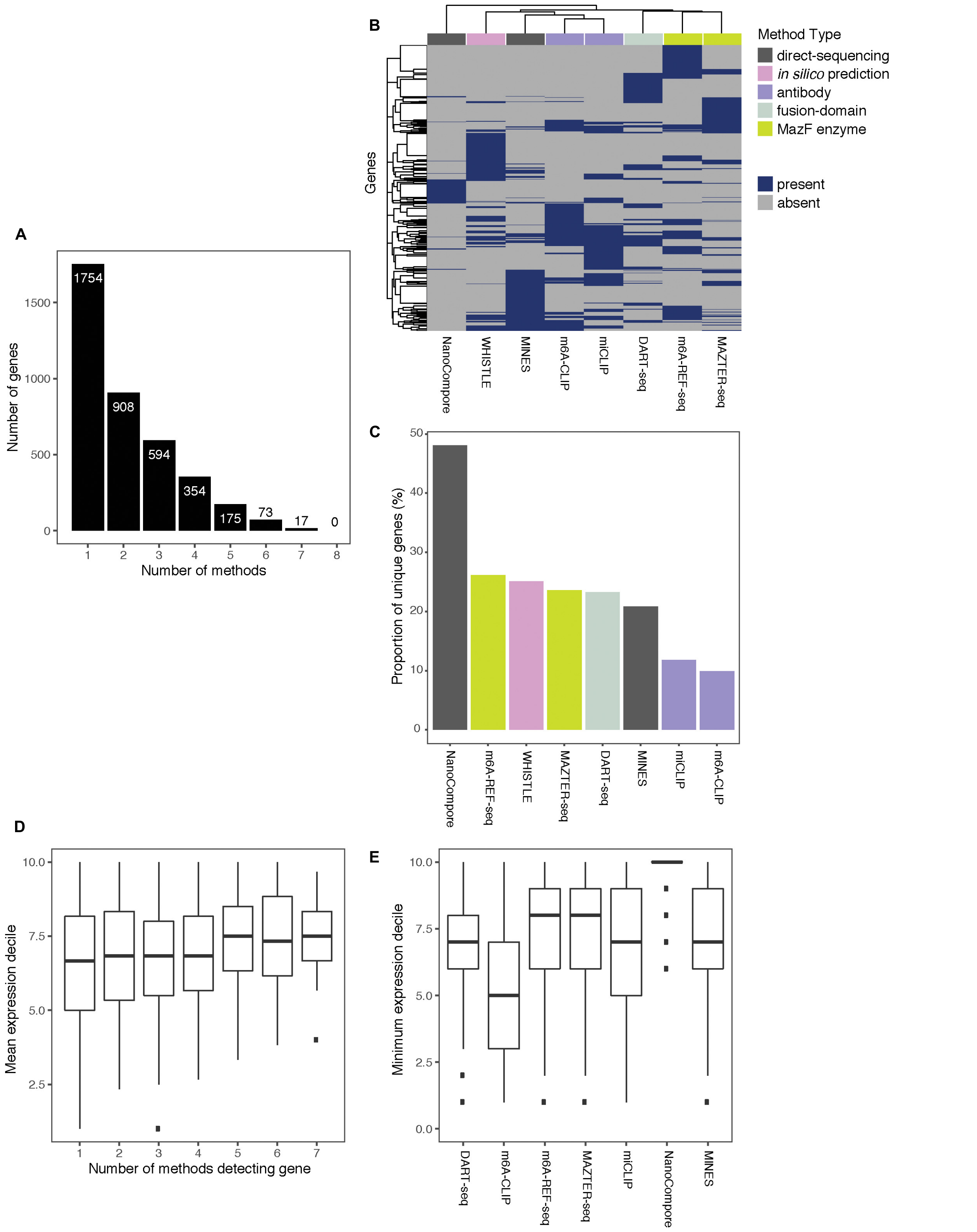


FIGURE 3 | Comparing the top-ranking target genes identified by eight methods. **(A)** Bar chart showing the numbers of top-ranking genes that overlap between the eight methods. **(B)** Heatmap showing overlap between the top targets. Dendrograms are produced by complete-linkage hierarchical clustering using the Jaccard index as the distance metric. Dark blue indicates presence of the gene among the top targets for a method, and gray indicates absence. Colored bars denote the category of the method. **(C)** Proportions of top targets that are unique to each method. **(D)** Number of methods detecting a target gene plotted against its mean expression decile across all studied cell lines. **(E)** Minimum expression deciles for the top ranked genes were plotted for each method.

between methods may arise partly from the expression-linked bias in m^{6}A detection and additional technical aspects of each method leading to different subsets of DRACH sites being detected.

DISCUSSION

Our analysis suggests that data coverage and mRNA expression are among the main biases for m^{6}A detection. With sufficient coverage, potential sites of m^{6}A modification can be detected in most mRNAs. However, in the absence of a gold standard, it is not possible at this point to estimate the false-positive rate of any single method for m^{6}A detection nor of integrated datasets. This will be important moving forward because it is clear that different studies display varying degrees of overlap. Determining the reasons behind this is valuable for the community, especially as several databases now give users access to repositories of miCLIP data (CVm^{6}A—Han et al., 2019; m^{6}AVar—Zheng et al., 2017) and algorithms trained on such data are being used to make conclusions about the functionality and disease relevance of m^{6}A sites (m^{6}AVar—Zheng et al., 2017; Deep-m^{6}A—Zhang S.-Y. et al., 2019; m^{6}Acomet—Wu et al., 2019; DeepM^{6}ASeq—Zhang and Hamada, 2018). Predictions will be limited by the validity of the training data, and it will be interesting to see how data from the newer non-antibody-based methods can be incorporated into such efforts.

In this review article, we performed analyses at the gene level as a tentative step to give the reader a broad perspective of the data types that are available for studies of m^{6}A RNA modifications. An important aspect for further analyses will be to compare individual sites within a transcript across methods, experimental conditions, and variants of DRACH motif. In this way, it will be possible to address the positional or sequence biases of methods, compare the dynamics of m^{6}A sites between conditions, cells or cellular compartments, and assess the modification rates of different DRACH sites. Such analysis could be approached in various ways, taking into account variable distances between sites assigned by different techniques and other method-specific issues. For such analyses, the use of unique molecular identifiers (UMIs) that control for PCR biases in library preparation—integrated into CLIP-based approaches—are particularly valuable. None of the antibody-free approaches currently use UMIs; therefore, quantifications of MazF and DART-seq datasets may be affected by variable PCR duplication rates. Direct RNA sequencing with Nanopores is not affected by PCR duplication, but the shallow sequencing depth may limit quantitative comparisons across large numbers of sites.

Finally, we have examined only m^{6}A sites that occur within DRACH motifs, in line with the computational approaches used in past studies. In the future, it will be interesting to analyze noncanonical sites: currently, the technical noise is often too high to reliably include such sites and therefore appropriate controls will be needed, such as METTL3 depletion. This would also help establish the methylation status of lowly expressed genes, which generally have lower sequencing coverage.

Ultimately, untangling the benefits and biases of each method in determining m^{6}A sites is crucial for the field as we move toward further understanding the mechanism, regulation, and function of m^{6}A methylation on a transcriptomic scale.

AUTHOR CONTRIBUTIONS

JU, NL, and CC conceptualized the work. CC curated and analyzed the data and produced all tables and figures. CC, JU, and PT-K wrote the initial draft, with review and editing from NL. JU and NL supervised the work. The manuscript was finalized with input from all authors.

FUNDING

This work was supported by funding from a Wellcome Trust Joint Investigator Award to NL and JU (215593/Z/19/Z). The Francis Crick Institute also receives its core funding from Cancer Research UK (FC010110), the United Kingdom Medical Research Council (FC010110), and the Wellcome Trust (FC010110). NL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support toward the establishment of the Francis Crick Institute and receives funding from the MRC eMedLab Medical Bioinformatics Infrastructure Award (MR/L016311/1) and core funding from the Okinawa Institute of Science and Technology Graduate University. PT-K is funded by a Leonard Wolfson Doctoral Training Fellowship in Neurodegeneration.

ACKNOWLEDGMENTS

We thank Flora Lee for the critical reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00398/full#supplementary-material

REFERENCES

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., et al. (2014). m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* 15, 707–719. doi: 10.1016/j.stem.2014.09.019

Bertero, A., Brown, S., Madrigal, P., Osnato, A., Ortmann, D., Yiangou, L., et al. (2018). The SMAD2/3 interactome reveals that TGFβ controls m6A mRNA methylation in pluripotency. *Nature* 555, 256–259. doi: 10.1038/nature25784

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi: 10.1080/07391102.2016.1157761

Cui, Q., Shi, H., Ye, P., Li, L., Qu, Q., Sun, G., et al. (2017). m(6)A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep.* 18, 2622–2634. doi: 10.1016/j.celrep.2017.02.059

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Doumpas, N., Lampart, F., Robinson, M. D., Lentini, A., Nestor, C. E., Cantù, C., et al. (2019). TCF/LEF dependent and independent transcriptional regulation of Wnt/β-catenin target genes. *EMBO J.* 38:e98873. doi: 10.15252/embj.201798873

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Fustin, J.-M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., et al. (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 155, 793–806. doi: 10.1016/j.cell.2013.10.026

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi: 10.1038/nmeth.4577

Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., et al. (2019). Deciphering the 'm6A code' via antibody-independent quantitative profiling. *Cell* 178, 731–747.e16. doi: 10.1016/j.cell.2019.06.013

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., et al. (2015). Stem cells m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* 347, 1002–1006. doi: 10.1126/science.1261417

Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., et al. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome Biol.* 18:7. doi: 10.1186/s13059-016-1130-x

Han, Y., Feng, J., Xia, L., Dong, X., Zhang, X., Zhang, S., et al. (2019). CVm6A: a visualization and exploration database for m6As in cell lines. *Cells* 8:168. doi: 10.3390/cells8020168

Hussain, S., Sajini, A. A., Blanco, S., Dietmann, S., Lombard, P., Sugimoto, Y., et al. (2013). NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.* 4, 255–261. doi: 10.1016/j.celrep.2013.06.029

Imanishi, M., Tsuji, S., Suda, A., and Futaki, S. (2017). Detection of N6-methyladenosine based on the methyl-sensitivity of MazF RNA endonuclease. *Chem. Commun.* 53, 12930–12933. doi: 10.1039/c7cc07699a

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., et al. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev.* 29, 2037–2053. doi: 10.1101/gad.269415.115

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., et al. (2017). m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev.* 31, 990–1006. doi: 10.1101/gad.301036.117

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database(Oxford)* 2011:bar030. doi: 10.1093/database/bar030

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., et al. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909–915. doi: 10.1038/nsmb.1838

Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief. Bioinform.* 14, 144–161. doi: 10.1093/bib/bbs038

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Barbieri, I., et al. (2019). RNA modifications detection by comparative nanopore direct RNA sequencing. *bioRxiv*[Preprint] doi: 10.1101/843136

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* 12, 767–772. doi: 10.1038/nmeth.3453

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., et al. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.* 10:4079. doi: 10.1038/s41467-019-11713-9

Liu, J., Dou, X., Chen, C., Chen, C., Liu, C., Xu, M. M., et al. (2020). N6-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription. *Science* 367, 580–586. doi: 10.1126/science.aay6018

Lorenz, D. A., Sathe, S., Einstein, J. M., and Yeo, G. W. (2019). Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base specific resolution. *RNA* 26, 19–28. doi: 10.1261/rna.072785.119

Merkurjev, D., Hong, W.-T., Iida, K., Oomoto, I., Goldie, B. J., Yamaguti, H., et al. (2018). Synaptic N6-methyladenosine (m6A) epitranscriptome reveals functional partitioning of localized transcripts. *Nat. Neurosci.* 21, 1004–1014. doi: 10.1038/s41593-018-0173-6

Meyer, K. D. (2019). DART-seq: an antibody-free method for global m6A detection. *Nat. Methods* 16, 1275–1280. doi: 10.1038/s41592-019-0570-0

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.redox.2018.11.018

Oxford Nanopore Technologies, (2018). *Tombo: Detection of Non-Standard Nucleotides Using the Genome-Resolved Raw Nanopore Signal*. Oxford: Oxford Nanopore Technologies.

Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., et al. (2018). Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* 46, D246–D251. doi: 10.1093/nar/gkx1158

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., et al. (2020). Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m6A modification. *eLife* 9:e49658. doi: 10.7554/eLife.49658

Patil, D. P., Pickering, B. F., and Jaffrey, S. R. (2018). Reading m6A in the transcriptome: m6A-binding proteins. *Trends Cell Biol.* 28, 113–127. doi: 10.1016/j.tcb.2017.10.001

Pei, S., Minhajuddin, M., Adane, B., Khan, N., Stevens, B. M., Mack, S. C., et al. (2018). AMPK/FIS1-mediated mitophagy is required for self-renewal of human AML stem cells. *Cell Stem Cell* 23, 86–100.e6. doi: 10.1016/j.stem.2018.05.021

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., et al. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* 155, 1409–1421. doi: 10.1016/j.cell.2013.10.047

Shi, H., Zhang, X., Weng, Y.-L., Lu, Z., Liu, Y., Lu, Z., et al. (2018). m6A facilitates hippocampus-dependent learning and memory through YTHDF1. *Nature* 563, 249–253. doi: 10.1038/s41586-018-0666-1

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PloS One* 14:e0216709. doi: 10.1371/journal.pone.0216709

Wei, C. M., Gershowitz, A., and Moss, B. (1976). 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry* 15, 397–401. doi: 10.1021/bi00647a024

Wei, C. M., and Moss, B. (1977). Nucleotide sequences at the N6-methyladenosine sites of HeLa cell messenger ribonucleic acid. *Biochemistry* 16, 1672–1676. doi: 10.1021/bi00627a023

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., et al. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*[Preprint] doi: 10.1101/459529

Wu, X., Wei, Z., Chen, K., Zhang, Q., Su, J., Liu, H., et al. (2019). m6Acomet: large-scale functional prediction of individual m6A RNA methylation sites from an RNA co-methylation network. *BMC Bioinform.* 20:223. doi: 10.1186/s12859-019-2840-3

Xiang, S., Liu, K., Yan, Z., Zhang, Y., and Sun, Z. (2016). RNAMethPre: a web server for the prediction and query of mRNA m6A sites. *PloS One* 11:e0162707. doi: 10.1371/journal.pone.0162707

Zaccara, S., Ries, R. J., and Jaffrey, S. R. (2019). Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* 20, 608–624. doi: 10.1038/s41580-019-0168-5

Zhang, S.-Y., Zhang, S.-W., Fan, X.-N., Meng, J., Chen, Y., Gao, S.-J., et al. (2019). Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput. Biol.* 15:e1006663. doi: 10.1371/journal.pcbi.1006663

Zhang, Y., and Hamada, M. (2018). DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinform.* 19(Suppl. 19):524. doi: 10.1186/s12859-018-2516-4

Zhang, Z., Chen, L.-Q., Zhao, Y.-L., Yang, C.-G., Roundtree, I. A., Zhang, Z., et al. (2019). Single-base mapping of m6A by an antibody-independent method. *Sci. Adv.* 5:eaax0250. doi: 10.1126/sciadv.aax0250

Zheng, Y., Nie, P., Peng, D., He, Z., Liu, M., Xie, Y., et al. (2017). m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res.* 46, D139–D145. doi: 10.1093/nar/gkx895

Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 09 June 2020
doi: 10.3389/fgene.2020.00545

LITHOPHONE: Improving lncRNA Methylation Site Prediction Using an Ensemble Predictor

Lian Liu [1], Xiujuan Lei [1], Zengqiang Fang [1], Yujiao Tang [2], Jia Meng [2] and Zhen Wei [2]**

[1] School of Computer Sciences, Shannxi Normal University, Xi'an, China, [2] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China

N^6-methyladenosine (m^{6}A) is one of the most widely studied epigenetic modifications, which plays an important role in many biological processes, such as splicing, RNA localization, and degradation. Studies have shown that m^{6}A on lncRNA has important functions, including regulating the expression and functions of lncRNA, regulating the synthesis of pre-mRNA, promoting the proliferation of cancer cells, and affecting cell differentiation and many others. Although a number of methods have been proposed to predict m^{6}A RNA methylation sites, most of these methods aimed at general m^{6}A sites prediction without noticing the uniqueness of the lncRNA methylation prediction problem. Since many lncRNAs do not have a polyA tail and cannot be captured in the polyA selection step of the most widely adopted RNA-seq library preparation protocol, lncRNA methylation sites cannot be effectively captured and are thus likely to be significantly underrepresented in existing experimental data affecting the accuracy of existing predictors. In this paper, we propose a new computational framework, **LITHOPHONE**, which stands for **l**ong noncod**i**ng RNA me**th**ylati**o**n sites **p**rediction from sequence c**h**aracteristics and gen**o**mic i**n**formation with an **e**nsemble predictor. We show that the methylation sites of lncRNA and mRNA have different patterns exhibited in the extracted features and should be differently handled when making predictions. Due to the used experiment protocols, the number of known lncRNA m^{6}A sites is limited, and insufficient to train a reliable predictor; thus, the performance can be improved by combining both lncRNA and mRNA data using an ensemble predictor. We show that the newly developed LITHOPHONE approach achieved a reasonably good performance when tested on independent datasets (AUC: 0.966 and 0.835 under full transcript and mature mRNA modes, respectively), marking a substantial improvement compared with existing methods. Additionally, LITHOPHONE was applied to scan the entire human lncRNAome for all possible lncRNA m^{6}A sites, and the results are freely accessible at: http://180.208.58.19/lith/.

Keywords: m^{6}A, lncRNA, site prediction, epitranscriptome, ensemble model

Edited by:
*Mattia Pelizzola,
Italian Institute of Technology (IIT), Italy*

Reviewed by:
*Qi Zhao,
Liaoning University, China
Vinicius Maracaja-Coutinho,
University of Chile, Chile*

***Correspondence:**
*Xiujuan Lei
xjlei@snnu.edu.cn
Zhen Wei
zhen.wei@xjtlu.edu.cn*

Specialty section:
*This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics*

Received: *19 September 2019*
Accepted: *06 May 2020*
Published: *09 June 2020*

Citation:
Liu L, Lei X, Fang Z, Tang Y, Meng J and Wei Z (2020) LITHOPHONE: Improving lncRNA Methylation Site Prediction Using an Ensemble Predictor. Front. Genet. 11:545. doi: 10.3389/fgene.2020.00545

INTRODUCTION

RNA modifications include more than 150 different types, among which N^6-methyladenosine (m^6A) has attracted the most attention due to its universality and various biological functions (Fu et al., 2014; Liu and Jia, 2014; Meyer and Jaffrey, 2014). The m^6A RNA methylation denotes that the amino group on the sixth carbon atom of adenine is modified by a methyl group, usually occurring in the conservative sequence RRACH (R = G, A; H = A, C, or U) or GGAC (Dominissini et al., 2012). The universality of m^6A is reflected in the following two aspects. On the one hand, it appears in almost all RNA transcripts, including coding and non-coding ones (Dominissini et al., 2012; Alarcón et al., 2015b). On the other hand, it is enriched near the stop codon, 3′ untranslated regions, and the last exon region of mRNA (Liu et al., 2014, 2015). Recent studies (Alarcón et al., 2015a; Roost et al., 2015) showed that as a common molecular tag, m^6A modification is involved in many important biological processes, including RNA localization and degradation (Wang et al., 2014), RNA structural dynamics (Roost et al., 2015; Song et al., 2020), variable splicing (Wang et al., 2014), primary microRNA process (Chen et al., 2015a; Geula et al., 2015), cell differentiation and adaptation, and circadian clock regulation (Fustin et al., 2013). It is also associated with protein translation, obesity, abnormal brain development, and a few other diseases (Peng et al., 2016).

Long non-coding RNA (lncRNA) refers to a class of RNAs that have no coding potentials and are of a length >200 nucleotides (nt). Studies have shown that lncRNA plays an important role in many life activities, such as dosage compensation effect, epigenetic regulation, cell cycle regulation, and cell differentiation regulation (Qureshi et al., 2010; Peng et al., 2016). Recent epitranscriptome analysis has shown that thousands of lncRNAs contain a large number of methylation sites (Shafik et al., 2016). For example, m^6A methylation is important for the silencing or inactivation of the X chromosome gene mediated by lncRNA XIST (Patil et al., 2016). The m^6A methylation of XIST is completed by recruiting the complex composed of RBM15 (RNA-binding motif protein 15)/RBM15B-WTAP-METTL3 to the specific region of XIST, the methylation recognition protein (reader) YTHDC1 then binds to this region and recruits silencing proteins to complete the whole gene suppression process. Moreover, the m^6A methylation of MALAT1 regulates pre-RNA synthesis. It was found that MALAT1 could carry this methylation in the stem ring structure. After m^6A methylation, the binding ability of the gene to the hnRNP C protein was enhanced (Nian et al., 2015). In addition, m^6A methylation can regulate lncRNA FOXM1-AS to promote the proliferation of cancer cells (Zhang et al., 2017; Song et al., 2020), and regulate lncRNA1281 to affect the differentiation of mouse embryonic stem cells (Yang et al., 2018).

With the development of high-throughput sequencing (HTS) technology, a new field of epitranscriptome analysis has emerged. The invention of MeRIP-Seq in 2012 (Meyer et al., 2012) presented the first technique to detect the m^6A spectrum in the whole transcriptome, during which RNA was randomly fragmented into short pieces of around 100 nt long; the fragments containing methylation modification were captured using the specific antibodies, and then subjected to sequencing to generate the IP samples; meanwhile, an input control sample was generated in parallel to serve as the background. Tools like MACS (Zhang et al., 2008), exomePeak (Meng et al., 2013), or other peak calling methods are usually used to detect m^6A peaks with a length of about 100 nt (Chen et al., 2017). It is possible to further narrow down the precise location of m^6A sites by searching for the m^6A conforming DRACH motif in the detected peaks. However, since these methods cannot distinguish the random DRACH motifs from the real m^6A-containing motifs nearby, a large number of false-positive m^6A methylation sites is reported by MeT-DB (Liu et al., 2018) and RMBase (Xuan et al., 2018), as previously reported (Zhang et al., 2019). In addition to MeRIP-Seq, technologies with a single base resolution such as miCLIP (Bastian et al., 2015) and m^6A-CLIP (Shengdong et al., 2015) have been developed. However, due to the high difficulty and cost of base-resolution experiments, these technologies have not been widely used compared with MeRIP-Seq.

In silico methods to predict methylation sites based on machine learning (ML) approaches have been increasingly popular in recent years. For example, Chen et al. proposed the first ML method to predict RNA methylation sites in 2015, called "iRNA-Methyl" (Chen et al., 2015b). This method used dinucleotide composition and physicochemical characteristics to construct the PseDNC in order to represent RNA sequences and used these as an input to support vector machines (SVMs) to predict the m^6A methylation sites of *Saccharomyces cerevisiae*. Later, Zhou et al. (2016) used a variety of features to represent the sequence information, including the features of sequence coding, K-nearest base pair similarity and base pair frequency, to train the predictive model with the random forest (RF) method for the m^6A methylation sites prediction in mammalians. MethyRNA (Chen et al., 2016) encoded RNA sequences using the nucleotides' chemical properties and their accumulated frequency information, and used SVM classifier to predict the methylation modification sites of *S. cerevisiae*. M6AMRFS (Qiang et al., 2018) represented the sequence features with dinucleotide binary encoding (DBE) and local position-specific dinucleotide frequency (LPDF), and predicted the methylation modification sites of *S. cerevisiae* m^6A based on an eXtreme Gradient Boosting (XGBoost) classifier. Besides, a number of methods used deep learning (DL) approaches to predict m^6A methylation sites. BERMP (Yu Huang et al., 2018) used the base coding and the frequency of each base in a sliding window of a certain length as the characteristics of the sequence information. Using trained Gated Recurrent Unit (GRU) classifier and RF classifier, the final prediction results are obtained by logical regression. In DeepM6ASeq (Zhang and Hamada, 2018), the sequence was encoded using a one-hot encoding scheme, and the methylation modification sites were then predicted using a deep learning model consisting of a convolutional neural network (CNN) layer and one bidirectional long short-term memory (BLSTM) layer. Gene2vec (Quan Zou et al., 2018) took the methylation status near the methylation site, a one-hot encoding, the RNA word embedding feature, and the context word embedding feature as sequence features, used them respectively as an input to a CNN, and used a devoting method to predict the location.

Deep-m6A (Zhang Sy et al., 2019) took the product of a one-hot encoding of the sequence characteristics and the sites' reads count in the IP samples as an input to predict m^{6}A sites using a CNN. In addition, PRNAm-PC (Liu et al., 2016), RAM-ESVM (Wei et al., 2017a), AthMethPre (Xiang et al., 2016), and other methods (Chen et al., 2015c; Li et al., 2016; Zhao et al., 2018; Liu et al., 2020) can also be used to predict m^{6}A methylation sites. Although all these methods can predict RNA methylation sites, they are entirely based on the sequence context information. Even when secondary structures or other advanced features are used, the information is still directly extracted from the sequence without considering other potential and useful genomic features, referring to genome-related features that are not directly derived from sequences, including the secondary structure, gene annotation, transcription type, conservation, and many more. Recently, the method of WHISTLE (Zhang et al., 2019) combined sequence and genomic features to predict m^{6}A sites and constructed the entire m^{6}A epitranscriptome, showing that genomic features can also be very effective in the prediction of these sites and should be considered in the prediction framework.

Although the aforementioned methods can all perform general RNA methylation sites prediction, none of them was specifically considered or optimized for lncRNA methylation sites detection. Most of the currently existing experimental data use polyA selection when constructing the RNA-seq library; thus, lncRNAs will not be effectively captured since many of them are non-polyadenylated, and many lncRNA methylation sites are likely to be missed in the data generated from such protocol that would mainly contain the methylation sites information of mRNAs. As a result, the performance of site predictors trained with such data is likely to be limited when they are applied for the lncRNA methylation sites prediction task. The interplay between lncRNA and RNA methylation is now of an increasing interest to the science community and it is needed to develop a lncRNA-specific methylation sites prediction tool.

In this paper, we propose a new computational framework, **LITHOPHONE**, which stands for **l**ong noncod**i**ng RNA me**th**ylati**o**n sites **p**rediction from sequence c**h**aracteristics and gen**o**mic i**n**formation with an **e**nsemble predictor. LITHOPHONE uses a RF classifier to predict m^{6}A methylation sites by extracting the physicochemical and frequency accumulation characteristics of the bases based on sequence information and multiple genomic features, and identify lncRNA methylation sites by combining the information from mRNA and lncRNA sites using an ensemble predictor.

MATERIALS AND METHODS

Dataset Construction

For predicting the m^{6}A methylation sites in lncRNA, we employed the ground truth data that was used in the WHISTLE project (Zhang et al., 2019), including six single-base resolution m^{6}A experiments from six datasets obtained from five cell types (see **Table 1**): HEK293T, MOLM13, A549, CD8T, and HeLa, respectively, where HEK293T has two samples. The annotation information of lncRNA was obtained through Bioconductor via the TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts R package.

TABLE 1 | Single-base resolution m^{6}A datasets in lncRNA m^{6}A prediction.

Cell	Note	References
HEK293T	Abacm antibody	Bastian et al., 2015
HEK293T	Sysy antibody	Bastian et al., 2015
MOLM13		Vu et al., 2017
A549		Shengdong et al., 2015
CD8T		Shengdong et al., 2015
HeLa		Ke et al., 2017

The positive m^{6}A sites were defined as under the DRACH consensus motifs in at least two of the six datasets. The negative m^{6}A sites were randomly selected from the non-positive DRACH adenosines on the full transcripts containing the positive sites. There were equal numbers of negative and positive sites for each set of the training data, and the underlying motifs were restricted on DRACH. In addition, no sites were reported from the regions that can be mapped to multiple genes.

Finally, 2,582 full transcript m^{6}A sites in lncRNA were collected, including 1,291 positive sites and 1,291 negative ones, while 2,214 m^{6}A sites were obtained in mature lncRNA mode with 1,107 positive sites and 1,107 negative ones. Four-fifths of the sites were randomly selected for training, and the rest was retained for testing under both full transcript and mature RNA modes, respectively. For comparison purposes, we also generated the matched data for mRNAs, including 57,105 positive sites and the same number of negative ones for the full transcript mode, and 54,476 positive sites and 54,476 negative ones for the mature RNA mode, respectively. There were many more mRNA methylation sites compared with the lncRNA sites, suggesting that the mRNA methylation sites usually dominate the epitranscriptome profiling results.

Feature Representation

In this work, the sequence and genomic features were simultaneously used to represent a m^{6}A site.

Sequence Features

A nucleotide in a 21-nt sequence around the DRACH motif was represented by a four-dimensional vector following the method of MethyRNA (Chen et al., 2016). Firstly, each kind of nucleotide in RNA, including adenine (A), guanine (G), cytosine (C), and uracil (U), was represented by three characteristics according to its different chemical characteristics. For example, there is only one ring structure in cytosine and uracil, while adenine and guanine have two rings; adenine and cytosine both contain an amino group, while guanine and uracil both contain a keto group; hydrogen bonds are strong in guanine and cytosine when forming the secondary structure, while they are weak in adenine and uracil. According to these three features, a three-dimensional vector $S = (x_i, y_i, z_i)$ could be used to represent a nucleotide:

$$x = \begin{cases} 1 \;\; if \;\; s \in \{A, G\} \\ 0 \;\; if \;\; s \in \{C, U\} \end{cases}, \; y = \begin{cases} 1 \;\; if \;\; s \in \{A, C\} \\ 0 \;\; if \;\; s \in \{G, U\} \end{cases}, \; z = \begin{cases} 1 \;\; if \;\; s \in \{A, U\} \\ 0 \;\; if \;\; s \in \{C, G\} \end{cases} \quad (1)$$

Therefore, based on the above-defined rules, the vectors (1,1,1), (0,1,0), (1,0,0), and (0,0,1) can be used to encode A, C, G, and U, respectively. Next, the base accumulation frequency was also considered to describe the distribution of each base in the sequence. This frequency was defined as the frequency of the *i*th base in the previous *i* bases. The density f_i of the *i*th base is calculated by $f_i = d_i/i$, where f_i is the frequency of the occurrence of the *i*th base before *i* position density, and d_i is defined as the sum of the occurrences of the *i*th base in the previous *i* bases. For a sequence like "ACCUGAAUUG," A occurs three times at the 1st, 5th, and 6th positions, so the cumulative frequencies are 1/1, 2/5, and 3/6, respectively. However, the cumulative frequencies of *C* are 1/2 and 2/3; those of *U* are 1/4, 2/8, and 3/9; and those of *G* are 1/5 and 2/10. According to the above-described chemical characteristics and frequency cumulative distribution characteristics, each base can be encoded using a four-dimensional vector.

Genomic Features

Sequence features can only reflect the characteristics of each base in the sequence, but they cannot represent the topological information of the RNA methylation sites; thus, 60 additional genomic features were generated to reflect this information for the RNA methylation prediction in lncRNA. These features are detailed as follows: genomic features 1–10 are the dummy variable features, which indicate whether the site is overlapped with the topological region on the major RNA transcript. In order to extract genomic features, the longest transcripts were selected to prevent the influence of transcription isoforms. All features were extracted using the transcriptional annotations of the hg19 TxDb package (Xuan et al., 2018). Genomic features 11–12 stand for the distances toward the splicing junctions. Features 13–14 represent the length of the transcript region containing the methylation site. Features 15–32 indicate the consistence motif to which the RNA methylation site belongs. Features 33–36 represent clustering indicators or motif clustering, which reflect the clustering effect of the RNA methylation sites. Features 37–40 are the scores related to the evolutionary conservation, including two Phast-Cons scores and two fitness consequences scores. Features 41–42 obtain the secondary structure information of the RNA using RNAfold (Gruber et al., 2015). RNA annotations related to m^6A biology are features 43–55. Feature 56 is a dummy variable indicating whether the lncRNA is a miRNA target. Finally, features 57–60 include two *z*-scores of the isoform and exon number, and two *z*-scores of the GC content. **Table S1** contains the detailed information of the genomic features considered in the prediction.

Evaluation Metrics

In order to measure the prediction effect of the model, we used the measurements of sensitivity (Sn), specificity (Sp), accuracy (ACC), and Matthews correlation coefficient (MCC) to show the results of the model. The four indicators are respectively defined as follows:

$$S_n = \frac{TP}{TP + FN} \tag{2}$$

$$S_p = \frac{TN}{TN + FP} \tag{3}$$

$$ACC = \frac{TP + TN}{TN + FP + TP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{5}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative values, respectively. The sensitivity reflects the success rate of the positive sample prediction, and the specificity reflects the success rate of the negative sample prediction. A good prediction system should have both a high sensitivity and a high specificity at the same time. If the sensitivity is very high and the specificity is low, the false positive will be very high, while if the specificity is very high and the sensitivity is low, the false negative will be very high. Therefore, the forecasting system needs to comprehensively consider these two indicators. Matthews correlation coefficient is a comprehensive performance evaluation index considering unbalanced datasets. In addition, we plotted the receiver operating characteristic (ROC) curves and calculated the areas under the curves (as called "AUC") to evaluate the prediction performance.

RESULTS AND DISCUSSION

Comparing RF and Other Algorithm Performance Through Cross-Validation

In order to compare the prediction results of different algorithms, five different classifiers were used: RF (Liu, 2017; Wei et al., 2017b), SVM (Song et al., 2018), K-nearest neighbor (KNN) (Jia et al., 2016), logistic regression (LR) (Cha et al., 2015) and XGBoost (Chen and Guestrin, 2016). RF is a popular ML algorithm used to predict m^6A RNA methylation, which was applied in SRAMP (Zhou et al., 2016) to predict mammalian m^6A sites. SVM is another ML algorithm applied in computational

TABLE 2 | Performance under 10-fold cross-validation.

Mode	Method	Evaluation metrics				
		Sn	Sp	ACC	MCC	AUC
Full transcript	RF	0.923	0.938	0.930	0.861	0.971
	SVM	0.884	0.942	0.913	0.828	0.964
	KNN	0.5	0.501	0.500	0.001	0.945
	LR	0.881	0.944	0.912	0.827	0.962
	XGBoost	0.907	0.940	0.924	0.848	0.955
Mature lncRNA	RF	0.784	0.724	0.754	0.511	0.827
	SVM	0.738	0.713	0.725	0.451	0.796
	KNN	0.499	0.501	0.500	0.001	0.727
	LR	0.602	0.807	0.704	0.418	0.789
	XGBoost	0.645	0.697	0.671	0.345	0.722

biology, based on which the methods of MethyRNA (Chen et al., 2016) and RAM-ESVM (Wei et al., 2017a) were developed to predict RNA methylation sites. KNN is one of the most powerful methods in the data mining classification technology, and LR is an ML method with a simple algorithm and a high performance. XGBoost is frequently used in competitions and industry, and can be effectively applied to the tasks of classification, regression, and ranking; it was used in M6AMRFS (Qiang et al., 2018) to predict m^6A sites in multiple species based on the sequence features. All methods were implemented using the corresponding R packages (see **Table S2**). In order to compare their performance, a 10-fold cross-validation was employed on the training datasets under the full transcript and mature lncRNA modes. The performance of the different classifiers is summarized in **Table 2**, which shows that RF achieved the best performance both under the full transcript mode and mature lncRNA mode with an AUC of 0.971 and 0.827, respectively.

TABLE 3 | Performance under independent test.

Mode	Training data	Testing data	Method	Evaluation metrics				
				Sn	Sp	ACC	MCC	AUC
Full transcript	lncRNA	lncRNA	RF	0.922	0.930	0.926	0.853	0.966
			SVM	0.903	0.934	0.919	0.838	0.963
			KNN	0.500	0.500	0.500	0.000	0.942
			LR	0.895	0.926	0.911	0.822	0.959
			XGBoost	0.922	0.903	0.913	0.826	0.947
	lncRNA	mRNA	RF	0.981	0.046	0.514	0.077	0.759
			SVM	0.984	0.051	0.518	0.098	0.678
			KNN	0.499	0.501	0.500	0.000	0.572
			LR	0.954	0.171	0.562	0.200	0.716
			XGBoost	0.908	0.250	0.579	0.209	0.697
	mRNA	lncRNA	RF	0.752	0.934	0.843	0.698	0.936
			SVM	0.744	0.899	0.822	0.651	0.905
			KNN	0.492	0.508	0.500	0.000	0.703
			LR	0.539	0.953	0.746	0.541	0.872
			XGBoost	0.721	0.891	0.806	0.622	0.869
	mRNA	mRNA	RF	0.846	0.833	0.839	0.679	0.913
			SVM	0.829	0.839	0.834	0.669	0.908
			KNN	0.499	0.501	0.500	0.001	0.798
			LR	0.717	0.896	0.806	0.623	0.898
			XGBoost	0.831	0.832	0.832	0.664	0.907
Mature RNA	lncRNA	lncRNA	RF	0.766	0.694	0.730	0.461	0.821
			SVM	0.712	0.689	0.700	0.401	0.789
			KNN	0.500	0.500	0.500	0.000	0.734
			LR	0.590	0.802	0.696	0.401	0.797
			XGBoost	0.757	0.703	0.730	0.460	0.784
	lncRNA	mRNA	RF	0.757	0.522	0.639	0.287	0.705
			SVM	0.814	0.424	0.619	0.258	0.717
			KNN	0.493	0.508	0.501	0.002	0.520
			LR	0.804	0.472	0.638	0.292	0.660
			XGBoost	0.652	0.527	0.590	0.181	0.615
	mRNA	lncRNA	RF	0.788	0.608	0.698	0.403	0.807
			SVM	0.761	0.631	0.696	0.395	0.774
			KNN	0.500	0.500	0.500	0.000	0.542
			LR	0.419	0.838	0.628	0.283	0.653
			XGBoost	0.694	0.694	0.694	0.387	0.749
	mRNA	mRNA	RF	0.858	0.825	0.841	0.683	0.916
			SVM	0.840	0.842	0.841	0.682	0.915
			KNN	0.499	0.501	0.500	0.001	0.800
			LR	0.742	0.895	0.819	0.645	0.908
			XGBoost	0.831	0.832	0.832	0.664	0.907

Independent Tests Suggest That lncRNA and mRNA Methylation Sites Possess Different Characteristics

Next, we independently tested the m^{6}A sites on lncRNA in the full transcript and mature lncRNA modes. It is worth mentioning that none of the existing sites prediction methods differentiated between lncRNA and mRNA sites. Since mRNA sites are significantly over-represented in the data, it should dominate the performance assessment results. In the following tests, the mRNA and lncRNA sites were explicitly separated in both training and testing phases. Specifically, we used m^{6}A sites from both mRNA and lncRNA for the training, and then as testing sites from the two categories as well. We used the training data in lncRNA to train in the full transcript mode, tested with the testing data of lncRNA and mRNA separately, then trained with the training data in mRNA and finally tested with the testing data of lncRNA and mRNA separately. The same method was used in the mature lncRNA mode. As shown in **Table 3**, the best performance was achieved when the training and testing data were matched, suggesting that lncRNA and mRNA methylation sites exhibited different characteristics. When using lncRNA data as training samples to predict m^{6}A sites in lncRNA, the prediction performance (AUC = 0.966 and AUC = 0.821, under full transcript and mature RNA modes, respectively) was better than when we used mRNA data as training samples to predict the sites of lncRNA (AUC = 0.936 and AUC = 0.807, under full transcript and mature RNA modes, respectively). Similarly, this situation also occurs in predicting the sites of mRNA. When mRNA sites were used for training, the results achieved for testing the sites of mRNA were better than those of lncRNA. In addition, it can be seen that the method of RF can achieve the best prediction results in both cross-validation and independent testing among the five different prediction methods. Therefore, RF is chosen as a classifier to predict the methylation sites in lncRNA.

Construction of an Ensemble Predictor

Since mRNA methylation sites can also be used for lncRNA site prediction and have achieved a reasonably good performance (**Table 3**), and considering that we only have a limited number of lncRNA methylation sites, which may not be sufficient for training, an ensemble model using mixed predictive results of mRNA and lncRNA was proposed in order to further improve the lncRNA sites prediction accuracy. The probability of lncRNA sites prediction in this model is defined as follows:

$$P_{en} = \alpha P_{\mathrm{m}} + (1-\alpha)P_{\mathrm{lnc}} \tag{6}$$

where P_{en} denotes the final prediction probability of the sites in the mature lncRNA mode, P_m represents the prediction probability of the sites when mRNA sites data were used for training, and P_{lnc} denotes the prediction probability of the sites when the lncRNA data were used for training. In order to optimize the value of α, which gives the models different weights, a grid search was performed $\alpha \in [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. The best performance was achieved when $\alpha = 0.3$ (AUC = 0.835) (see **Figure 1**), which indicates that the relatively small number (1,107) of lncRNA sites plays a major role in the ensemble predictor (weight = 0.7), while the very large number (54,476) of mRNA methylation sites plays a minor role (weight = 0.3). The results comparing the mRNA and lncRNA models are shown in **Table 4**.

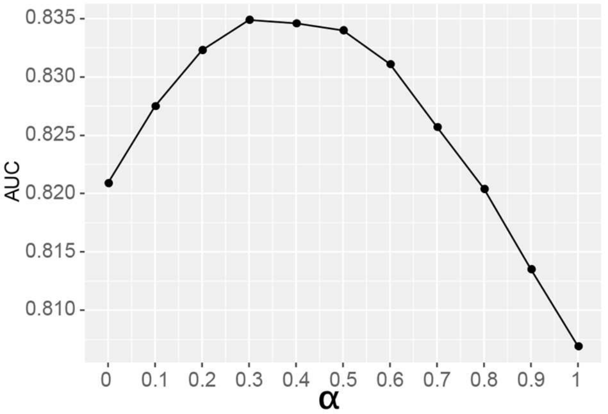


FIGURE 1 | Search for optimal parameter of the ensemble predictor. The optimal result was achieved when α=0.3. When $\alpha = 0$, only lncRNA sites were used for training; while when $\alpha = 1$, only mRNA sites were considered.

TABLE 4 | Comparison of ensemble model and lncRNA trained model.

Predictor	Evaluation metrics				
	Sn	Sp	ACC	MCC	AUC
mRNA trained	0.788	0.608	0.698	0.403	0.807
lncRNA trained	0.766	0.694	0.730	0.461	0.821
Ensemble (α = 0.3)	0.797	0.689	0.743	0.489	0.835

Feature Selection

To further optimize the prediction results, we used feature selection to obtain the most effective feature set to predict the methylation sites on lncRNA, and a greedy search was implemented. Firstly, we ranked the features according to their importance through the results of AUC with 10-fold cross validation. Then, one feature was added to the training set each time from the sorted feature set, and the prediction results were obtained using 10-fold cross-validation. The optimal feature set was obtained through the highest AUC. As shown in **Figures 2C,D**, the first 134 features composed the optimal feature set in the m^{6}A sites prediction in the full transcript mode, while the top 41 features can get the highest AUC when predicting m^{6}A sites in the mature RNA mode. In addition, it can be seen from **Figures 2A,B** that the top five features when predicting lncRNA m^{6}A sites under the full transcript mode are whether the site is overlapped with the intron (intron), the distance to the downstream (3′ end) splicing junction (dist_sj_3_p2000), the z-score of the isoform num (isoform_num), whether the site is

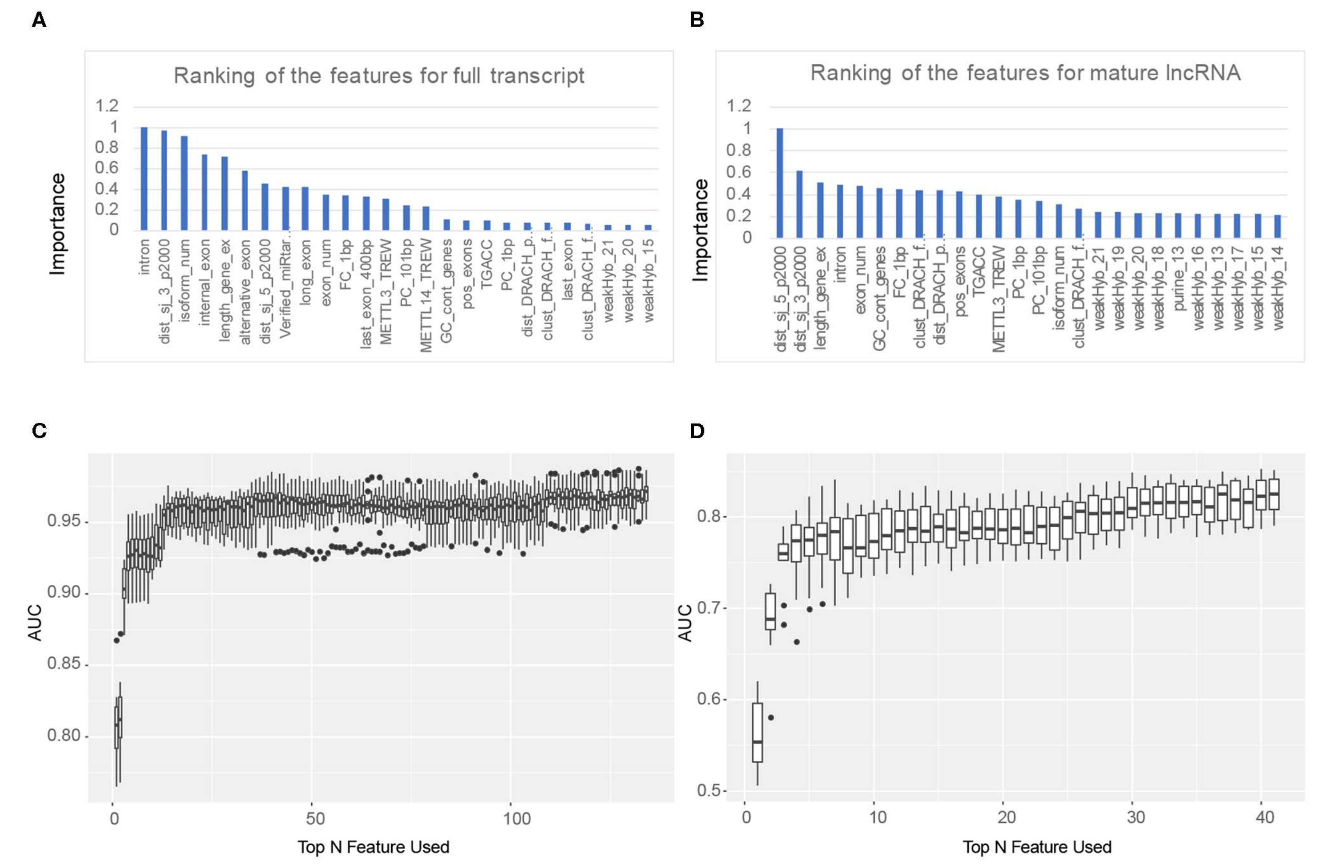


FIGURE 2 | Feature selection results. **(A)** The ranking of the features for full transcript m^{6}A site prediction. **(B)** The ranking of the features for mature lncRNA m^{6}A site prediction. **(C)** Top 134 features were selected for full transcript m^{6}A site prediction. **(D)** Top 41 features were selected for mature lncRNA m^{6}A site prediction.

overlapped with the internal exon (internal_exon), and the *z*-score of the gene length exons (length_gene_ex). On the other hand, the five most importance features in the prediction sites under the mature RNA mode are the distance to the upstream (5′ end) splicing junction (dist_sj_5_p2000), the distance to the downstream (3′ end) splicing junction (dist_sj_3_p2000), the *z*-score of the gene length exons (length_gene_ex), whether the site is overlapped with the intron (intron), and the *z*-score of the exon num (exon_num). Although some of the first five features are identical in predicting RNA methylation sites in both full transcript and mature lncRNA modes, different characteristics reflect the inherent differences between the two modes.

Comparison With Existing Methods

In order to further verify the validity of the proposed algorithm, we compared it with the methods of SRAMP that uses RF to predict mRNA m^{6}A sites, MethyRNA that uses the same sequence features as we do, but uses SVM for prediction, and the deep learning method of Gene2vec. These methods have available prediction tools. The results are summarized in **Table 5** and the ROC curves of the four methods are shown in **Figure 3**. The results show that the proposed method is superior to the current popular methods in predicting lncRNA methylation sites.

TABLE 5 | Performance comparison for lncRNA m^{6}A site prediction.

Mode	Method	Evaluation metrics				
		Sn	Sp	ACC	MCC	AUC
Full transcript	SRAMP	0.705	0.791	0.748	0.498	0.827
	MethyRNA	0.717	0.752	0.734	0.469	0.801
	Gene2vec	0.798	0.813	0.805	0.611	0.865
	LITHOPHONE	0.922	0.930	0.926	0.853	0.966
Mature RNA	SRAMP	0.604	0.748	0.676	0.355	0.749
	MethyRNA	0.622	0.644	0.633	0.266	0.679
	Gene2vec	0.778	0.689	0.734	0.469	0.806
	LITHOPHONE	0.797	0.689	0.743	0.489	0.835

LncRNAome-Wide m^{6}A Site Prediction

In order to obtain a complete map of all the human lncRNA methylation sites, we searched the entire lncRNAome for all the DRACH motifs, which represent candidate lncRNA methylation sites, under both full transcript and mature RNA modes, and

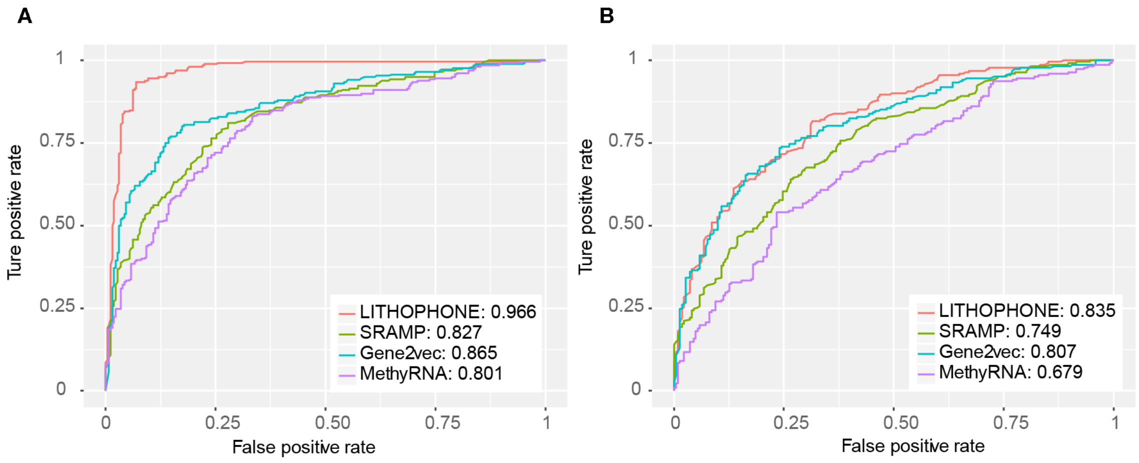


FIGURE 3 | ROC for lncRNA methylation site prediction. The proposed approach substantially outperformed competing approaches. **(A)** The ROC curve for the full transcript mode. **(B)** The ROC curve for the mature RNA mode.

used the proposed method to predict the probability of lncRNA methylation sites. Finally, 330,564 out of the total 4,046,330 DRACH motifs were predicted to contain m^{6}A RNA methylation sites under the full transcript mode with a probability greater than 0.5, and 114,093 out of the total 313,458 DRACH motifs from 29,687 lncRNAs were predicted as putative lncRNA methylation sites under the mature RNA mode. The prediction results can be freely accessed at: http://180.208.58.19/lith/. In addition, the data and code used in this article can be obtained from https://github.com/lianliu09/lncRNA-m6a.git.

CONCLUSION

With the rapid development of high-throughput sequencing and RNA methylation profiling technologies, people can now study RNA modifications with a high accuracy in the full transcriptome range. In recent years, a number of RNA methylation sites prediction methods have been developed. However, to the best of our knowledge, none of them considered the experimental bias induced in the current epitranscriptome data, which can significantly affect the performance of these predictors.

In this paper, we presented LITHOPHONE, an ensemble framework to predict m^{6}A epitranscriptome in lncRNA. Unlike other methods that rely only on sequence information, LITHOPHONE extracts the physicochemical and frequency accumulation characteristics of the bases, combining 60 genomic characteristics to predict the m^{6}A methylation modification sites under both full transcript and mature RNA modes on lncRNA using the RF algorithm. To the best of our knowledge, LITHOPHONE is the first m^{6}A sites predictor that is optimized for lncRNA. We showed that lncRNA and mRNA exhibit different predictive characteristics, and how LITHOPHONE outperforms competing approaches in lncRNA methylation site prediction. Additionally, we searched the entire lncRNAome in human for all possible m^{6}A sites located on lncRNAs and predicted 330,564 m^{6}A sites on pre-lncRNA and 114,093 sites on mature lncRNA. We built a website to query the prediction results of lncRNA methylation sites and it is freely accessible at: http://180.208.58.19/lith/. The LITHOPHONE framework can be easily extended to other RNA modifications, such as m^{1}A, as well as other species, such as the mouse.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/lianliu09/lncRNA-m6a.git.

AUTHOR CONTRIBUTIONS

ZW and LL initialized the project. LL, XL, ZW, and JM designed the research plan. ZW constructed the genomic features considered in site prediction. LL performed the site prediction and drafted the manuscript. ZF and YT built the website. All authors read and critically revised and approved the final manuscript.

FUNDING

This work has been supported by National Natural Science Foundation of China (61902230, 61972451, and 31671373); China Postdoctoral Science Foundation (2018M640949); Fundamental Research Funds for the Central Universities (GK201903083 and GK201901010); XJTLU Key Program Special Fund (KSF-T-01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00545/full#supplementary-material

REFERENCES

Alarcón, C. R., Hyeseung, L., Hani, G., Nils, H., and Tavazoie, S. F. (2015a). N6-methyladenosine marks primary microRNAs for processing. *Nature* 519, 482–485. doi: 10.1038/nature14281

Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N., and Tavazoie, S. F. (2015b). N6-methyladenosine marks primary microRNAs for processing. *Nature* 519, 482–485. doi: 10.1038/nature14281

Bastian, L., Grozhik, A. V., Olarerin-George, A. O., Cem, M., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* 12:767. doi: 10.1038/nmeth.3453

Cha, S., Yu, H., Park, A. Y., Oh, S. A., and Kim, J. Y. (2015). The obesity-risk variant of FTO is inversely related with the So-Eum constitutional type: genome-wide association and replication analyses. *Bmc Complement. Alternative Med.* 15:120. doi: 10.1186/s12906-015-0609-4

Chen, T., and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco).

Chen, T., Hao, Y. J., Zhang, Y., Li, M. M., Wang, M., Han, W., et al. (2015a). m 6 A RNA methylation is regulated by MicroRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* 16:289. doi: 10.1016/j.stem.2015.01.016

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K. C. (2015b). iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem* 490:26. doi: 10.1016/j.ab.2015.08.021

Chen, W., Hong, T., Liang, Z., Lin, H., and Zhang, L. (2015c). Identification and analysis of the N6-methyladenosine in the Saccharomyces cerevisiae transcriptome. *Sci. Reports* 5:13895. doi: 10.1038/srep13859

Chen, W., Tang, H., and Lin, H. (2016). MethyRNA: a web-server for identification of N(6)-methyladenosine sites. *J. Biomol. Struct. Dyn* 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, X., Sun, Y. Z., Liu, H., Zhang, L., Li, J. Q., and Meng, J. (2017). RNA methylation and diseases: experimental results, databases, Web servers and computational models. *Brief Bioinform*. 20, 896–917. doi: 10.1093/bib/bbx142 bbx142

Dominissini, D., Moshitchmoshkovitz, S., Schwartz, S., Salmondivon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485:201. doi: 10.1038/nature11112

Fu, Y., Dan, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.* 15, 293–306. doi: 10.1038/nrg3724

Fustin, J. M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., et al. (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 155, 793–806. doi: 10.1016/j.cell.2013.10.026

Geula, S., Moshitchmoshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmondivon, M., et al. (2015). Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science* 347:1002. doi: 10.1126/science.1261417

Gruber, A. R., and Bernhart, S. H., Lorenz, R. (2015). RNA bioinformatics. *Springer* 307—326. doi: 10.1007/978-1-4939-2291-8_19

Jia, C. Z., Zhang, J. J., and Gu, W. Z. (2016). RNA-MethylPred: a high accuracy predictor to identify N6-methyladenosine in RNA. *Anal. Biochem.* 510, 72–75. doi: 10.1016/j.ab.2016.06.012

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., et al. (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. doi: 10.1101/gad.301036.117

Li, G. Q., Liu, Z., Shen, H. B., and Yu, D. J. (2016). TargetM6A: identifying N6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans. Nanobiosci.* 15, 674–682. doi: 10.1109/TNB.2016.2599115

Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform*. 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, H., Wang, H., Wei, Z., Zhang, S., Hua, G., Zhang, S. W., et al. (2018). MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res.* 46, D281–D287. doi: 10.1093/nar/gkx1080

Liu, J., and Jia, G. (2014). Methylation modifications in eukaryotic messenger RNA. *J. Genet. Genom.* 41, 21–33. doi: 10.1016/j.jgg.2013.10.002

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* 10:93. doi: 10.1038/nchembio.1432

Liu, L., Lie, X., Meng, J., and Wei, Z. (2020). WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features. *Curr. Genomics*. 21, 67–76. doi: 10.2174/1389202921666200211104140

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518:560. doi: 10.1038/nature14234

Liu, Z., Xiao, X., Yu, D. J., Jia, J., Qiu, W. R., and Chou, K. C. (2016). pRNAm-PC: predicting N 6 -methyladenosine sites in RNA sequences via physical–chemical properties. *Anal. Biochem.* 497:60. doi: 10.1016/j.ab.2015.12.017

Meng, J., Cui, X., Rao, M. K., Chen, Y., and Huang, Y. (2013). Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics* 29, 1565–1567. doi: 10.1093/bioinformatics/btt171

Meyer, K. D., and Jaffrey, S. R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* 15:313. doi: 10.1038/nrm3785

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Nian, L., Qing, D., Guanqun, Z., Chuan, H., Marc, P., and Tao, P. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518, 560–564. doi: 10.1038/nature14234

Patil, D. P., Chen, C. K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M., et al. (2016). m(6)A RNA methylation promotes XIST-mediated transcriptional repression. *Nature* 537:369. doi: 10.1038/nature19342

Peng, L., Yuan, X., Jiang, B., Tang, Z., and Li, G. C. (2016). LncRNAs: key players and novel insights into cervical cancer. *Tumor Biol.* 37, 2779–2788. doi: 10.1007/s13277-015-4663-9

Qiang, X., Chen, H., Ye, X., Su, R., and Wei, L. (2018). M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front. Genet.* 9:495. doi: 10.3389/fgene.2018.00495

Quan Zou, P. X., Leyi, W., and Bin, L. (2018). Gene2vec: gene subsequence embedding for prediction of mammalian N6-Methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Qureshi, I. A., Mattick, J. S., and Mehler, M. F. (2010). Long non-coding RNAs in nervous system function and disease. *Brain Res.* 1338, 20–35. doi: 10.1016/j.brainres.2010.03.110

Roost, C., Lynch, S. R., Batista, P. J., Qu, K., Chang, H. Y., and Kool, E. T. (2015). Structure and thermodynamics of N6-Methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc* 137:2107. doi: 10.1021/ja513080v

Shafik, A., Schumann, U., Evers, M., Sibbritt, T., and Preiss, T. (2016). The emerging epitranscriptomics of long noncoding RNAs. *Biochim. Biophys. Acta* 1859:S187493991500231X. doi: 10.1016/j.bbagrm.2015.10.019

Shengdong, K., Alemu, E. A., Claudia, M., Emily Conn, G., Fak, J. J., Aldo, M., et al. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev.* 29, 2037–2053. doi: 10.1101/gad.269415.115

Song, B., Tang, Y., Wei, Z., Liu, G., Su, J., Meng, J., et al. (2020). PIANO: a web server for pseudouridine site (Ψ) identification and functional annotation. *Front. Genet.* 11:88. doi: 10.3389/fgene.2020.00088

Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N. D., Webb, G. I., et al. (2018). iProt-Sub: a comprehensive tool for accurately mapping and predicting protease-specific substrates and cleavage sites. *Phys. Rev. E* 97:28. doi: 10.1093/bib/bby028

Vu, L. P., Pickering, B. F., Cheng, Y., Zaccara, S., Nguyen, D., Minuesa, G., et al. (2017). The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat. Med.* 23, 1369–1376. doi: 10.1038/nm.4416

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wei, C., Xing, P., and Quan, Z. (2017a). Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep* 7:40242. doi: 10.1038/srep40242

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2017b). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Xiang, S., Yan, Z., Liu, K., Zhang, Y., and Sun, Z. (2016). AthMethPre: a web server for the prediction and query of mRNA m(6)A sites in Arabidopsis thaliana. *Mol. Biosyst* 11:e0162707. doi: 10.1039/C6MB00536E

Xuan, J. J., Sun, W. J., Lin, P. H., Zhou, K. R., Liu, S., Zheng, L. L., et al. (2018). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* 46:D327. doi: 10.1093/nar/gkx934

Yang, D., Qiao, J., Wang, G., Lan, Y., Li, G., Guo, X., et al. (2018). N6-Methyladenosine modification of lincRNA 1281 is critically required for mESC differentiation potential. *Nucleic Acids Res.* 46:130. doi: 10.1093/nar/gky130

Yu Huang, N. H., Yu, C., Zhen, C., and Lei, L. (2018). BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci* 14, 1669–1677. doi: 10.7150/ijbs.27819

Zhang Sy, Z. S., Fan, X.n, Meng, J., Chen, Y., Gao, S.j, and Huang, Y. (2019). Global analysis of N6-methyladenosine functions and its disease association using deep learning and network-based methods. *PLoS Comput. Biol.* 15:e1006663. doi: 10.1371/journal.pcbi.1006663

Zhang, Q., Chen, K., Wu, X., Wei, Z., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47:e41. doi: 10.1093/nar/gkz074

Zhang, S., Zhao, B. S., Zhou, A., Lin, K., Zheng, S., Lu, Z., et al. (2017). m 6 A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program. *Cancer Cell* 31:591. doi: 10.1016/j.ccell.2017.02.013

Zhang, Y., and Hamada, M. (2018). DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinform.* 19(Suppl.19):524. doi: 10.1186/s12859-018-2516-4

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhao, Z., Hui, P., Lan, C., Yi, Z., Liang, F., and Li, J. (2018). Imbalance learning for the prediction of N6-Methylation sites in mRNAs. *BMC Genomics* 19:574. doi: 10.1186/s12864-018-4928-y

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 August 2020
doi: 10.3389/fgene.2020.00863

N6-Methyladenosine RNA Methylation Regulators Have Clinical Prognostic Values in Hepatocellular Carcinoma

Wei Liu[1], Cuiqing Zhong[2], Deguan Lv[3], Mengjie Tang[4] and Feng Xie[5]**

[1] Department of Pharmacy, The Third Xiangya Hospital, Central South University, Changsha, China, [2] Center for Reproductive Medicine, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [3] Division of Regenerative Medicine, Department of Medicine, University of California, San Diego, San Diego, CA, United States, [4] Hunan Cancer Hospital, The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, China, [5] Department of Pharmacy, The Nanshan District Maternity and Child Healthcare Hospital of Shenzhen, Shenzhen, China

OPEN ACCESS

Edited by:
Mattia Pelizzola,
Italian Institute of Technology (IIT), Italy

Reviewed by:
Deli Liu,
Weill Cornell Medicine, Cornell University, United States
Nan Li,
China-Japan Friendship Hospital, China
Zhuoxian Rong,
Xiangya Hospital, Central South University, China

***Correspondence:**
Mengjie Tang
tangmengjie@hnca.org.cn
Feng Xie
xiefeng101787@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 December 2019*
Accepted: *15 July 2020*
Published: *12 August 2020*

Citation:
Liu W, Zhong C, Lv D, Tang M and Xie F (2020) N6-Methyladenosine RNA Methylation Regulators Have Clinical Prognostic Values in Hepatocellular Carcinoma. Front. Genet. 11:863. doi: 10.3389/fgene.2020.00863

Although it is widely accepted that N6-methyladenosine (m^6A) RNA methylation plays critical roles in tumorigenesis and progression, the values of m^6A modification are less known in hepatocellular carcinoma. The major purpose of our current studies is to investigate the role of m^6A regulators in hepatocellular carcinoma and whether it can affect the prognosis of hepatocellular carcinoma. Here we demonstrate that most of the m^6A regulators are highly expressed in hepatocellular carcinoma. Furthermore, we cluster hepatocellular carcinoma into two subgroups (cluster 1/2) by applying consensus clustering to m^6A regulators. Compared with the cluster 1 subgroup, the cluster 2 subgroup was significantly associated with a higher pathological grade and survival. Based on these findings, we reveal a risk signature by using three m^6A regulators, which are not only an independent prognostic marker but also a predictor of the clinicopathological features in hepatocellular carcinoma. In conclusion, m^6A regulators are crucial participants in the malignant progression of hepatocellular carcinoma and are potential targets for prognosis.

Keywords: m^6A modification, m^6A regulators, hepatocellular carcinoma, a risk signature, prognostic marker

INTRODUCTION

RNA modification was first discovered in the 1960s and was considered to be another epigenetic form analogous to DNA and histone modification (Jia et al., 2013). Among more than 100 kinds of RNA modifications known so far, N6-methyladenosine (m^6A) methylation is the most abundant RNA epigenetic modification in RNA, which is dynamically regulated by methyltransferases ("writers"), binding proteins ("readers"), and demethylases ("erasers") (Niu et al., 2013; Yang et al., 2018). The prominent methyltransferases complex catalyzes the formation of m^6A, which contain at least six "writer" proteins: methyltransferase like 3 (METTL3), methyltransferase like 14 (METTL14), WT1-associated protein (WTAP), VIRMA (KIAA1429), zinc finger CCCH domain-containing protein 13 (ZC3H13), and RNA binding motif protein 15 (RBM15) (Liu and Pan, 2016). The demethylases catalyze the demethylation of m^6A, which mainly include fat mass- and obesity-associated protein (FTO) and α-ketoglutarate-dependent dioxygenase alkB homolog 5 (ALKBH5)

(Ding et al., 2018; Piette and Moore, 2018). The binding proteins, which recognize and bind with m^6A, are consisting of YTH domain family proteins and heterogeneous nuclear ribonucleoprotein C (HNRNPC) (Duan et al., 2019). The biological functions of m^6A RNA methylation are involved in regulating all stages of the RNA life cycle, including pre-mRNA splicing, pri-miRNA processing, nuclear output, RNA translation regulation, and RNA degradation (Roignant and Soller, 2017).

The transcriptome-wide mapping of m^6A focuses on investigating the landscapes and the functions of the reversible m^6A modification in the last decade (Bi et al., 2019). Recently, more and more scientists focus on exploring the association between m^6A and human diseases, especially in tumors (Dai et al., 2018; Pan et al., 2018). A growing appreciation of the biological significance of m^6A RNA methylation implied that m^6A contributed to tumorigenesis and tumor progression (Deng et al., 2018). The dislocation of m^6A is closely associated with various kinds of cancers, such as glioblastoma (GBM), colorectal carcinoma (CRC), pancreatic cancer (PC), and hepatocellular carcinoma (HCC) (Chen et al., 2018; Chai et al., 2019; Zhang et al., 2019; Zhou et al., 2019). Notably, the roles of m^6A regulators in tumors are controversial. METTL3 serves as a tumor suppressor gene in GBM and is considered as an oncogene in CRC or non-small cell lung carcinoma (Li et al., 2019; Wei et al., 2019; Liu et al., 2020). YTHDF2 acts as a tumor suppressor gene in lung cancer and supposed to be an oncogene in PC (Chen et al., 2017; Sheng et al., 2019). The controversial roles of m^6A regulators in tumors suggest that the functions of m^6A modification in tumors are complicated. Moreover, the literature does not have comprehensive m^6A regulator expression and prognosis analysis in tumors.

In this study, we systematically analyze the expression data of 13 m^6A modification regulators in HCC from The Cancer Genome Atlas (TCGA) datasets. We demonstrate that most of the 13 m^6A regulators are highly expressed among HCC. Moreover, we also find that the m^6A regulators are crucial participants in the malignant progression of HCC and a signature with three selected m^6A regulators is designed to stratify the prognosis of HCC.

MATERIALS AND METHODS

Data Acquisition and Processing

The RNA-seq transcriptome and clinical data of 407 HCC samples and 58 adjacent tissue samples were obtained from TCGA[1]. The workflow type is fragments per kilobase million. The R package "limma" was used to process and delete duplicate genes. The expression of m^6A regulators in HCC was extracted from RNA-seq transcriptome. The Wilcoxon test was used to analyze the differential expression of these m^6A regulators ($p < 0.05$ was considered as significant). Incomplete samples of survival data were removed, and finally, 403 samples with complete clinical information were obtained for subsequent analysis. The flow chart of this study is shown in **Figure 1**.

Identify the Role of m6A Regulators in HCC

Gene mutation and copy number variation data were downloaded from the cbioport database[2]. The interaction and the correlation among m^6A regulators were analyzed using the R package "corrplot." The HCC patients were divided into two subgroups based on the expression of m^6A regulators using a cluster analysis method with "ConsensusClusterPlus"[3]. The R package "ggplot2" is used for principal component analysis (PCA). The R package "survival" was used to plot Kaplan–Meier survival curves. A $p < 0.05$ was considered as statistically significant.

Construction of a Signature Associated With Prognosis

The roles of m^6A regulators in the prognosis of HCC patients were identified by univariate Cox regression analysis; $p < 0.05$ was considered as significant. A risk signature was built by the least absolute shrinkage and selection operator (LASSO) Cox regression algorithm, and multivariate Cox regression analysis. The signature is expressed as follows: risk score = (coefficient gene 1 × gene 1 expression) + (coefficient gene 2 × expression of gene 2) + ... + (coefficient gene n × expression gene n). The median risk score served as a cutoff value to classify patients into high-risk and low-risk groups. The R package "survival ROC" was used to perform time-dependent receiver operating characteristic (ROC) curve analysis to assess the accuracy of the predicted genetic features of time-dependent cancer death. The area under the curve (AUC) was calculated to evaluate the accuracy of the risk prediction model. The R package "survival" was used to plot Kaplan–Meier survival curves.

Independence of Prognostic Factors From Other Clinical Parameters in TCGA

Complete information on the 403 samples included relevant clinical data for univariate and multivariate Cox regression analyses. $p < 0.05$ was considered as statistically significant.

Construction of a Predictive Nomogram

The independent prognostic factors were chosen as the prognostic model to construct a nomogram in the entire TCGA cohort. The calibration plot and the concordance index (C-index) were used to investigate the calibration and the discrimination of the nomogram.

RESULTS

m^6A Regulators in HCC Patients Are Highly Expressed

More and more reports have shown that m^6A regulators such as METTL14 (Li et al., 2020), YTHDF1 (Zhao et al., 2018), YTHDF2 (Chen et al., 2018), and WTAP (Chen et al., 2019)

[1] https://portal.gdc.cancer.gov/

[2] http://www.cbioportal.org/

[3] http://www.bioconductor.org/

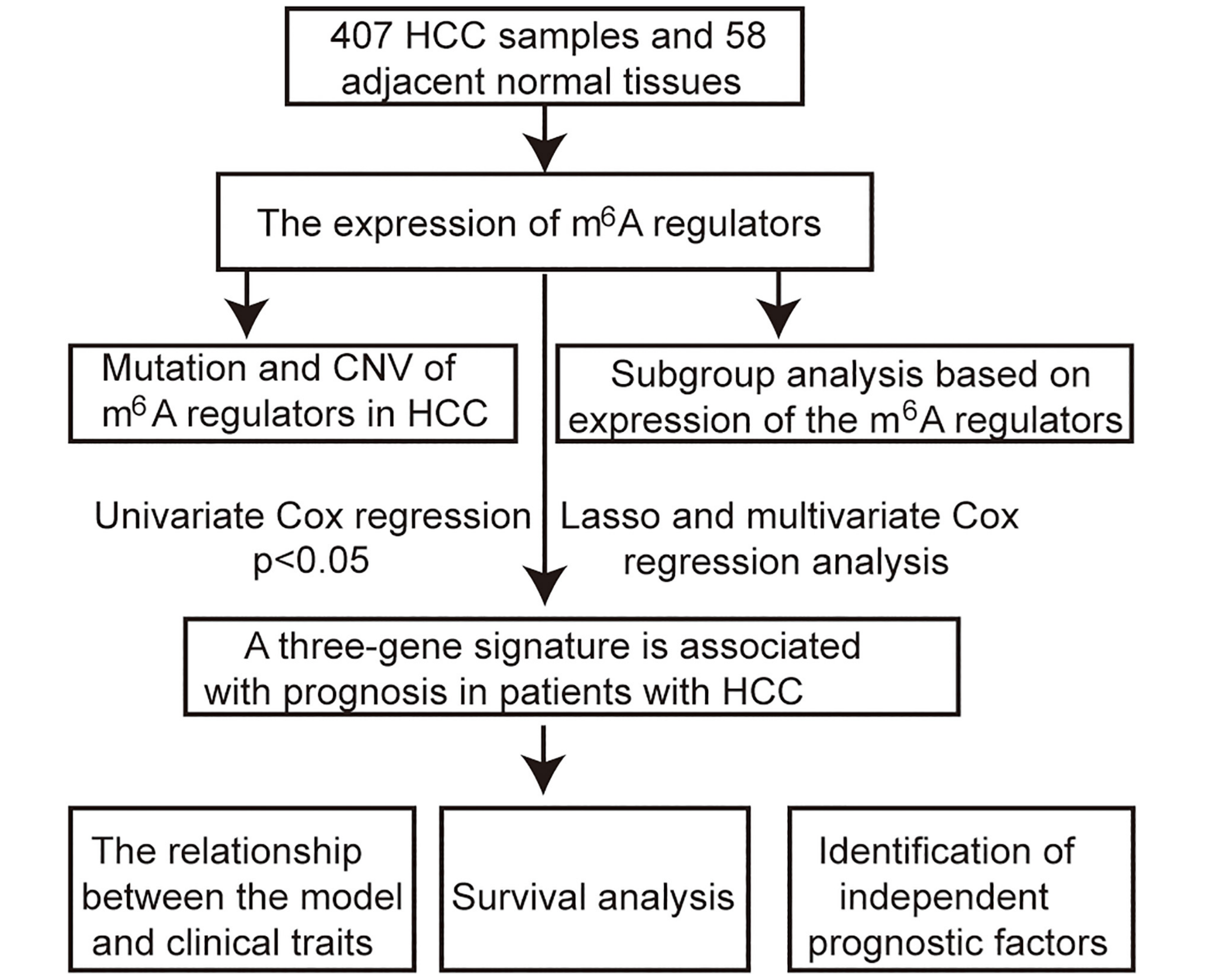

407 HCC samples and 58 adjacent normal tissues
The expression of m6A regulators
Mutation and CNV of m6 A regulators in HCC
Subgroup analysis based on expression of the m6A regulators
Univariate Cox regression p<0.05
Lasso and multivariate Cox regression analysis
A three-gene signature is associated with prognosis in patients with HCC
The relationship between the model and clinical traits
Survival analysis
Identification of independent prognostic factors

FIGURE 1 | Flow chart of the approach utilized in the current study.

are essential for the deterioration and the progression of HCC. To further confirm the role of all m^{6}A regulators in HCC, we systematically investigated the expression of 13 m^{6}A regulators (including six writers: KIAA1429, METTL3, METTL14, RBM15, WTAP, and ZC3H13; two erasers: ALKBH5 and FTO; and five readers: HNRNPC, YTHDC1, YTHDC2, YTHDF1, and YTHDF2) in 403 HCC samples and 58 adjacent normal tissue samples from the TCGA database. Information on these m^{6}A regulators is shown in **Table 1**. Similar to the results of Li's report (Li et al., 2020), we found that KIAA1429, METTL3, and HNRNPC are highly expressed in HCC tumor samples. Contrary to Li's findings, our results show that the expression of METTL14, YTHDC1, YTHDC2, and FTO was also increased in HCC, while the expression of ZC3H13 has no difference between the tumor samples and the adjacent normal tissue samples. In detail, HNRNPC had the highest expression, followed by ALKBH5 and YTHDF1($p < 0.05$) (**Figures 2A,B**). The inconsistent results between Li's study and our research may be caused by different sample data.

Mutation and Copy Number Variation of m^{6}A Regulatory Genes in HCC

We then completely analyzed the different mutation and copy number variation (CNV) patterns of m6A regulatory genes in HCC from the cbioport database[4]. It included gene mutation, amplification, deep deletion, mRNA expression change, and other multiple alterations. The result revealed that m^{6}A regulators were highly expressed in most HCC samples; meanwhile, m^{6}A regulators had gene mutations and CNV (**Figure 3A**). Specifically, the m^{6}A "writer" gene VIRMA (KIAA1429) had the highest mutation and CNV frequency

[4] http://www.cbioportal.org/

TABLE 1 | Information on 13 m^{6}A regulators.

Types	Gene symbol	HGNC symbol	Full name
Readers	*HNRNPC*	5035	Heterogeneous nuclear ribonucleoprotein C
	YTHDC1	30626	YTH domain containing 1
	YTHDC2	24721	YTH domain containing 2
	YTHDF1	15867	YTH N6-methyladenosine RNA binding protein 1
	YTHDF2	31675	YTH N6-methyladenosine RNA binding protein 2
Writers	*KIAA1429*	24500	vir like m6A methyltransferase associated
	METTL3	17563	Methyltransferase like 3
	METTL14	29330	Methyltransferase like 14
	RBM15	14959	RNA binding motif protein 15
	WTAP	16846	WT1 associated protein
	ZC3H13	20368	Zinc finger CCCH-type containing 13
Erasers	*ALKBH5*	25996	alkB homolog 5, RNA demethylase
	FTO	24678	FTO alpha-ketoglutarate dependent dioxygenase

(40%). as well as "readers" YTHDF1 (18%), ALKBH5 (17%), and WTAP (17%), respectively (**Figure 3B**).

Interaction and Correlation Among m^{6}A Regulators in HCC

Next, we evaluated the interaction and the correlation among m^{6}A regulators. In the cbioport database, we found that there were close interactions among m^{6}A regulators (**Figure 3C**). Furthermore, we analyzed the expression correlation of these genes in detail based on the expression profile of m^{6}A regulators. The result showed that there was a significant positive correlation between the expressions of most m^{6}A regulators. However, there might be no correlation between YTHDC2 and ALKBH5, ZC3H13 and ALKBH5, ZC3H13 and KIAA1429, and ZC3H13 and YTHDF1 (**Figure 3D**). These results reveal that, except for a few m^{6}A regulators, most of them may play roles together in HCC.

Classification of HCC Samples Based on the Expression of m^{6}A Regulators

To study whether m^{6}A regulators type HCC samples well, by inputting the expression profile of the m^{6}A regulators, we performed a cluster analysis with the R package "ConsensusClusterPlus" (k = 2–9, **Figure 4A**). The results revealed that it was most appropriate to divide the patients into two subtypes (**Figure 4A**). These two subtypes were defined as cluster 1 and cluster 2 in order to further verify the accuracy of the two subtypes. We input all gene expression profiles and subtype information and used the R packages "limma" and "ggplot2" for the PCA of HCC. The PCA results also showed that the HCC sample could be well divided into two subtypes (**Figure 4B**). Moreover, a significantly shorter survival curve in the cluster 2 subgroup was observed (**Figure 4C**). Furthermore, the clinical characteristics of the two subtypes are shown in **Table 2**. These two subgroups were significantly correlated with the WHO grade, gender, age, and lymph node metastasis ($p < 0.05$) (**Figure 4D**). These findings further indicate that m^{6}A regulators have a key role in HCC categories. However, the specific molecular differences or other effects between these two subtypes needed further research.

A Risk Signature Built Using Three Selected m^{6}A Regulators

The previous results revealed that m^{6}A regulators play an important role in HCC. In order to explore whether m^{6}A regulators predict the survival prognosis of HCC patients, we combined the expression profile and the clinical data of m^{6}A regulators for univariate Cox regression analysis. The results revealed that a total of seven genes (YTHDF2, KIAA1429, HNRNPC, WTAP, YTHDF1, YTHDC1, and METTL3) were significantly associated with survival prognosis ($p < 0.05$, **Figure 5A**). The hazard ratio values of these seven genes were all more than 1 (**Figure 5A**), indicating that they may be negative prognostic factors for HCC patients.

Then, we further analyzed these seven genes through LASSO regression analysis, and the results showed that three m^{6}A regulators (YTHDF1, YTHDF2, and KIAA1429) might be able to construct a prognostic model (**Figures 5B,C**). A multivariate Cox regression analysis was used to construct a risk signature based on the expression of these three genes (**Figure 5D**). The univariate and multivariate Cox regression results are shown in **Table 3**. Risk score = 0.038 × expression of YTHDF1 + 0.064 × expression of YTHDF2 + 0.067 × expression of KIAA1429. The patients were divided into high-risk and low-risk groups by the median risk score (0.939), which served as the cutoff value. The model constructed with the risk signature showed that the AUC values of the time–ROC curve for 3-year overall survival (OS) was 0.665 (**Figure 6A**). As the risk score increased, the mortality rate increased gradually (**Figure 6B**). OS in the high-risk group was significantly shorter than in the low-risk group ($p < 0.05$, **Figure 6C**). The clinical characteristics of the high- and low-risk groups are shown in **Table 2**. The high- and low-risk groups were found to correlate significantly with age, grade, and lymph node metastasis in HCC ($p < 0.05$, **Figure 6D**). To further assess whether risk score can be used as an independent prognostic indicator, we performed univariate Cox and multivariate Cox regression analyses on the risk score. By univariate analysis, we found that the risk score, WHO grades, and TNM stages were all correlated with the OS ($P < 0.001$) (**Figure 6E**). Including these factors into the multivariate Cox regression, the risk score remained significantly associated with the OS ($p < 0.001$) (**Figure 6F**). All the results suggest that the prognostic survival models based on these three genes are useful for prognosis in HCC patients. The expression level of these three genes can be used as independent prognostic factors for HCC in the clinic.

Construction of a Prognostic Nomogram

To further evaluate this risk signature, we used the ROC curve to evaluate the model to predict the survival status of HCC for

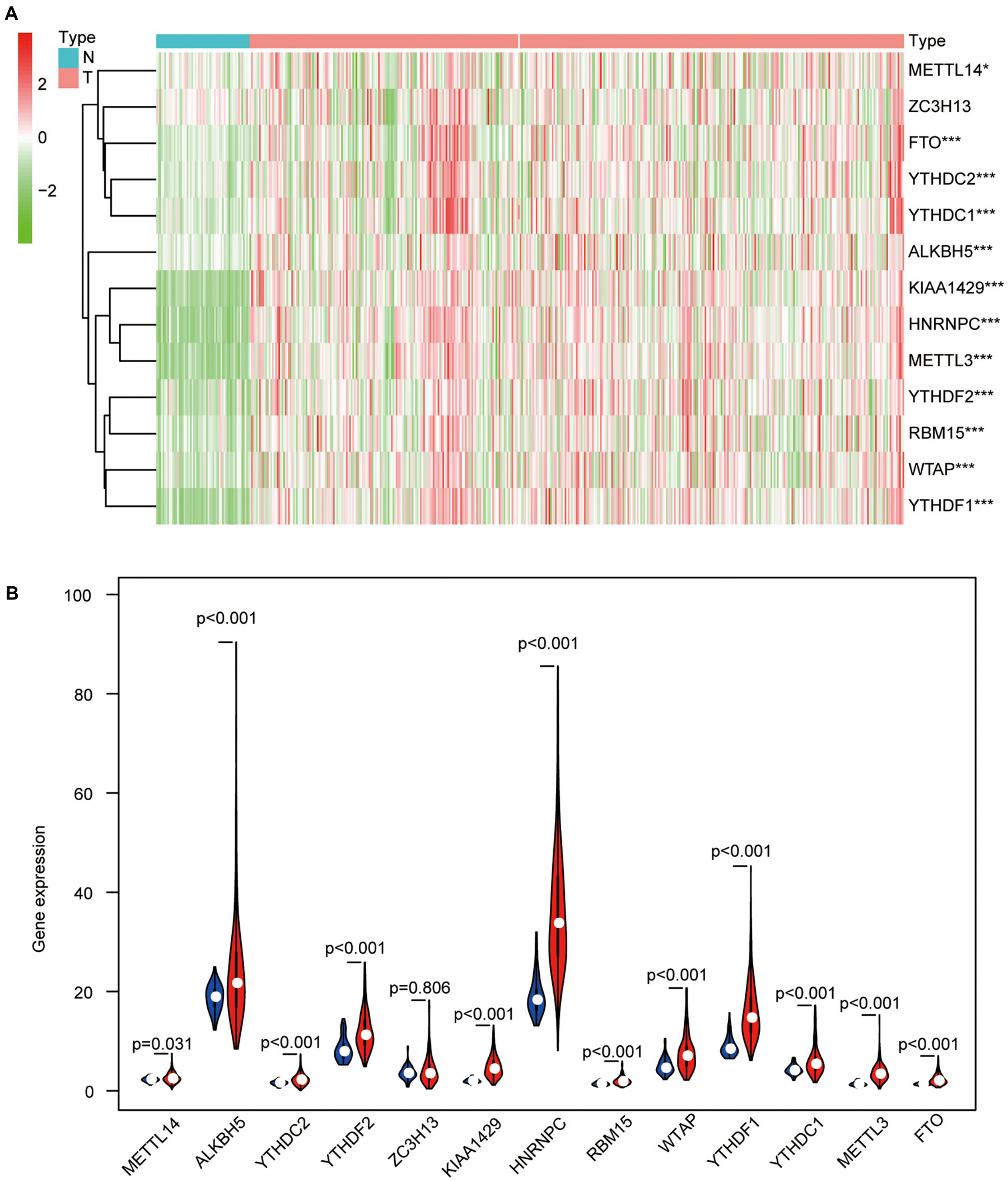


FIGURE 2 | The differential expression levels of 13 m^{6}A regulators in hepatocellular carcinoma (HCC) tissues and adjacent normal tissues. **(A)** The results of the heat map show the expression levels of 13 m^{6}A regulators in 407 HCC samples and 58 adjacent normal tissues (Wilcoxon test). **(B)** The histogram shows the differential expression levels of 13 m^{6}A regulators in 407 HCC samples and 58 adjacent normal tissues (Wilcoxon test). **p* < 0.05 and ****p* < 0.001.

1, 3, and 5 years, respectively. The results showed that the AUC value for 1 year is 0.72, the AUC value for 3 years is 0.665, and the AUC value for 5 years is 0.599 (**Figure 7A**). This result shows that the risk signature has a good prognosis for 1 and 3 years, but for the 5-year survival status, the prediction is not so accurate. The reason may be that the number of HCC patients in the TCGA data set who survived more than 5 years is too small. It may be better to add more samples for analysis.

Then, we constructed a nomogram to predict OS in patients with HCC based on risk scores (**Figure 7B**). The calibration plots showed that the performance of the nomogram was best in predicting 1-, 3-, and 5-year OS (**Figure 7C**).

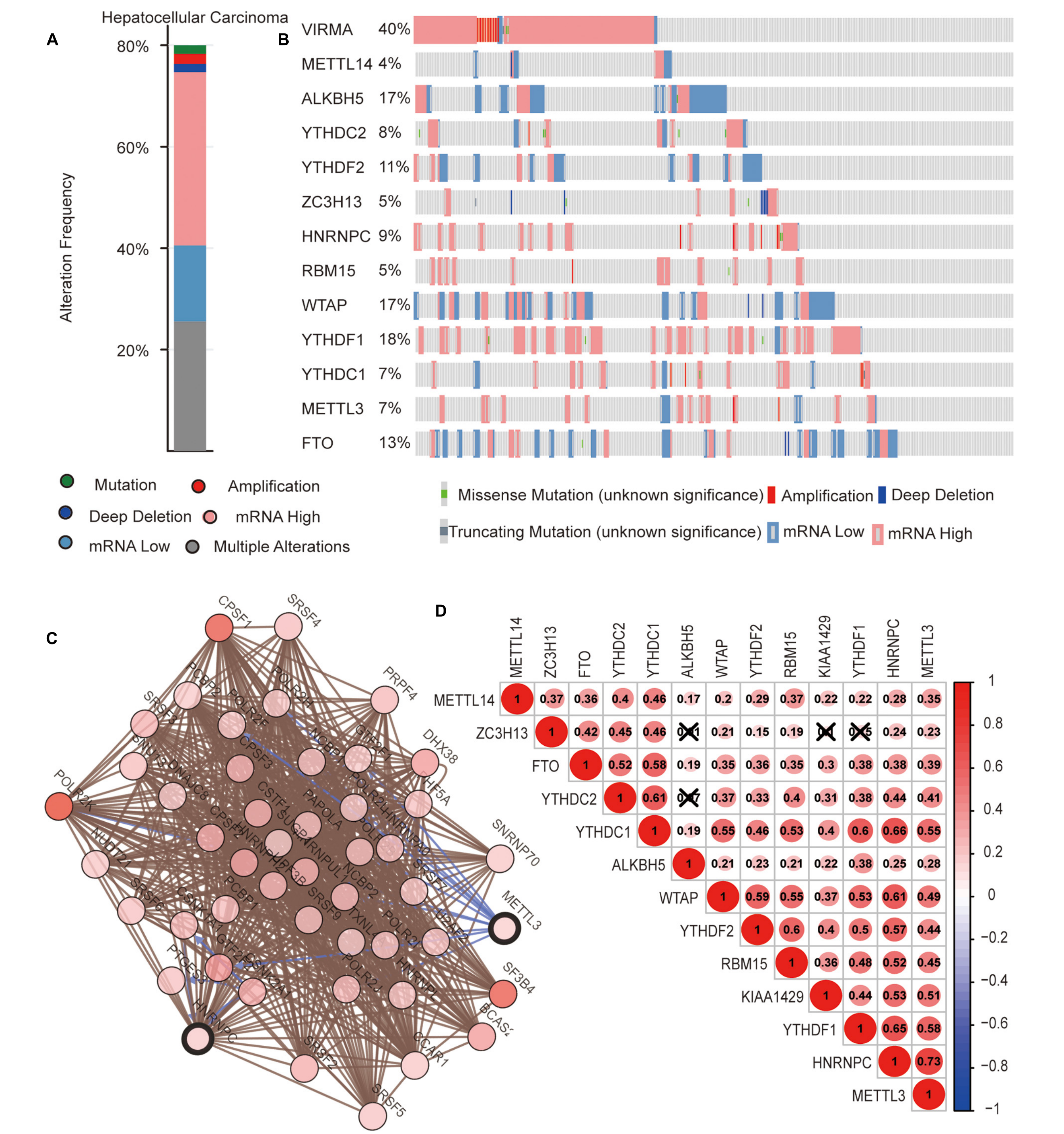


FIGURE 3 | The interaction among m^{6}A regulators in hepatocellular carcinoma (HCC). **(A,B)** The copy number variations and mutation of 13 m^{6}A regulators in HCC from cbioport database. **(C)** The correlation of m6A regulator protein expression in HCC. **(D)** Correlation between the expression of 13 m6A regulators mRNA in HCC.

Consequently, an independent prognostic risk signature was built based on three m^{6}A regulators (YTHDF1, YTHDF2, and KIAA1429) in HCC (**Figure 8**).

DISCUSSION

Accumulating evidence shows that the m^{6}A modification was observed in diverse cancers, which is important for cancer stem

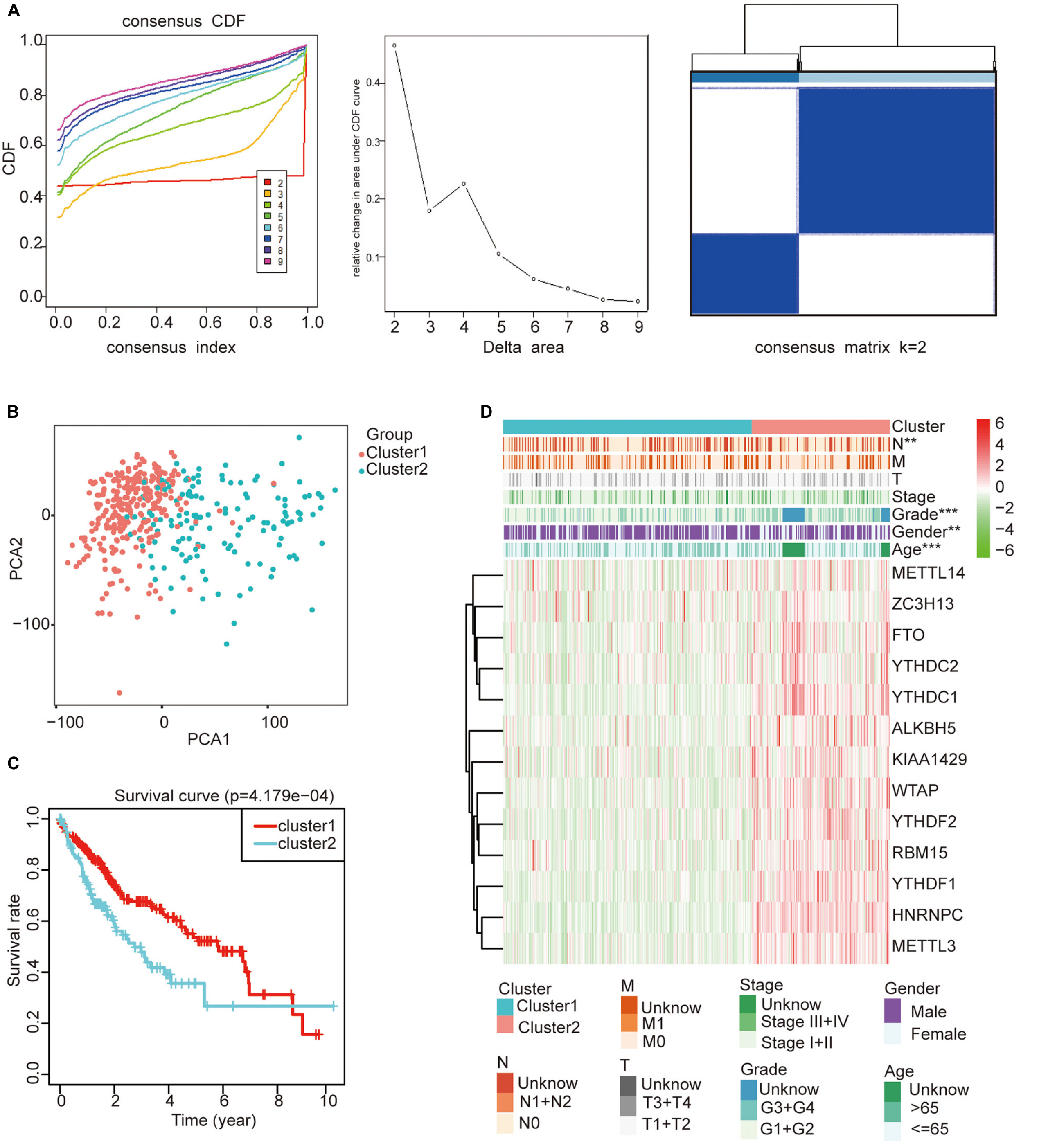


FIGURE 4 | The role of subtypes based on m^{6}A regulator expression profiling in hepatocellular carcinoma (HCC). **(A)** The relative change of the cumulative distribution function and the area under the curve of k = 2–9 for consensus clustering. This result shows that, when K = 2, the m^{6}A regulators can well divide HCC into two types. **(B)** The results show that the subtypes identified based on the expression profile of the m6A modulator can well distinguish HCC into two clusters. **(C)** The comparison of survival curves between cluster 1 and cluster 2 subgroups. **(D)** The comparison of clinicopathological features between cluster 1 and cluster 2 subgroups (Wilcoxon test). $^{**}p < 0.01$ and $^{***}p < 0.001$.

cells self-renewal, cancer cell proliferation, and radiotherapy or chemotherapy resistance (Pan et al., 2018). The formation of m^{6}A is catalyzed by the prominent "writer" proteins (Duan et al., 2019). The downstream cellular functions of m^{6}A rely on its "readers" (Wang et al., 2015; Kretschmer et al., 2018). In addition, HNRNPC is considered as an "m^{6}A switch" to

TABLE 2 | The clinical features of hepatocellular carcinoma.

Variables	Cluster 1 (*n* = 260)	Cluster 2 (*n* = 143)	High risk (*n* = 201)	Low risk (*n* = 202)
Age (years)				
≤ 65	154	78	113	119
> 65	105	33	56	82
Unkonw	1	32	32	1
Gender				
Female	76	64	77	63
Male	184	79	124	193
Grade				
G1 + G2	183	49	86	146
G3 + G4	72	61	81	52
Unknown	5	33	34	4
Tumor invasion (T)				
T1 + T2	197	105	174	155
T3 + T4	60	38	54	44
Unknown	3	0	0	3
Lymph node (N)				
N0	177	100	147	130
N1 + N2	1	7	6	2
Unknown	83	36	48	70
Metastasis (M)				
M0	184	110	153	140
M1	3	4	4	3
Unknown	73	30	44	59
Tumor stage				
Stages I + II	188	95	137	146
Stages III + IV	56	40	53	43
Unknown	16	8	11	13

improve the accessibility of RNA binding proteins (Liu et al., 2015). Some reports show that METTL14 is supposed to be an oncogene in acute myeloid leukemia (Weng et al., 2018). WTAP also acts as an oncogene for the development of malignant tumors and a target for immunotherapy of cancer patients (Xie et al., 2019). KIAA1429 acts as an oncogenic factor in breast cancer and contributes to liver cancer progression (Lan et al., 2019; Qian et al., 2019).

Currently, increasing evidence indicates that m^{6}A regulators are involved in the progression of HCC (Ma et al., 2017; Yang et al., 2017). The "writer" METTL3 contributes to HCC progression by repressing SOCS2 expression (Chen et al., 2018). The "writers" METTL14 acts as an adverse prognosis factor for HCC by promoting miR126 processing (Ma et al., 2017). KIAA1429 is involved in liver cancer progression and regulates the invasion of HCC by altering the m^{6}A modification of ID2 and GATA3 (Qian et al., 2019; Cheng et al., 2019). The "reader" YTHDF2 was closely associated with the malignancy of HCC modulated by MiR145 (Yang et al., 2017). Our results are consistent with these reports. All m^{6}A regulators, except ZC3H13, are highly expressed in HCC, indicating that m^{6}A regulators have key roles in HCC. The PCA results show that m^{6}A regulators can divide hepatocellular carcinoma patients into two types well, and two clustering subgroups have significant differences in WHO grade, gender, age, and lymph node metastasis. All these results suggest that m^{6}A regulators may be a useful diagnostic classification tool for HCC. However, we only explore the relevance of these two types and clinical features. More detailed studies of m^{6}A regulatory factors in the diagnostic classification of HCC are needed.

There is an important question of whether the m^{6}A regulator expression level can act as a prognostic marker in HCC. Li et al. show that KIAA1429, METTL3, and HNRNPC are highly expressed in HCC tissues, while METTL14, ZC3H13, YTHDC1, YTHDC2, and FTO expressions are lower than those in normal tissues. A three-gene (CSAD, GOT2, and SOCS2) signature regulated by METTL14 is efficient for the prognostication of HCC (Li et al., 2020), which suggests that m^{6}A regulators have a clinical prognostic impact in HCC. In our present study, we get similar results that the m^{6}A regulator expression levels are essential for hepatocellular carcinoma prognosis. Differently, in our study, we derive the HCC prognostic signature from the expression of three m^{6}A regulators (YTHDF1, YTHDF2, and KIAA1429). As we have observed, the three-gene signature generated by risk score can stratify the OS for HCC patients. In our results, the expression of all m^{6}A regulators, except for ZC3H13, is higher in the tumor samples than in the adjacent normal tissue. Inconsistent results may result from different sample amounts and sources. More samples are used in our study than in their research, and all our study data of 407 samples are from the TCGA database, while 64 of 307 patients included in their report are from the GSE116174 dataset (others are from the TCGA database). Moreover, that report focuses on studying the function of METTL14 and establishing a METTL14-regulated three-gene (CSAD, GOT2, and SOCS2) signature and nomogram to predict the OS of HCC. However, in our study, the HCC prognostic signature derives from directly using three m^{6}A regulators (YTHDF1, YTHDF2, and KIAA1429). The three regulators are considered to be useful markers for the diagnosis and the treatment of HCC patients in the clinic. Because the signature is generated based on the expression level of m^{6}A regulators which do not involve the downstream target genes, additional trials are needed to find the target genes and the signaling pathways of these three regulators. That should be a good strategy to treat HCC by targeting YTHDF1, YTHDF2, and KIAA1429 combined with targeting their downstream genes.

In our results, a very surprising one is that ZC3H13 expression has no difference between tumor samples and adjacent tissue samples. In addition, ZC3H13 is not correlated with ALKBH5, KIAA1429, and YTHDF1. The previous report shows that the expression of ZC3H13 is lower than those in normal tissues (Li et al., 2020). ZC3H13 is a classical CCCH zinc finger protein localized in human chromosome 13q14.139 (Ouna et al., 2012). As an m^{6}A methylation writer, the role of ZC3H13 in tumors is controversial. A report shows that ZC3H13 serves as a tumor suppressor protein in colon carcinoma and colorectal cancer by regulating the Ras-ERK signaling pathway (Zhu et al., 2019). Other reports consider it as an oncogenic protein by binding with K-ras and activating the NF-κB signal (Knuckles et al., 2018). The controversial roles of ZC3H13 in tumors give us a clue that the essentiality and the functions of m^{6}A RNA methylation in tumors

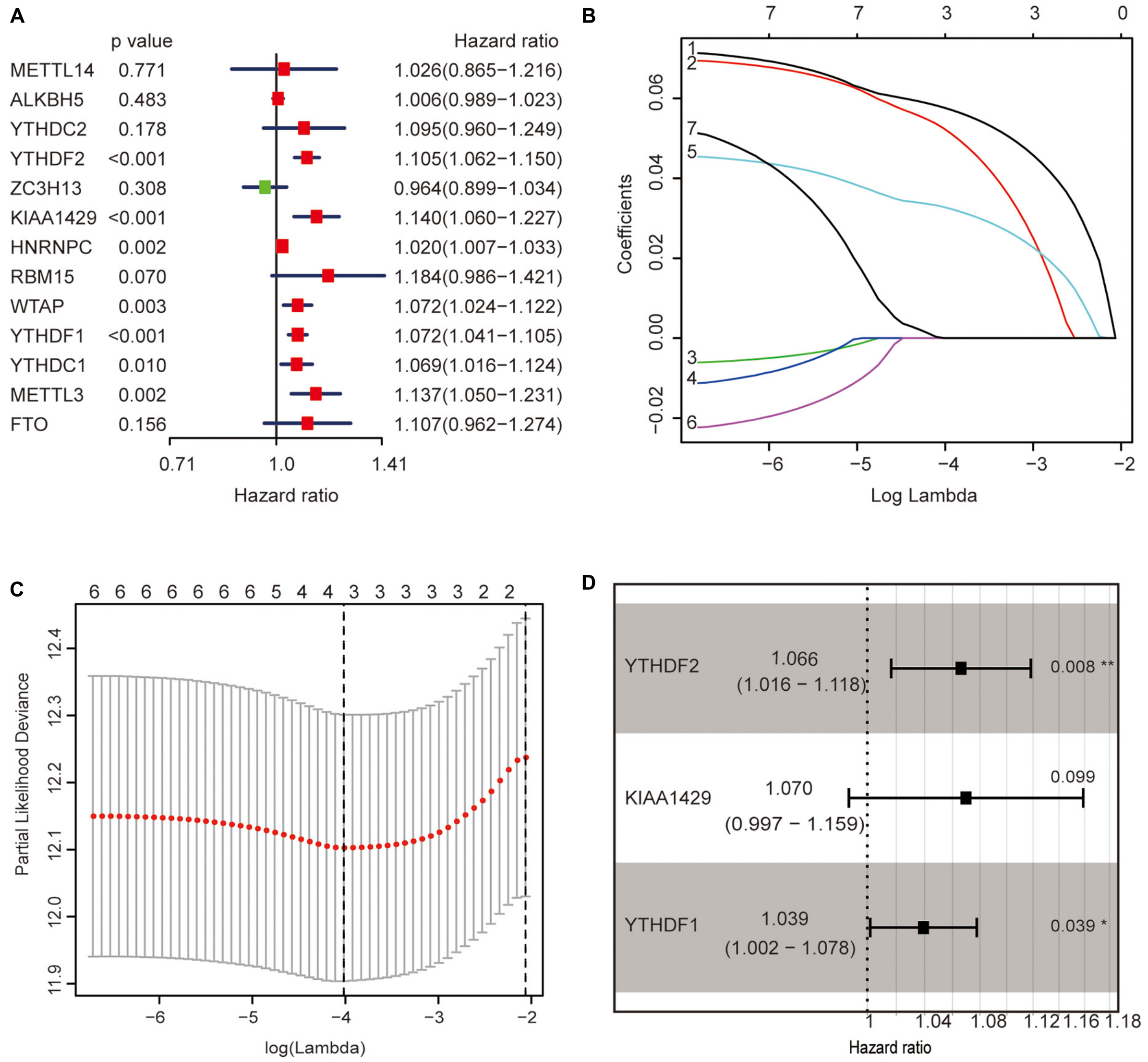


FIGURE 5 | Identification of m^{6}A regulators associated with hepatocellular carcinoma (HCC) prognosis. **(A)** The univariate Cox regression models identified seven m^{6}A regulators associated with overall survival. **(B,C)** Three m^{6}A regulators were identified by LASSO regression analysis. **(D)** Three m^{6}A regulators were identified for constructing a prognostic model by multivariate Cox regression analysis. *$p < 0.05$ and **$p < 0.01$.

TABLE 3 | Univariate and multivariate Cox regression analyses of three m^{6}A regulators in hepatocellular carcinoma.

Variables	Univariate analysis		Multivariate analysis		
	Hazard ratio (HR) (95% CI)	***P*-value**	**Coefficient**	**HR (95% CI)**	***P*-value**
YTHDF2	1.105 (1.062–1.150)	<0.001	0.064	1.066 (1.016–1.118)	0.008
YTHDF1	1.072 (1.041–1.105)	<0.001	0.038	1.039 (1.002–1.078)	0.039
KIAA1429	1.140 (1.060–1.227)	<0.001	0.067	1.070 (0.997–1.159)	0.099

are complicated, and further studies are needed to focus on its prognostic value in HCC.

Another interesting result is that RNA binding protein HNRNPC expression is elevated in HCC. This is consistent with the previous report (Li et al., 2020). The essentiality of HNRNPC in tumors is not clear. Certain studies show that HNRNPC promotes cell proliferation, apoptosis, and tumor growth (Kleemann et al., 2018; Wu et al., 2018). In addition, a high expression of HNRNPC has a poor prognosis and may act as a candidate biomarker for chemoresistance in gastric cancer (Huang et al., 2016). Besides that, HNRNPC also acts as a dengue virus NS1-interacting protein and plays

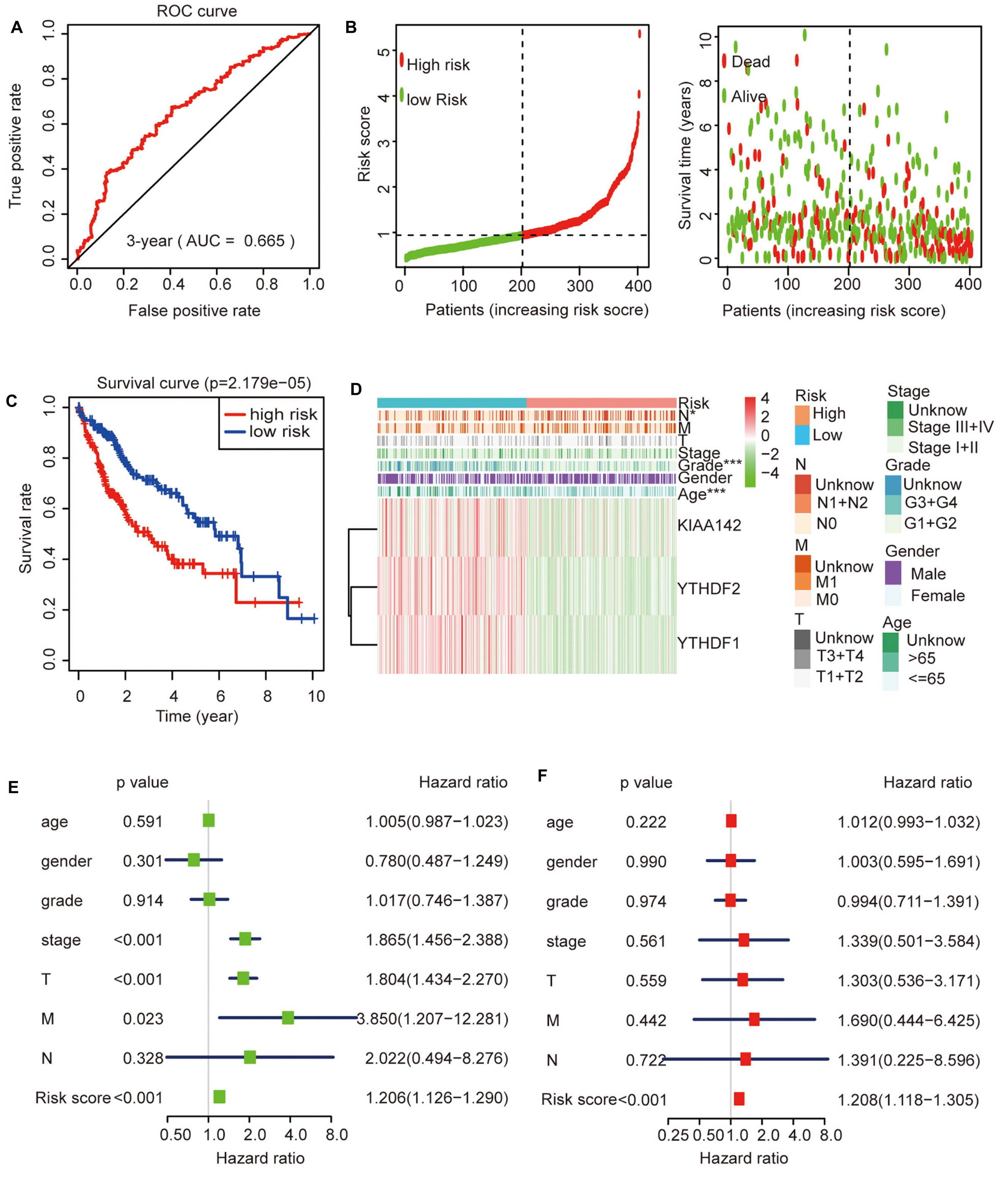


FIGURE 6 | The role of this risk signature in hepatocellular carcinoma (HCC). **(A)** Receiver operating characteristic curve analysis predicts the accuracy of the 3-year survival of risk signature in HCC. **(B)** The risk score analysis of this risk signature in HCC. **(C)** The comparison of survival curves between high- and low-risk groups. **(D)** The comparison of clinicopathological features between high- and low-risk groups (Wilcoxon test). **(E)** The association between clinicopathological factors (including the risk score) and overall survival by univariate Cox regression analyses. **(F)** The association between clinicopathological factors (including the risk score) and overall survival by multivariate Cox regression analyses. **p* < 0.05 and ****p* < 0.001.

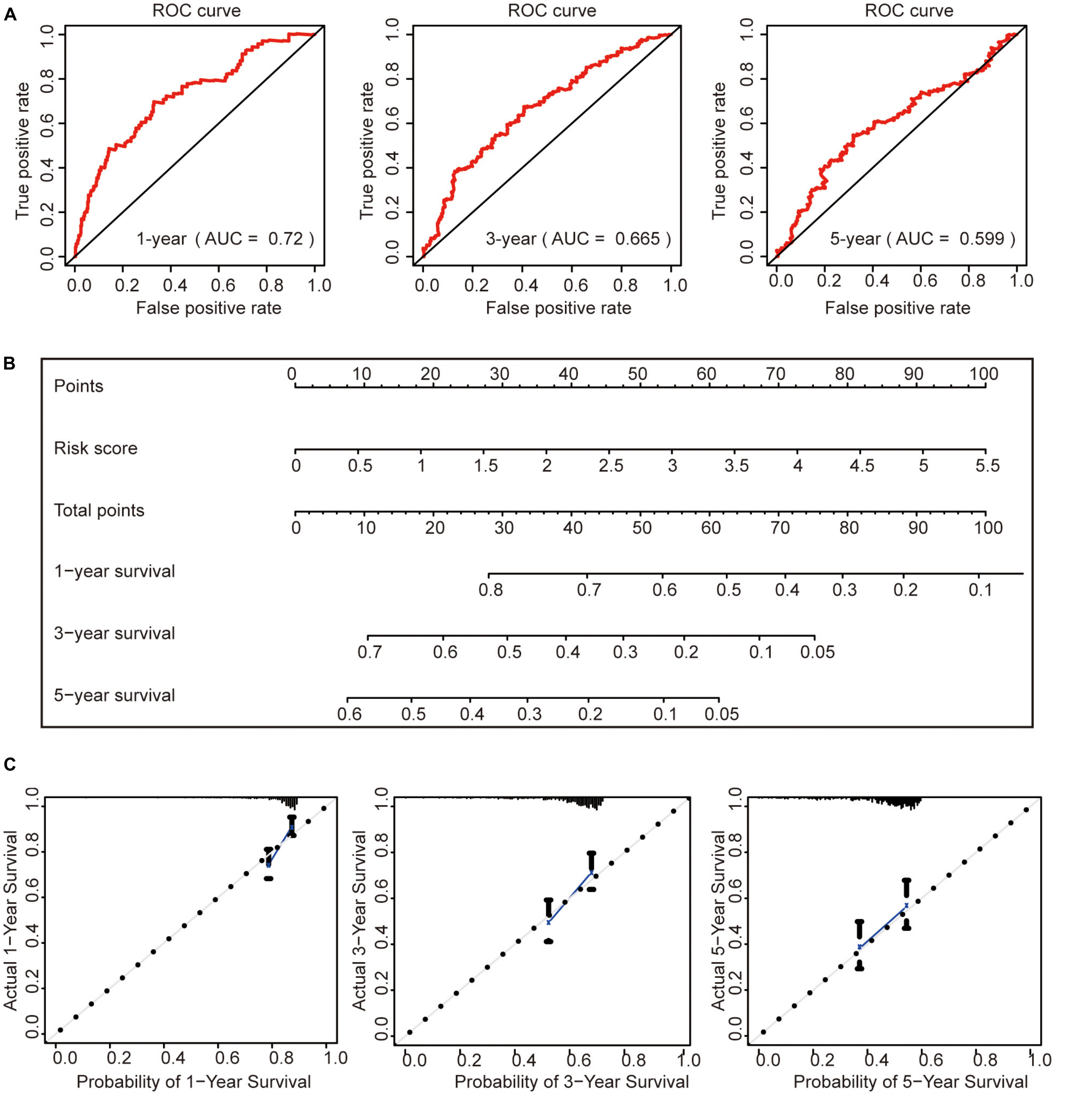


FIGURE 7 | Construction of a prognostic model. **(A)** Receiver operating characteristic curve analysis of the ability of this risk signature to predict hepatocellular carcinoma (HCC) 1-, 3-, and 5-year survival status. **(B)** The construction of the nomogram was based on the risk score in HCC. **(C)** The calibration plot for internal validation of the nomogram.

an important role during the replication of the hepatitis C virus and hepatitis delta virus (Noisakran et al., 2008; Casaca et al., 2011). Our results imply that HNRNPC is a candidate biomarker for HCC. More work is needed to verify the relevant regulatory pathways.

Among 13 m^6A RNA methylation regulators, the m^6A methylation writer VIRMA (KIAA1429) has the most obvious mutation in HCC. VIRMA is identified as the component associated with WTAP in mammalian cells and involved in the regulation of m^6A methylation events in 3’UTR and near the stop codon (Lobo et al., 2019). Certain studies show that KIAA1429 contributes to liver cancer progression through N6-methyladenosine-dependent post-transcriptional modification of GATA3 and regulates the migration and the invasion of HCC by altering the m^6A modification of ID2 mRNA (Cheng et al., 2019; Qian et al., 2019). It is necessary to study the

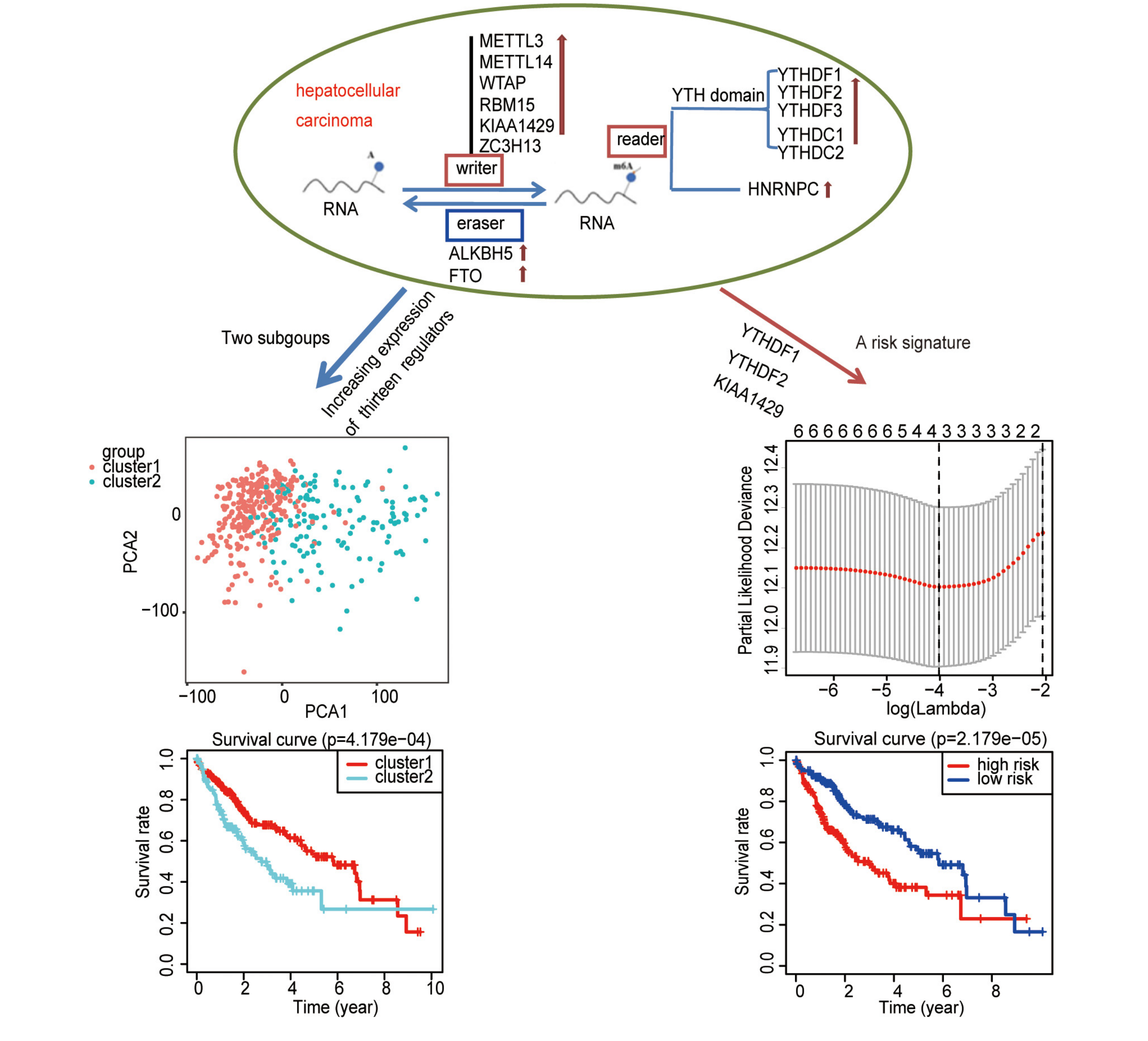


FIGURE 8 | Summary of the potential prognostic value of m^6A regulators in hepatocellular carcinoma.

roles of obvious mutation of VIRMA in HCC occurrence and progression.

CONCLUSION

In conclusion, a high expression of m^6A regulators implies that dysregulated m^6A play important roles in HCC. Furthermore, two clustering subgroups indicate that m^6A RNA methylation plays essential roles in the prognosis and the clinicopathological features of HCC. In addition, a prognostic risk signature with three selected m^6A RNA methylation regulators gives us a clue that m^6A RNA methylation regulators are potentially useful for prognostic stratification and targeting treatment in HCC.

DATA AVAILABILITY STATEMENT

The RNA-seq transcriptome and clinicopathological datas of 407 HCC samples and 58 adjacent normal tissue samples were obtained from TCGA (https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

WL and MT designed the study. WL and FX performed the analysis and drafted the manuscript. WL, MT, CZ, DL, and FX contributed to the editing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (no. 81503074), the Shenzhen Nanshan District Health and Family Planning Bureau (Shenzhen Nanshan District Science and Technology Project, no. 2018045), and the Scientific Research Project of Hunan Health Committee (no. 20200985).

REFERENCES

Bi, Z., Liu, Y., Zhao, Y., Yao, Y., Wu, R., and Liu, Q. (2019). A dynamic reversible RNA N(6) -methyladenosine modification: current status and perspectives. *J. Cell Physiol.* 234, 7948–7956. doi: 10.1002/jcp.28014

Casaca, A., Fardilha, M., da Cruz e Silva, E., and Cunha, C. (2011). The heterogeneous ribonuclear protein C interacts with the hepatitis delta virus small antigen. *Virol. J.* 8:358. doi: 10.1186/1743-422x-8-358

Chai, R. C., Wu, F., Wang, Q. X., Zhang, S., Zhang, K. N., and Liu, Y. Q. (2019). m(6)A RNA methylation regulators contribute to malignant progression and have clinical prognostic impact in gliomas. *Aging* 11, 1204–1225. doi: 10.18632/aging.101829

Chen, J., Sun, Y., Xu, X., Wang, D., He, J., and Zhou, H. (2017). YTH domain family 2 orchestrates epithelial-mesenchymal transition/proliferation dichotomy in pancreatic cancer cells. *Cell Cycle* 16, 2259–2271. doi: 10.1080/15384101.2017.1380125

Chen, M., Wei, L., Law, C. T., Tsang, F. H., Shen, J., and Cheng, C. L. (2018). RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent post-transcriptional silencing of SOCS2. *Hepatology* 67, 2254–2270. doi: 10.1002/hep.29683

Chen, Y., Peng, C., Chen, J., Chen, D., Yang, B., and He, B. (2019). WTAP facilitates progression of hepatocellular carcinoma via m6A-HuR-dependent epigenetic silencing of ETS1. *Mol. Cancer* 18:127. doi: 10.1186/s12943-019-1053-8

Cheng, X., Li, M., Rao, X., Zhang, W., Li, X., and Wang, L. (2019). KIAA1429 regulates the migration and invasion of hepatocellular carcinoma by altering m6A modification of ID2 mRNA. *Oncotargets Ther.* 12, 3421–3428. doi: 10.2147/OTT.S180954

Dai, D., Wang, H., Zhu, L., Jin, H., and Wang, X. (2018). N6-methyladenosine links RNA metabolism to cancer progression. *Cell Death Dis.* 9:124. doi: 10.1038/s41419-017-0129-x

Deng, X., Su, R., Weng, H., Huang, H., Li, Z., and Chen, J. (2018). RNA N(6)-methyladenosine modification in cancers: current status and perspectives. *Cell Res.* 28, 507–517. doi: 10.1038/s41422-018-0034-6

Ding, C., Zou, Q., Ding, J., Ling, M., Wang, W., and Li, H. (2018). Increased N6-methyladenosine causes infertility is associated with FTO expression. *J. Cell Physiol.* 233, 7055–7066. doi: 10.1002/jcp.26507

Duan, H. C., Wang, Y., and Jia, G. (2019). Dynamic and reversible RNA N(6) -methyladenosine methylation. *Wiley Interdiscip. Rev. RNA* 10:e1507. doi: 10.1002/wrna.1507

Huang, H., Han, Y., Zhang, C., Wu, J., Feng, J., and Qu, L. (2016). HNRNPC as a candidate biomarker for chemoresistance in gastric cancer. *Tumour. Biol.* 37, 3527–3534. doi: 10.1007/s13277-015-4144-1

Jia, G., Fu, Y., and He, C. (2013). Reversible RNA adenosine methylation in biological regulation. *Trends Genet.* 29, 108–115. doi: 10.1016/j.tig.2012.11.003

Kleemann, M., Schneider, H., Unger, K., Sander, P., Schneider, E. M., and Fischer-Posovszky, P. (2018). MiR-744-5p inducing cell death by directly targeting HNRNPC and NFIX in ovarian cancer cells. *Sci. Rep.* 8:9020. doi: 10.1038/s41598-018-27438-6

Knuckles, P., Lence, T., Haussmann, I. U., Jacob, D., Kreim, N., and Carl, S. H. (2018). Zc3h13/Flacc is required for adenosine methylation by bridging the mRNA-binding factor Rbm15/Spenito to the m(6)A machinery component Wtap/Fl(2)d. *Genes Dev.* 32, 415–429. doi: 10.1101/gad.309146.117

Kretschmer, J., Rao, H., Hackert, P., Sloan, K. E., Hobartner, C., and Bohnsack, M. T. (2018). The m(6)A reader protein YTHDC2 interacts with the small ribosomal subunit and the 5'-3' exoribonuclease XRN1. *RNA* 24, 1339–1350. doi: 10.1261/rna.064238.117

Lan, T., Li, H., Zhang, D., Xu, L., Liu, H., and Hao, X. (2019). KIAA1429 contributes to liver cancer progression through N6-methyladenosine-dependent post-transcriptional modification of GATA3. *Mol. Cancer* 18, 186–205. doi: 10.1186/s12943-019-1106-z

Li, T., Hu, P. S., Zuo, Z., Lin, J. F., Li, X., and Wu, Q. N. (2019). METTL3 facilitates tumor progression via an m(6)A-IGF2BP2-dependent mechanism in colorectal carcinoma. *Mol. Cancer* 18:112. doi: 10.1186/s12943-019-1038-7

Li, Z., Li, F., Peng, Y., Fang, J., and Zhou, J. (2020). Identification of three m6A-related mRNAs signature and risk score for the prognostication of hepatocellular carcinoma. *Cancer Med.* 9, 1877–1889. doi: 10.1002/cam4.2833

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518, 560–564. doi: 10.1038/nature14234

Liu, N., and Pan, T. (2016). N6-methyladenosine-encoded epitranscriptomics. *Nat. Struct. Mol. Biol.* 23, 98–102. doi: 10.1038/nsmb.3162

Liu, T., Yang, S., Sui, J., Xu, S. Y., Cheng, Y. P., and Shen, B. (2020). Dysregulated N6-methyladenosine methylation writer METTL3 contributes to the proliferation and migration of gastric cancer. *J. Cell Physiol.* 235, 548–562. doi: 10.1002/jcp.28994

Lobo, J., Costa, A. L., Cantante, M., Guimaraes, R., Lopes, P., Antunes, L., et al. (2019). M^6A RNA modification and its writer/reader VIRMA/YTHDF3 in testicular germ cell tumors: a role in seminoma phenotype maintenance. *J. Transl. Med.* 17:79. doi: 10.1186/s12967-019-1837-z

Ma, J. Z., Yang, F., Zhou, C. C., Liu, F., Yuan, J. H., and Wang, F. (2017). METTL14 suppresses the metastatic potential of hepatocellular carcinoma by modulating N(6) -methyladenosine-dependent primary MicroRNA processing. *Hepatology* 65, 529–543. doi: 10.1002/hep.28885

Niu, Y., Zhao, X., Wu, Y. S., Li, M. M., Wang, X. J., and Yang, Y. G. (2013). N6-methyl-adenosine (m6A) in RNA: an old modification with a novel epigenetic function. *Genomics Proteomics Bioinformatics* 11, 8–17. doi: 10.1016/j.gpb.2012.12.002

Noisakran, S., Sengsai, S., Thongboonkerd, V., Kanlaya, R., Sinchaikul, S., and Chen, S. T. (2008). Identification of human hnRNP C1/C2 as a dengue virus NS1-interacting protein. *Biochem. Biophys. Res. Commun.* 372, 67–72. doi: 10.1016/j.bbrc.2008.04.165

Ouna, B. A., Stewart, M., Helbig, C., and Clayton, C. (2012). The Trypanosoma brucei CCCH zinc finger proteins ZC3H12 and ZC3H13. *Mol. Biochem. Parasitol.* 183, 184–188. doi: 10.1016/j.molbiopara.2012.02.006

Pan, Y., Ma, P., Liu, Y., Li, W., and Shu, Y. (2018). Multiple functions of m(6)A RNA methylation in cancer. *J. Hematol. Oncol.* 11:48. doi: 10.1186/s13045-018-0590-8

Piette, E. R., and Moore, J. H. (2018). Identification of epistatic interactions between the human RNA demethylases FTO and ALKBH5 with gene set enrichment analysis informed by differential methylation. *BMC Proc.* 12(Suppl. 9):59. doi: 10.1186/s12919-018-0122-0

Qian, J. Y., Gao, J., Sun, X., Cao, M. D., Shi, L., and Xia, T. S. (2019). KIAA1429 acts as an oncogenic factor in breast cancer by regulating CDK1 in an N6-methyladenosine-independent manner. *Oncogene* 38(Suppl. 7), 1–19. doi: 10.1038/s41388-019-0861-z

Roignant, J. Y., and Soller, M. (2017). m(6)A in mRNA: an Ancient Mechanism for Fine-Tuning Gene Expression. *Trends Genet.* 33, 380–390. doi: 10.1016/j.tig.2017.04.003

Sheng, H., Li, Z., Su, S., Sun, W., Zhang, X., and Li, L. (2019). YTH domain family 2 promotes lung cancer cell growth by facilitating 6-phosphogluconate dehydrogenase mRNA translation. *Carcinogenesis* 41, 541–550. doi: 10.1093/carcin/bgz152

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., and Ma, H. (2015). N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell* 161, 1388–1399. doi: 10.1016/j.cell.2015.05.014

Wei, W., Huo, B., and Shi, X. (2019). miR-600 inhibits lung cancer via downregulating the expression of METTL3. *Cancer Manag. Res.* 11, 1177–1187. doi: 10.2147/cmar.S181058

Weng, H., Huang, H., Wu, H., Qin, X., Zhao, B. S., and Dong, L. (2018). METTL14 inhibits hematopoietic Stem/Progenitor differentiation and promotes leukemogenesis via mRNA m(6)A modification. *Cell Stem Cell* 22, 191.e9–205.e9. doi: 10.1016/j.stem.2017.11.016

Wu, Y., Zhao, W., Liu, Y., Tan, X., Li, X., and Zou, Q. (2018). Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response. *Embo J.* 37:e99017. doi: 10.15252/embj.201899017

Xie, W., Wei, L., Guo, J., Guo, H., Song, X., and Sheng, X. (2019). Physiological functions of Wilms' tumor 1-associating protein and its role in tumourigenesis. *J. Cell Biochem.* 120, 10884–10892. doi: 10.1002/jcb.28402

Yang, Y., Hsu, P. J., Chen, Y. S., and Yang, Y. G. (2018). Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* 28, 616–624. doi: 10.1038/s41422-018-0040-8

Yang, Z., Li, J., Feng, G., Gao, S., Wang, Y., and Zhang, S. (2017). MicroRNA-145 modulates N(6)-Methyladenosine levels by targeting the 3'-Untranslated mRNA Region of the N(6)-Methyladenosine binding YTH domain family 2 Protein. *J. Biol. Chem.* 292, 3614–3623. doi: 10.1074/jbc.M116.749689

Zhang, J., Bai, R., Li, M., Ye, H., Wu, C., and Wang, C. (2019). Excessive miR-25-3p maturation via N(6)-methyladenosine stimulated by cigarette smoke promotes pancreatic cancer progression. *Nat. Commun.* 10:1858. doi: 10.1038/s41467-019-09712-x

Zhao, X., Chen, Y., Mao, Q., Jiang, X., Jiang, W., and Chen, J. (2018). Overexpression of YTHDF1 is associated with poor prognosis in patients with hepatocellular carcinoma. *Cancer Biomark* 21, 859–868. doi: 10.3233/CBM-170791

Zhou, J., Wang, J., Hong, B., Ma, K., Xie, H., and Li, L. (2019). Gene signatures and prognostic values of m6A regulators in clear cell renal cell carcinoma - a retrospective study using TCGA database. *Aging* 11, 1633–1647. doi: 10.18632/aging.101856

Zhu, D., Zhou, J., Zhao, J., Jiang, G., Zhang, X., and Zhang, Y. (2019). ZC3H13 suppresses colorectal cancer proliferation and invasion via inactivating Ras-ERK signaling. *J. Cell Physiol.* 234, 8899–8907. doi: 10.1002/jcp.27551

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 09 October 2020
doi: 10.3389/fgene.2020.518949

tRNA Fragments Populations Analysis in Mutants Affecting tRNAs Processing and tRNA Methylation

Anahi Molla-Herman[1], Margarita T. Angelova[2]†, Maud Ginestet[1], Clément Carré[2]†, Christophe Antoniewski[3]*† and Jean-René Huynh[1]*

[1] Collège de France, CIRB, CNRS Inserm UMR 7241, PSL Research University, Paris, France, [2] Transgenerational Epigenetics & Small RNA Biology, Sorbonne Université, CNRS, Laboratoire de Biologie du Développement - Institut de Biologie Paris Seine, Paris, France, [3] ARTbio Bioinformatics Analysis Facility, Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Paris, France

OPEN ACCESS

Edited by:
Pavel V. Baranov,
University College Cork, Ireland

Reviewed by:
Isidore Rigoutsos,
Thomas Jefferson University,
United States
Lluís Ribas De Pouplana,
Institute for Research in Biomedicine,
Spain

****Correspondence:***
Anahi Molla-Herman
anahi.molla-herman@college-de-france.fr
Christophe Antoniewski
christophe.antoniewski@upmc.fr

†ORCID:
Margarita T. Angelova
orcid.org/0000-0002-0046-9891
Clément Carré
orcid.org/0000-0002-9668-4288
Christophe Antoniewski
orcid.org/0000-0001-7709-2116

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *10 December 2019*
Accepted: *03 September 2020*
Published: *09 October 2020*

Citation:
Molla-Herman A, Angelova MT, Ginestet M, Carré C, Antoniewski C and Huynh J-R (2020) tRNA Fragments Populations Analysis in Mutants Affecting tRNAs Processing and tRNA Methylation. Front. Genet. 11:518949. doi: 10.3389/fgene.2020.518949

tRNA fragments (tRFs) are a class of small non-coding RNAs (sncRNAs) derived from tRNAs. tRFs are highly abundant in many cell types including stem cells and cancer cells, and are found in all domains of life. Beyond translation control, tRFs have several functions ranging from transposon silencing to cell proliferation control. However, the analysis of tRFs presents specific challenges and their biogenesis is not well understood. They are very heterogeneous and highly modified by numerous post-transcriptional modifications. Here we describe a bioinformatic pipeline (tRFs-Galaxy) to study tRFs populations and shed light onto tRNA fragments biogenesis in *Drosophila melanogaster*. Indeed, we used small RNAs Illumina sequencing datasets extracted from wild type and mutant ovaries affecting two different highly conserved steps of tRNA biogenesis: 5′pre-tRNA processing (RNase-P subunit Rpp30) and tRNA 2′-O-methylation (dTrm7_34 and dTrm7_32). Using our pipeline, we show how defects in tRNA biogenesis affect nuclear and mitochondrial tRFs populations and other small non-coding RNAs biogenesis, such as small nucleolar RNAs (snoRNAs). This tRF analysis workflow will advance the current understanding of tRFs biogenesis, which is crucial to better comprehend tRFs roles and their implication in human pathology.

Keywords: *Drosophila*, Nm methylation, RNase P, tRNA, tRFs, oogenesis

INTRODUCTION

Transfer RNAs (tRNAs) are molecules of ∼75 nt transcribed by RNA polymerase III that adopt a typical cloverleaf secondary structure. They are ancient molecules required for protein translation and are encoded by hundreds of genes (∼300 in *Drosophila*, ∼400 in humans) localized in clusters throughout the genome in some species (Haeusler and Engelke, 2006; Willis and Moir, 2018). tRNAs can be transcribed in the nucleus or in mitochondria. Once transcribed, tRNA precursors (pre-tRNAs, ∼125 nt) are processed by the highly conserved ribozymes RNAse P and Z, to cleave the 5′ leader and the 3′ trailer, respectively (Jarrous, 2017). Then, a CCA trinucleotide tag is added at the 3′ end of mature tRNAs by a specific enzyme (RNA polymerase ATP(CTP):tRNA nucleotidyltransferase) present in all kingdoms of life. CCA tag plays a role in tRNA amino-acylation, tRNA export toward the cytoplasm, and tRNA quality control (Wellner et al., 2018). RNase P is formed by one RNA molecule and several protein subunits such as Rpp30, highly conserved throughout evolution (Jarrous, 2017). In some species, RNAse P can also cleave non-canonical targets such as rRNA, snoRNA, some long non-coding RNA and RNAs containing N6-methyladenosine (m^6A) (Coughlin et al., 2008; Jarrous, 2017; Park et al., 2019).

Importantly, tRNA biogenesis involves the production of small RNA molecules, hereafter referred to as tRNA fragments (tRFs), derived either from tRNA precursors or from cleavage of mature tRNAs. tRFs are found in a wide variety of organisms and tissues and are associated with several pathologies such as cancer and neurodegeneration (reviewed Kumar et al., 2016; Soares and Santos, 2017; Shen et al., 2018). Despite recent efforts to develop tools describing tRFs populations (Thompson et al., 2008; Kumar et al., 2014b; Selitsky and Sethupathy, 2015; Pliatsika et al., 2016; Loher et al., 2017a; Schorn et al., 2017; Kuscu et al., 2018; Liu et al., 2018; Guan et al., 2019) tRFs analyses from different laboratories remain difficult to compare (**Supplementary Table 1**). Indeed, finding consensus tools to study different species and tissues is difficult for several reasons (Telonis et al., 2016). First, different factors can vary in RNA sample preparation (protocol, tissue, species, sex, population. . .) as well as in library preparation. Secondly, tRFs nomenclature, bioinformatics workflows, bioinformatics softwares and parameters vary depending on the laboratory. Thirdly, tRNAs-genome references are different in each species[1] and their construction to get all tRFs types can vary depending on the study. Fourthly, it has been suggested that very small RNAs (14–16 nt) could originate not only from tRNA molecules, but also from highly repeated regions unrelated with tRNA, or from incomplete (truncated) pseudo-tRNAs in some organisms, with different copy numbers and genomic localizations (Telonis et al., 2014). Also, tRNAs can be substrates for the production of other types of small ncRNA such as miRNAs or piRNAs (Maute et al., 2013; Keam et al., 2014; Honda et al., 2017). This problem can be addressed by studying tRFs that match the "non-tRNA-space," which corresponds to the whole genome excluding tRNA genes (Telonis et al., 2016; Loher et al., 2017b). Importantly, while trying to exclude false positive tRFs, one could increase false negative error rate, since it is difficult to know the real origin of tRFs: "tRNA space," "non-tRNA space," or both. In addition, some nuclear tRNAs can be similar to mitochondrial tRNAs in vertebrates (especially in primates). These tRNAs, called tRNA-lookalikes, could be a source of tRFs, whose origin is difficult to determine. However, no tRNAs-lookalike were found in *Drosophila* using perfect match alignments, and only one tRNA-lookalike was found allowing mismatches (Telonis et al., 2014, 2015a). Finally, several tRNAs corresponding to the same amino-acid share the same sequence[2]. Thus, these tRNAs will generate different types of tRFs which can be attributed randomly to one of these tRNAs or to all of them (ex.tRNA:Val-CAC-2-1 to 2-6). This problem can be solved by collapsing tRNA sequences to obtain unique tRNA mature sequences. However, this collapse cannot be done with the extended sequences of tRNAs (25 nt and 80 nt flanking mature tRNA) since these sequences are different. Besides, only some bioinformatic analysis have tried to validate tRFs profiles in parallel, by performing Northern Blot (Torres et al., 2019) (**Supplementary Table 1**).

The impact of tRFs levels in various biological processes is currently under investigation and multiple processes have already been identified, amongst which stands gene expression and translation control, transposon silencing, ncRNA processing, histone levels control, cell proliferation and DNA damage response modulation (Goodarzi et al., 2015; Sharma et al., 2016, 2018; Kuscu et al., 2018; Li et al., 2018; Liu et al., 2018; Schorn and Martienssen, 2018; Shen et al., 2018; Boskovic et al., 2019; Guan et al., 2019; Su et al., 2019).

In wild type condition, when RNAse P cleaves the 5′ trailer of tRNA-precursor, the resulting fragment is believed to be degraded by the ribonuclease translin–TRAX complex (C3PO) (Li Z. et al., 2012; **Figure 1A**). Then, RNase Z cleaves the 3′ trailer forming tRFs-1 (also called tRFs-3′U because the Poly-U tract is typically found at 3′ of pre-tRNAs) (Rossmanith, 2012; Kumar et al., 2015). Once mature, tRNAs can be cleaved forming small fragments: tRFs-5 (or 5′tRFs, originating from 5′) or tRFs-3 (or 3′tRFs, originating from 3′ including CCA tag). These cleavages could be done by Dicer or by other endonucleases that remain to be discovered (Cole et al., 2009; Sobala and Hutvagner, 2011; Li L. et al., 2012; Kuscu et al., 2018; Shen et al., 2018; Su et al., 2019). Internal tRFs (i-tRFs) are contained to the interior of the mature tRNA sequence and can straddle the anticodon (Telonis et al., 2015b). Also, mature tRNA molecules can be cut in 2 halves (tRNA halves ∼35 nt) which play important roles in different stress conditions, such as hypoxia or temperature changes (Fu et al., 2008; Thompson et al., 2008; Shen et al., 2018; Akiyama et al., 2020). Intriguingly, in some neuropathologies, tRNA precursors can be cleaved and generate tRNA fragments (∼40 nt) which include the 5′ trailer (Hanada et al., 2013). Spanner-tRFs are another class of tRFs that occur rarely and can be formed before the RNase Z cleavage and before CCA addition, spanning the CCA editing point. Finally, transcription termination associated tRNA fragments (taRFs) are formed when RNA Pol-III does not finish transcription properly. Interestingly, altered tRF populations have been discovered in mouse mutants for RNase Z (ELAC2), which have cardiomyopathy and premature death (Siira et al., 2018). However, it is still not known whether RNase P also plays a role in tRFs formation.

Aberrant tRFs populations could have *trans* effects on gene expression. They could target different RNAs by sequence complementarity, by guiding Argonaute proteins similarly to other small non-coding RNAs like miRNAs (microRNAs), siRNAs (small interfering RNAs) and piRNAs (Piwi-interacting RNAs) (Kim et al., 2009; Kumar et al., 2014a; Yamanaka and Siomi, 2015). miRNAs are small RNAs known to cleave mRNAs or inhibit mRNA translation (Jonas and Izaurralde, 2015). piRNAs and siRNAs are small RNAs known to silence transposable elements (TEs) (Czech et al., 2018). Among tRFs targets, some TEs and gene sequences have been identified, linking tRFs to several cellular processes and pathologies, such as translation control, cell signaling, development, proteasome regulation or metabolism (Goodarzi et al., 2015;

[1] http://gtrnadb.ucsc.edu/

[2] http://gtrnadb.ucsc.edu/genomes/eukaryota/Dmela6/Dmela6-align.html

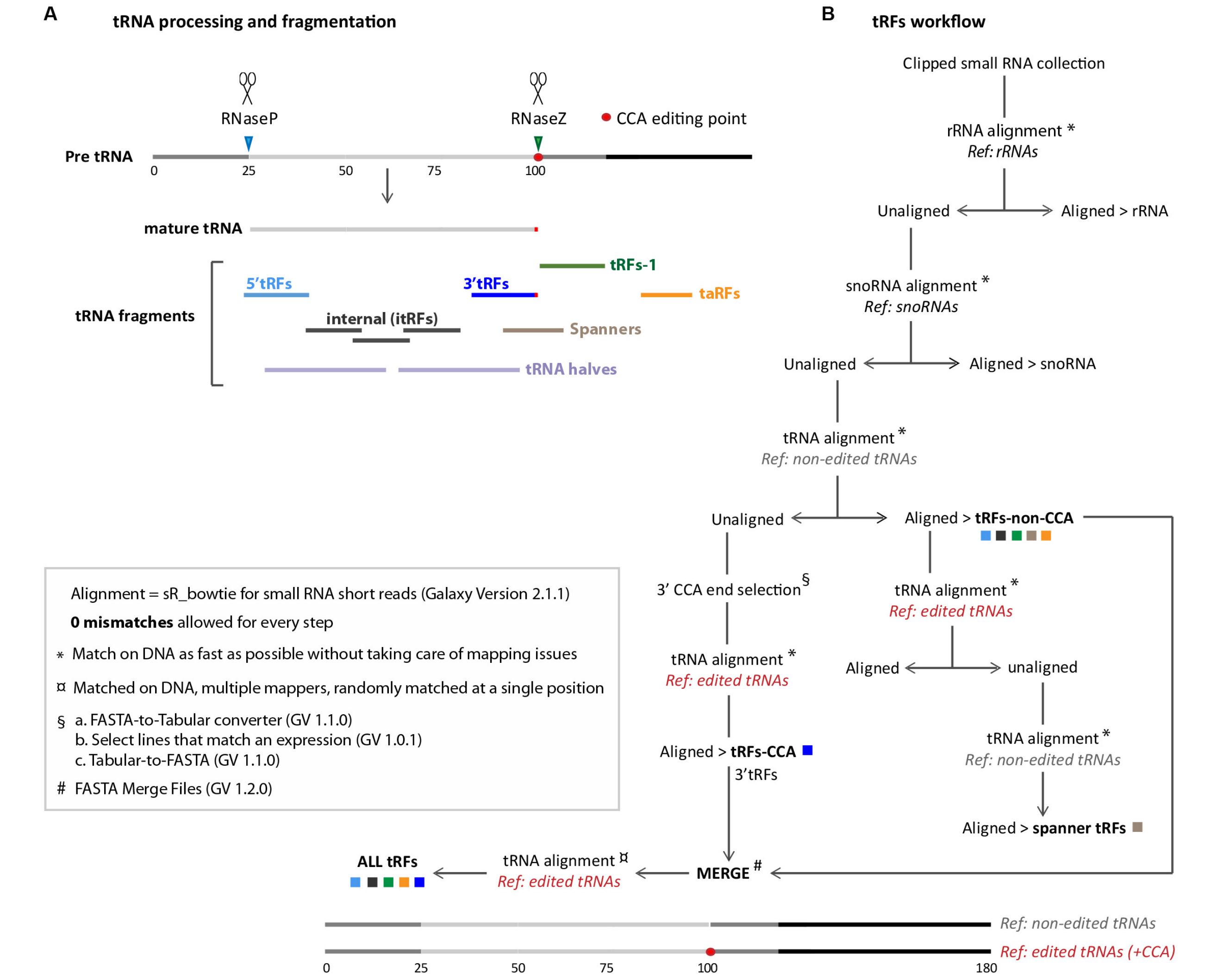


FIGURE 1 | General workflow for tRNA fragments (tRFs) classes extraction: **(A)** tRNA processing and tRNA fragments are depicted. The 5′ tail of pre-tRNAs is cleaved by RNase P (blue arrowhead) and the 3′ tail is cleaved by RNase Z (green arrowhead). 5′ cleavage product is believed to be degraded whereas RNase Z cleavage product forms tRFs-1 (green line). Mature tRNAs (light gray line) is edited by the addition of 3′-tRFs motif (red dot). Several types of tRFs can be generated from mature RNAs, such as 5′-tRFs (light blue line), 3′-tRFs (dark blue line), and inner tRFs (i-tRFs) belonging to the anticodon region (dark gray lines). Spanner-tRFs can be formed before the addition of CCA from tRNA-precursors, spanning the CCA region (light brown line). Transcription associated (taRFs, orange line) can be formed from downstream regions of tRNAs. Longer tRNA halves are represented with light purple lines. **(B)** Galaxy-developed workflow for extraction of all tRFs classes, described in **A**. Alignments were done with SR_Bowtie tool for small RNA short reads (version 2.1.1) using two types of matching: * Match on DNA as fast as possible or ¤ Match on DNA, multiple mappers. "*Ref.*" are the different genome references used for alignments in this pipeline: rRNA, snoRNA, tRNA-non-edited or tRNA-CCA-edited. For tRNA-non-edited reference construction, mature tRNAs (75 nt) were compared with tRNA-precursors (125 nt) to determine RNase P and RNase Z cleavage points. 25 nt were added upstream at 5′, and 80 nt downstream, right after the RNase Z cleavage point (25 + 75 + 80 = 180 nt approximately). For tRNA-CCA-edited reference construction, a CCA motif was added to the non-edited reference, precisely at the 3′CCA edition point (red dot). tRFs CCA or non-CCA can be treated separately or altogether (ALL-tRFs).

Karaiskos and Grigoriev, 2016; Sharma et al., 2016; Martinez, 2017; Schorn et al., 2017; Kim et al., 2019; Mo et al., 2019; Telonis et al., 2019). tRFs thus emerge as potential biomarkers and therapeutic targets for human pathologies (Balatti et al., 2017; Zhu et al., 2018).

Currently, around 150–170 RNA modifications are known, and recent reports show that RNA modifications defects play an important role in tRFs production in different organisms. Epitranscriptomics have recently emerged as a new field to comprehend the mechanisms underlying RNA modifications and their role in gene expression. Indeed, tRNAs are the most extensively modified RNAs in cells (up to 25% of nucleotides per tRNA) (Delaunay and Frye, 2019; Ontiveros et al., 2019; Guzzi and Bellodi, 2020). These marks are believed to help tRNAs to respond to a wide range of environmental cues, stimuli and stress. They play crucial roles at all tRNA biogenesis steps, such as sequence maturation, folding, recycling and degradation. Interestingly, there is a crosstalk between the different modification pathways and a large amount of tRNA modification enzymes defects have been linked to human pathologies (Angelova et al., 2018; Sokołowski et al., 2018; Lyons et al., 2018; Dimitrova et al., 2019).

In *Drosophila* it has been recently shown that methylation marks protect tRNAs from cleavage and that aberrant tRFs populations accumulate in methylation mutants: on the one hand, Dnmt2 mutation impairs m^5C methylation (Schaefer et al., 2010; Durdevic et al., 2013a; Genenncher et al., 2018). On the other hand, dTrm7_34 (CG7009) and dTrm7_32 (CG5220) mutation impairs 2′-O-methylation (Angelova et al., 2020). 2′-O-methylation is one of the most common RNA modifications and consists in the addition of a methyl group to the 2′ hydroxyl of the ribose moiety of a nucleoside, being also known as Nm. It is found in tRNAs, rRNAs, snRNAs (small nuclear RNAs), at the 3′ end of some small non-coding RNAs (such as piRNAs), and at some sites on mRNAs (Ontiveros et al., 2019). This modification plays a wide range of roles in RNA structure, stability and interactions (Dimitrova et al., 2019). It has been recently shown that *Drosophila* proteins dTrm7_34 and dTrm7_32 are the functional orthologs of yeast TRM7 (Pintard, 2002) and human FTSJ1 (Guy et al., 2015) respectively, which are involved in 2′-O-methylation of the anticodon loop of several conserved tRNAs substrates (tRNA-Leu, Trp, Phe). Mutations of these tRNAs methyltransferases in *Drosophila* lead to lifespan reduction, small non-coding RNA pathways dysfunction and increased sensitivity to RNA virus infections, besides specific tRFs accumulation (Angelova et al., 2020).

Despite their abundance, only a very limited subset of RNA modifications can be detected and quantified by current high-throughput analytical techniques such as ARM-seq, and substantial efforts are being invested for the development of this field (Cozen et al., 2015; Dai et al., 2017). Some modifications, such as 2′-O-methylation, can have an impact on classical sequencing techniques during library preparation (reverse transcription blocking) and could introduce a bias in the analyses, such as in the type of tRFs preferentially sequenced which can have different degrees of modification (Motorin and Helm, 2019). However, one study have reported that tRNA modifications only have a limited impact on data mining when studying tRFs in The Cancer Genome Atlas (Telonis et al., 2019). Indeed, we still do not know the impact of each RNA modification on small RNA sequencing, and thousands of small RNA datasets have already been generated with Illumina sequencing techniques. Thus, a wide range of wild type and mutant datasets from different species are available[3,4] and their analysis can bring important new information on tRFs biogenesis and/or stability.

Since tRFs biogenesis remains obscure, we developed and describe a user-friendly tRFs-pipeline for *Drosophila melanogaster* based on Galaxy environment (tRFs-Galaxy), with workflows and tools that can be easily shared with the scientific community. To do so, we took advantage of several *Drosophila* datasets (15–29 nt) generated in our laboratories: *Rpp30* mutants, which affect tRNA processing, and dTrm7_34 and dTrm7_32, which affect tRNA Nm methylation (Molla-Herman et al., 2015; Mollà-Herman et al., 2019; Angelova et al., 2020). We believe that this study will help to better understand the known pathways of tRFs biogenesis as well as to uncover new tRFs biogenesis factors and unexpected crosstalks between different RNA regulatory mechanisms, crucial for gene expression.

MATERIALS AND METHODS

Fly Stocks

Fly stocks are described in Molla-Herman et al. (2015) and Angelova et al. (2020).

RNA Extraction From Ovaries

RNA was extracted from *Drosophila* ovaries following standard methods detailed in Molla-Herman et al. (2015) and Angelova et al. (2020).

Small RNA Sequencing

RNA samples of 3–5 μg were used for High-throughput sequencing using Illumina HiSeq, 10% single-reads lane 1 × 50 bp (Fasteris). 15–29 nt RNAs sequences excluding rRNA (riboZero) were sequenced. All the analyses were performed with Galaxy tools[5]. Workflows are available upon request. Data set deposition is described in Molla-Herman et al. (2015) and Angelova et al. (2020). European Nucleotide Archive (ENA) of the EMBL-EBI[6], accession numbers are: PRJEB10569 (Rpp30 mutants), PRJEB35301 and PRJEB35713 (Nm mutants).

Clipping and Concatenation

Raw data were used for clipping the adaptors [Clip adapter (Galaxy-Version 2.3.0, owner: artbio)] and FASTQ quality control was performed [FastQC Read Quality reports (Galaxy-Version 0.72)]. Since replicates were homogeneous in quality and analysis (replicates for heterozygous and homozygous *dTrm7_34** flies and triplicates for *dTrm7_34*- dTrm7_32** double mutants) we merged them [Concatenate multiple datasets tail-to-head (Galaxy-Version 1.4.1, owner: artbio) to have single fasta files. *dTrm7_34*/Def9487 as well as* $Rpp30^{18.2}$, mnk^{P6} *homozygous and* $Rpp30^{PE}/Rpp30^{18.2}$ datasets were used to obtain normalization numbers but are not shown in the figures for simplicity (**Supplementary Figure 8**).

Data Normalization Using DeSeq miRNA Counts

Data were normalized with library Normalization Factors (NF) obtained by using [DESeq geometrical normalization (Galaxy-Version 1.0.1, owner: artbio)] with miRNA counts obtained using [miRcounts (Galaxy-Version 1.3.2)], allowing 0 mismatch (MM). Then, 1/NF values were used in Galaxy small RNA maps (**Supplementary Figures 8A,B**).

[3] www.ebi.ac.uk/ena

[4] www.ncbi.nlm.nih.gov/geo

[5] https://mississippi.snv.jussieu.fr

[6] http://www.ebi.ac.uk/ena

Data Normalization With DeSeq Using tRFs Counts

To create tRFs expression heatmaps, all-tRFs read counts were normalized using [DESeq Normalization (Galaxy-Version 1.0.1, owner: artbio)] giving rise to a Normalized Hit Table.

Genome References

rRNA, snoRNA, miRNA, ncRNA, intergenic, genic references and Transposable Elements (Ensemble canonical TE) were obtained from Ensembl Biomart[7]. For tRNAs, we created a genome reference of extended pre-tRNAs adding 25 nt upstream and 80 nt downstream of tRNAs genome annotations. These sequences referred to as "non-edited tRNAs" have an average length of ∼180.3 nt (Standard Deviation 14.9 nt) for nuclear tRNAs and ∼170 nt (Standard Deviation 6.2 nt) for mitochondrial tRNAs. Sixteen tRNA sequences have an intron that has to be spliced. To analyze tRFs carrying 3′CCA motif we inserted a CCA in the genomic precursor sequence, at the position where tRNAs are edited after pre-tRNA maturation. We called this reference "CCA-edited-tRNAs." To study the "non-tRNA space" we created a reference genome excluding known tRNAs gene segments. To avoid multimapping of tRFs to several tRNAs with similar sequences we collapsed tRNAs mature sequences into "Unique Mature tRNAs" and we added CCA tag. We split the snoRNA sequences in two reference sets, one with box C/D snoRNAs whose mature sequences are equal or less than 120 nt long, the other with box H/ACA snoRNAs whose mature sequences are more than 120 nt long.

General Small RNA Annotation

Small RNA reads files were first depleted from rRNAs by discarding reads aligning to rRNA genome reference. Then, we annotated the small RNAs by iterative alignments to the various references using the tool [Annotate smRNA dataset (Galaxy-Version 2.4.0, owner: artbio)] and allowing 0 mismatches. For annotation cascades, iterative alignments were performed in the following order: tRNA, tRNA-CCA-edited, miRNA, TE-derived, all-ncRNA, all genes and all intergenic. The number of alignments for each class were visualized with Pie-Charts whose sizes reflect the respective depth (total aligned reads) of the libraries (see **Supplementary Figure 8C**).

Specific tRFs Classes Extraction

Small RNA reads trimmed off from their adapter sequences were first aligned to the rRNA reference using the Galaxy tool [sR_bowtie (Galaxy-Version 2.1.1, owner: artbio)] and the option "Match on DNA as fast as possible." Unaligned reads were retrieved and aligned to the snoRNA reference, and snoRNA alignments were visualized using the tool [small RNA maps (Galaxy-Version 2.16.1, owner: artbio)].

Next, unaligned reads were retrieved and realigned to the non-edited tRNA reference. Matching reads in this step correspond to tRFs without CCA (tRF-non-CCA) including 5′-tRFs, tRFs-1, spanners and internal tRFs. On the contrary, edited 3′-tRFs did not match in this step, because the CCA motif is not encoded in the genome and we did not allow mismatches (see below). To retrieve these unmatched tRFs, we selected unaligned reads with 3′ end CCA and realigned these reads to the CCA-edited-tRNA reference.

Finally, we merged non-CCA tRFs and 3′ tRF using the tool [FASTA Merge Files (Galaxy-Version 1.2.0)] and realigned those reads to the CCA-edited-tRNA reference. Matched reads ("all-tRFs") were visualized (see **Figure 1B**) using the tool [small RNA maps (Galaxy-Version 2.16.1, owner: artbio)].

In order to isolate spanner tRFs, aligned non-CCA-tRFs were realigned using CCA-edited tRNAs as reference. Unaligned reads in this step are tRFs that span the editing point. These reads were realigned using non-edited-tRNA reference, allowing to retrieve spanner-tRFs maps.

Importantly, we could not reliably detect tRNA Halves (> 30 nt) since our original libraries were prepared using RNA size selection between 15 and 29 nt.

tRFs Global Size Distribution, Coverage and tRF Logo

All-tRFs, non-CCA-tRFs or 3′-tRFs datasets were used to generate small RNA maps and read size distributions taking into account the normalization factors for the different genotypes. Read coverage of tRNA sequences was generated using the tool [BamCoverage (Galaxy-Version 3.1.2.0.0, owner: bgruening)]. Briefly, we first used sR_bowtie with the options "matched on DNA, multiple mappers randomly matched at a single position," "0 mismatch allowed," and tRNA-CCA-edited as a reference. Bam alignment files from this step were used with the BamCoverage tool to generate BigWig coverage files, using the library normalization factors as scale factors. The tool [computeMatrix (Galaxy-Version 3.1.2.0.0, owner: bgruening)] was then used to prepare the data for plotting heatmaps or a profile of given regions. We used four Bed files with this tool to visualize Nuclear tRNAs, 5′-tRFs, 3′-tRFs and Mitochondrial tRNAs (see **Supplementary Figure 9A**). To obtain a Logo, tRFs FASTA files were treated to obtain the last 15 nt of every sequence then we used the tool [Sequence Logo (Galaxy-Version 3.5.0, owner: devteam)].

tRFs Expression Heatmap and Ratio Calculation

To visualize tRFs expression levels we created Heatmaps. With all-tRFs collection list, we used sR_Bowtie (for small RNA short reads Galaxy-Version 2.1.1, matched on DNA, multiple mappers, randomly matched at a single position, 0 mismatch allowed) and we used tRNA-CCA-edited as reference. Then we used the tool [Parse items in sR_Bowtie alignment (Galaxy-Version 1.0.6)]. We did a DESeq2 normalization of hit lists (geometrical method Galaxy-Version 1.0.1, see above). We cut columns from the Normalized Hit table (Galaxy-Version 1.0.2) and we used Sort data in ascending or descending order tool (Galaxy-Version 1.0.0), generating a table with the tRFs counts for the different genotypes. We used Plot Heatmap with high number of rows (Galaxy-Version 1.0.0) to create the expression profiles. We used

[7] http://www.ensembl.org/biomart/martview

Log2(value + 1) and Blue-White-Red colors to reflect reads from minimal to maximal expression. We created Heatmaps using the "tRNA-extended-CCA-edited" genome of reference to have all types of tRFs represented. This method leads to multimapping issues of several tRFs that match different tRNAs genes with similar sequences. We thus also created Heatmaps using the "Unique tRNA mature CCA-edited" genome of reference that avoids multimapping but leads to the loss of tRFs-1 originating from the precursor. To detect important changes of tRFs between genotypes, we cut columns corresponding to counts of *white*$^{-}$ and *Rpp30*$^{18.2}$ mutants, or *dTrm7_34*/TbSb* heterozygous and *dTrm7_34** homozygous mutants. We calculated the ratio of tRFs expression between them, using Compute an expression on every row tool (Galaxy-Version 1.2.0). The obtained data were treated with Microsoft Office Excel to better observe ratio differences by using conditional formatting tool, obtaining a three color code (Blue-White-Red from minimal to maximal value, see **Supplementary Figure 9B**).

snoRNAs Global Size Distribution and Coverage

To represent all the reads along a canonical snoRNA molecule we analyzed the Bam Coverage, by first using sR_Bowtie for small RNA short reads (Galaxy-Version 2.1.1), matched on DNA, multiple mappers, randomly matched at a single position. 0 mismatches were allowed, using snoRNA as genome reference. Then BamCoverage tool generates a coverage BigWig file from a given BAM file (Galaxy-Version 3.1.2.0.0) that we normalized using the scale factors. Afterward, Compute Matrix prepares data for plotting a heatmap or profiles of given regions (Galaxy-Version 3.1.2.0.0). We had three Bed files to plot: snoRNAs > 120 nt Bed file; snoRNAs < 120 nt Bed file; and both together (see **Supplementary Figure 9C**).

RESULTS

How to Study Different tRFs Categories

In this study we have developed user friendly and easy to share workflows using Galaxy[5] allowing to extract all major classes of tRFs (tRFs-Galaxy) (**Figure 1** and see section "MATERIALS AND METHODS"): 5′-tRFs, 3′-tRFs and inner-tRFs, corresponding to fragments derived from mature tRNA transcripts; tRFs-1, formed by RNase-Z cleavage of tRNA precursors; spanner tRFs, spanning the CCA region and created before CCA addition; and transcription associated tRFs (taRFs), formed due to problems in transcription termination. The presented pipeline allows to study them separately or altogether.

tRFs Description in *Drosophila* Ovaries

To describe tRFs general populations in *wild type* ovaries from young flies, we first performed a cascade of annotations of small RNA populations, to the exclusion of rRNA fragments which were previously depleted from the sequence datasets (**Figure 2A** and **Supplementary Figure 8**) (rRNA were "bioinformatically depleted"). A high percentage of small RNA reads correspond to transposable elements (TEs, yellow), representing piRNAs and/or siRNA that match TE sequences. To distinguish tRFs carrying a 3′CCA motif from non-CCA-tRFs (5′-tRFs, i-tRFs, spanners, taRFs and tRFs-1) we used two different reference genome files (see below). In *white*$^{-}$ ovaries there are twice as much non-CCA-tRFs than 3′-tRFs (**Figure 2A**: 1.15% *vs* 0.52%). However, since some sequences can be matched by multiple types of small ncRNAs, the mapping order in the cascade annotation tool can introduce a bias, as observed in the MINTmap tRFs study of Loher et al. (2017b). Thus, we used different tools to study tRFs populations in detail.

To have a general overview of tRNA fragments, we aligned all tRFs along canonical nuclear or mitochondrial tRNA precursors, belonging to 290 different nuclear tRNAs and 21 different mitochondrial tRNAs (**Figure 2B**). Nuclear tRFs coverage shows that in *white*$^{-}$ control ovaries there is a majority of tRFs-1. In addition, we observe a significant population of 3′-tRFs and a minor population of 5′-tRFs and inner tRFs. Mitochondrial tRFs seem more abundant at the 5′ part of tRNAs molecules and around the anticodon region. In addition, global size distribution analysis showed that in control ovaries, non-CCA-tRFs are heterogeneous, ranging from 15 to 25 nt, whereas 3′-tRFs are mostly 17 nt long (**Figure 2C**). The presence of a CCA signature could be easily identified by analyzing the Logo of the last 15 nt of tRFs populations (**Figure 2D**).

We next interrogated which type of tRNAs molecules could generate these tRFs. In *Drosophila melanogaster* there are 21 mitochondrial tRNAs (one per amino-acid) and 290 nuclear tRNAs, comprising several tRNAs per isotype with different anticodon sequences (between 5 and 22 tRNAs per amino acid)[1]. For example, there are 15 tRNAs for Valine with different anticodons: 6 tRNA:Val-AAC, 7 tRNA:Val-CAC and 2 tRNA:Val-TAC. Among tRNA genes, 16 tRNAs carry an intron (tRNA:Leu-CAA, Ile-TAT and Tyr-GTA). Since tRNA genes are redundant, the physiological importance of expression levels variations of individual tRNA genes is not well understood. However, it has been recently shown that differential tRNA gene expression results in changes in the abundance of tRFs but not of mature tRNAs, suggesting that different expression levels of tRNA genes may regulate non-canonical tRNA functions through tRFs (Torres et al., 2019).

Moreover, it has been shown in some organisms that small tRFs sequences could originate from genome regions similar to tRNAs, which are not true tRNA genes. These regions can be tRNA-lookalikes, truncated tRNA genes or repeated elements and they form the "non-tRNA-space" (Telonis et al., 2016; Loher et al., 2017b) (**Supplementary Figure 1A**). Thus, it is difficult to know the genomic origin of tRFs: if they belong to the "tRNA-space" or to the "non-tRNA-space." Indeed, in *white*-control *Drosophila* ovaries we observe a fraction of 15–17 nt long tRFs matching to the non-tRNA space (**Supplementary Figure 1B**). This proportion increases in *Rpp30*$^{18.2}$ mutants (**Supplementary Figure 1C**). Interestingly, if we run the same analysis excluding smallest tRFs (15–16 nt) profiles are similar in control (*w-)* while 5′tRFs accumulation in *Rpp30* mutants is less dramatic (**Supplementary Figure 1E**). Another problem in determining the origin of tRFs is that several tRNAs from

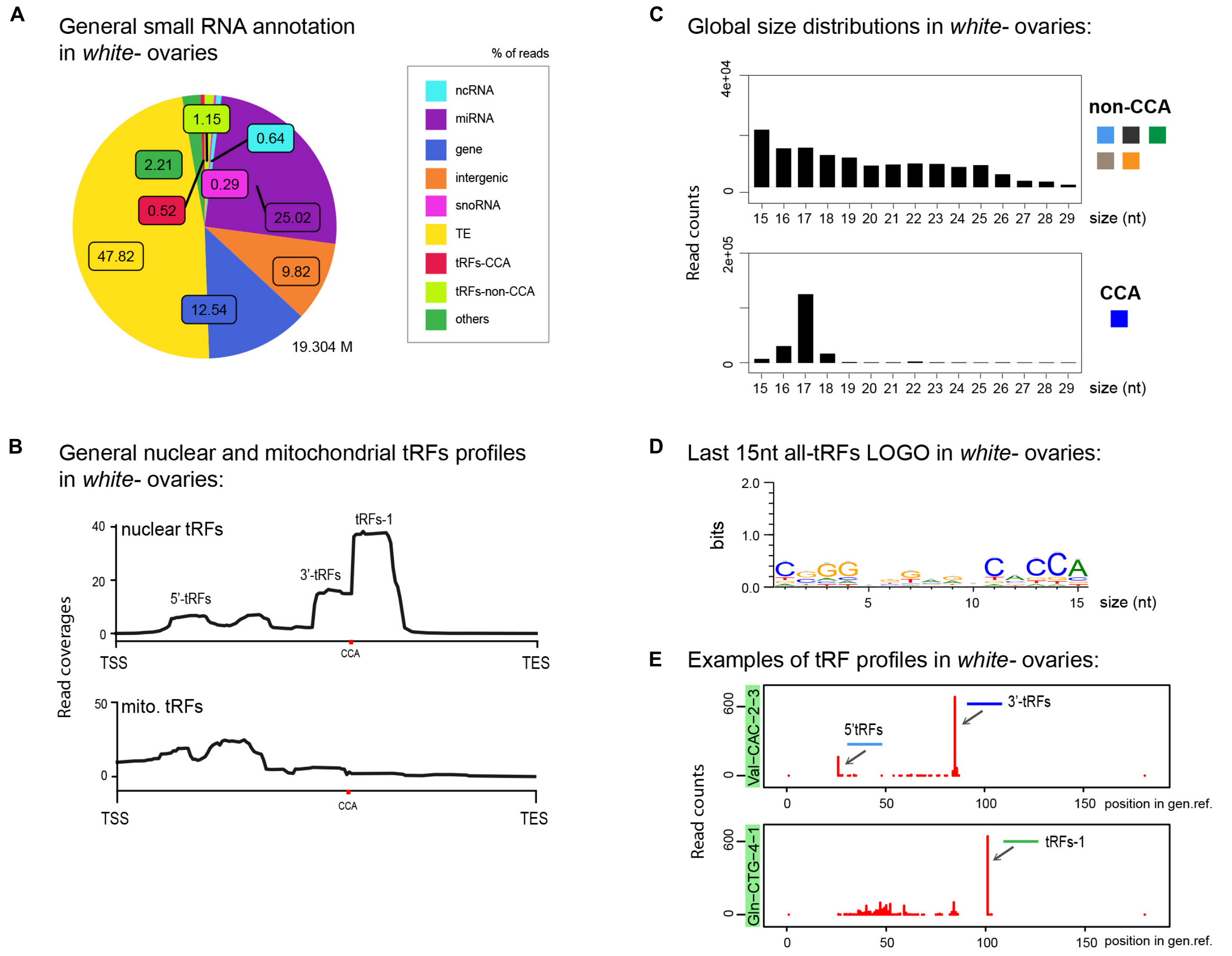


FIGURE 2 | tRFs description in control *Drosophila* ovaries: **(A)** Small RNAs sequences from 15–29 nt were analyzed to distinguish different categories with the help of an annotation cascade tool in the following order: miRNA, ncRNA, intergenic, genes, TE (piRNA, siRNA), snoRNAs, tRFs-non-CCA or tRFs-CCA. The percentage of reads is shown in a pie-chart, which size reflects the bank's depth (M = Millions of reads). **(B)** Nuclear and mitochondrial tRFs coverages of 15–29 nt tRFs were analyzed in *white*- control ovaries using scaling factors (see section "MATERIALS AND METHODS"). CCA edition point is shown with a red dot. The different types of tRFs are shown along the coverage profile from the beginning of the pre-tRNA molecule (TSS transcription start site) to the end of the extended edited genome reference (TES, transcription extended site). **(C)** General size distribution (15–29 nt) of normalized read counts corresponding to different categories of tRFs in *white*- control ovaries. Color-codes on the right are the same as in **Figure 1B** for tRFs categories. **(D)** Logo for the last 15 nt of *white*- tRFs sequences (all categories included, issued from *fasta* files). **(E)** Examples of tRFs readmap profiles in *white*- control ovaries originating from two different tRNAs. Red peaks reflect read counts (using scaling factors). The position of the peak along the edited tRNA reference genome reflects the beginning of the reads sequences. 0: beginning of the pre-tRNA. 100: position of RNase Z cleavage. 5′-tRFs are in light blue, 3′-tRFs are in dark blue, tRFs-1 are in green.

the same amino-acid share the same sequence at different parts of the molecule[1] (see alignments). Thus, sometimes we cannot distinguish if 5′-tRFs, 3′-tRFs or i-tRFs derived from a single or several tRNA molecules.

To analyze the expression of tRFs and have an idea of tRNA type forming tRFs, we made a tRNA heatmap reflecting the expression levels of all tRFs (all types comprised) belonging to a given tRNA isotype (**Supplementary Figure 2**) by using two different reference genomes: the "unique tRNA mature CCA-edited" (**Supplementary Figures 2A,B**) and the "tRNA extended CCA-edited" (**Supplementary Figure 2C**). By using the collapsed "unique tRNA mature sequences," tRFs-1 originating from tRNAs precursors cannot be studied, neither taRFs or spanner tRFs.

In *white*⁻ control ovaries, among the most abundant tRFs originating from mature tRNA sequences we could observe: tRNA:Phe-GAA, tRNA:Val-AAC or TAC, tRNA:Lys-CTT, tRNA-Gly-GCC, tRNA:Pro-AGG or CGG, tRNA:His-GTG, and tRNA:Glu-CTC (**Supplementary Figure 2A**). If we study all types of tRFs by using the "tRNA extended CCA-edited" sequences we observe that tRFs from tRNA:Val-TAC or AAC were the most abundant, followed by tRFs mapping tRNA:Glu-CTC, several tRNA:Phe-GAA, and tRNA:Pro-CGG or AGG. tRFs corresponding to mature tRNA:Val-CAC, tRNA:Ala-TGC, tRNA:Lys-TTT or tRNA:Gln-CTG were also abundant. It is important to note that, as mentioned, tRNA modifications can induce sequencing bias allowing preferential sequencing to certain tRNA and tRFs types over others, since some tRNA (and potentially also tRFs) modification patterns are isoacceptor specific. Thus, the biological meaning of this tRFs abundance pattern remains to be explored.

To describe in more detail the most relevant tRFs profiles corresponding to each individual tRNAs, we developed a multidimensional tRFs map which displays the name of the tRNA molecule, the read counts and the tRFs position along the tRNA molecule (**Figure 2E**). For example, in control ovaries, highly expressed tRFs from tRNA:Val-CAC-2-3 produce mostly 3′-tRFs (dark blue) and 5′-tRFs to a lesser extent (light blue). Moreover, tRNA:Gln-CTG-4-1, a tRNA which generates high amounts of tRFs in control ovaries, has a clear majority of tRFs-1 (green).

In conclusion, our analysis describes in detail the population of tRFs present in control *Drosophila* ovaries in a global manner (annotation, coverage, size distribution, logo and heatmap tools), as well as the specific tRFs profiles of each tRNA isotype (multidimensional tRFs maps). We find that tRFs-1 are highly present, followed by 3′-tRFs and 5′-tRFs.

tRNA Processing Defects Lead to tRFs Accumulation

We recently discovered that *Drosophila Rpp30* mutations lead to tRNA processing and early oogenesis arrest, producing atrophied small ovaries full of early arrested stages (Molla-Herman et al., 2015). As control, we chose *white*- young (freshly hatched) ovaries described above, since they are full of early stages. Besides, we observed that *Rpp30* mutants have a defect in piRNA production. In accordance, cascade annotations showed that *Rpp30*$^{18.2}$ homozygous ovaries have highly decreased TE-matching sequences compared to *white*$^{-}$ (**Figure 3A**), which is rescued in *Rpp30*$^{18.2;}$ *ubiRpp30GFP* ovaries, showing the specificity of the phenotype. Intriguingly, we observed a substantial increase of small RNAs derived from snoRNA (pink, 6.42% in the ovaries from *Rpp30*$^{18.2}$ homozygous flies compared to 0.29%, observed in *white*$^{-}$ controls). Moreover, we found that in *Rpp30*$^{18.2}$ homozygous ovaries, both non-CCA and CCA-tRFs were present in equal quantities (1.99 *vs* 2.09%), whereas in control ovaries non-CCA-tRFs were more represented than CCA-tRFs (1.15% *vs* 0.52%). This suggests an increase of CCA-tRFs and/or a decrease of some non-CCA-tRFs in *Rpp30*$^{18.2}$ homozygous mutants.

Nuclear tRFs coverage (**Figure 3B**, left panel) showed that in *Rpp30*$^{18.2}$ homozygous ovaries, there is a substantial increase of 5′-tRFs, i-tRFs and 3′-tRFs), and a drastic decrease of tRFs-1, when compared to control. Importantly, rescued *Rpp30*$^{18.2;}$ *Rpp30GFP* ovaries (purple line) showed a similar profile to *white*$^{-}$, demonstrating that Rpp30 overexpression is able to recover tRFs formation in *Rpp30*$^{18.2}$ homozygous mutants. In parallel, mitochondrial-tRFs coverages (**Figure 3B**, right) showed that *Rpp30*$^{18.2}$ homozygous individuals have a high accumulation of different tRFs types in their ovaries.

Next, global size distribution (**Figure 3C**) indicated that tRFs accumulate in *Rpp30*$^{18.2}$ homozygous ovaries compared to *white*$^{-}$. Indeed, non-CCA-tRFs range from 15 to 22 nt whereas 3′-tRFs are on average 17 nt long in mutants (**Figures 3C,D**). Finally, spanner-tRFs, which are a very minor population in *Drosophila white-* ovaries, are heterogeneous in size and do not show important changes in mutants when compared to control (**Figure 3C**, lower panels).

In conclusion, our analysis shows that tRNA processing defects alter tRFs biogenesis and/or stability in *Rpp30* mutants: increase of (5′-tRFs, i-tRFs and 3′tRFs), and tRFs-1 decrease. Since there are more than 300 tRNAs genes in *Drosophila*, we wondered if these defects were due to tRFs originating from a particular tRNA type.

tRFs Expression Levels Are Altered in *Rpp30* Mutants

As mentioned, tRFs heatmaps showed that *white*$^{-}$ control ovaries have abundant tRFs derived from tRNA-Val, Glu, Phe, Pro, Ala, Lys, Gln (**Supplementary Figure 2**). Importantly, the general heatmap profile is highly changed in *Rpp30*$^{18.2}$ homozygous but is partially rescued in *Rpp30*$^{18.2;}$ *ubiRpp30GFP* (**Supplementary Figure 2**). To easily detect the most drastic changes in tRFs populations we calculated the ratio of tRF-counts between *Rpp30*$^{18.2}$ homozygous and *white-* ovaries (**Supplementary Figure 2B**). For example, tRFs derived from tRNA:Val-AAC-2-1 are highly decreased in *Rpp30*$^{18.2}$ homozygous ovaries compared to *white-*, with a ratio of 0.05 (**Supplementary Figure 2B**).

From this ratio data, we selected tRNA profiles in which tRFs were increased, decreased or unchanged in mutants when compared to *white-* (**Figure 4**). For example, in *Rpp30*$^{18.2}$ mutants: tRNA:Leu−TAA−1−1, tRNA:Thr−AGT−1−6, tRNA:Ser−GCT−2−1, tRNA:Gly−TCC−1−2 and tRNA:Pro−AGG−1−5 show an increase of 3′-tRFs. In addition, tRNA:Ala−CGC−1−1 accumulates 3′−tRFs and 5′−tRFs. tRNA:Ser−AGA−2−2 shows a drastic increase in only 5′−tRFs. Indeed, all tRNA:Ser−AGA/CGA (12 different tRNAs) behave similarly. tRNA:Leu−CAA−2−2 has an important increase in 5′−tRFs as all tRNA:Leu−CAA. It should be noted that Leu-CAA group have an intron of 40–44 nt, that is why 3′-tRFs are located offset in tRFs maps. Next, tRNA-Gly-GCC-2-1 is similar in *white-* and *Rpp30*$^{18.2}$ mutants. Finally, several tRFs types decreased in *Rpp30*$^{18.2}$ mutants: tRFs-1 generated from tRNA:Glu-CTC-3-8 and tRNA:Gln-CTG-4-1; 5′-tRFs generated from tRNA:Val-CAC-2-3; 3′-tRFs generated from tRNA:Val-CAC-2-2 and 2-3. We also compared tRFs profiles by selecting tRNAs having the mostly expressed tRFs (up to heatmaps) in *white-* and we compared them to mutants (**Supplementary Figure 3**).

Overall, we find that in *Drosophila* ovaries, tRFs originate from diverse isotypes of tRNAs and show heterogeneous profiles. In general, as shown in **Figure 3B**, we find that tRFs-1 are decreased in *Rpp30* mutants, whereas tRFs originating from mature tRNA are accumulated. tRNA processing by RNase P is the first step of tRNA biogenesis following transcription. We thus wondered whether other downstream events could also affect

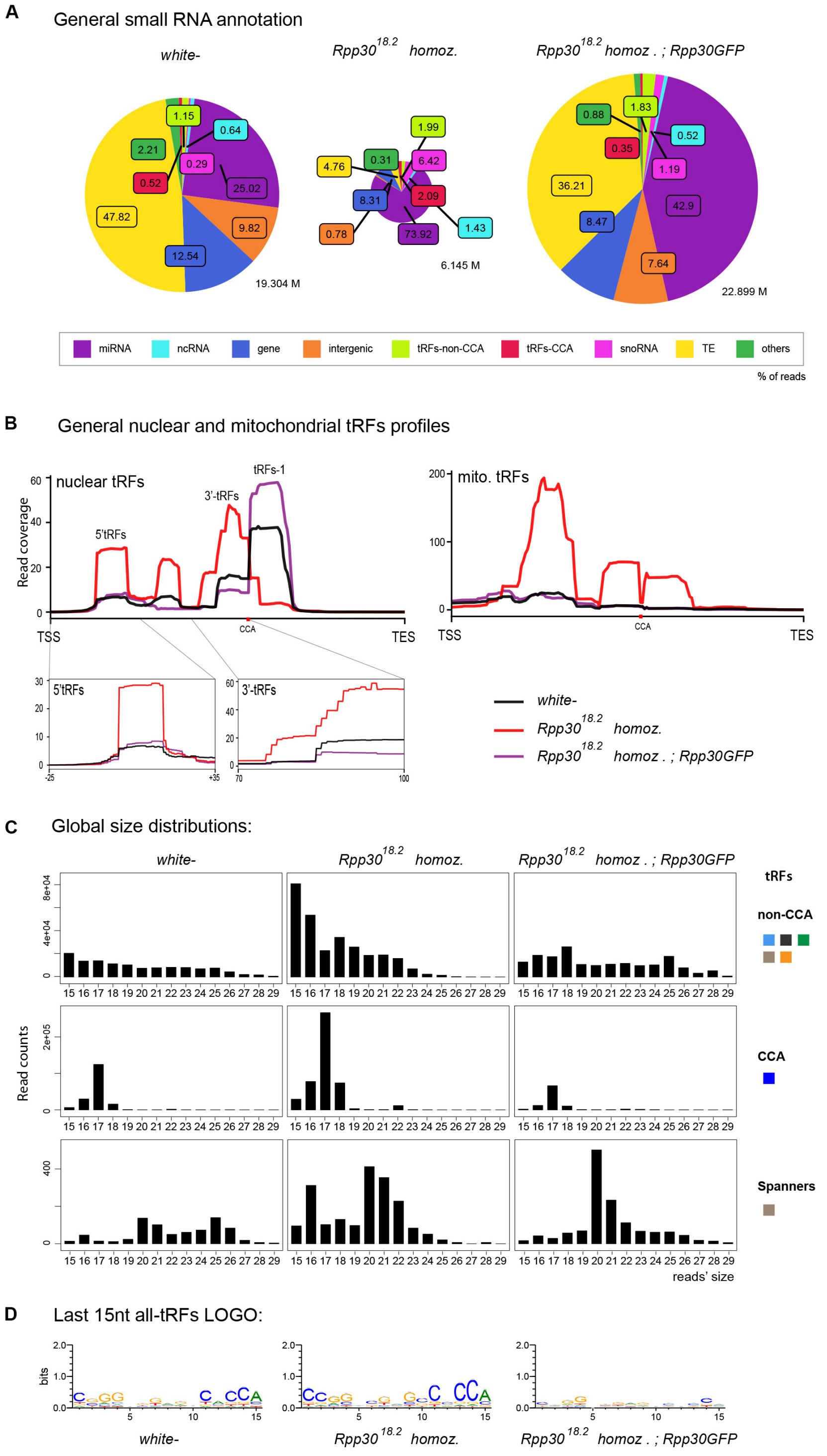


FIGURE 3 | tRNA processing plays a role in nuclear and mitochondrial tRFs formation: **(A)** Small RNAs sequences (from 15–29 nt) were analyzed in different genotypes to distinguish categories with the help of an annotation cascade tool in the following order: miRNA, ncRNA, intergenic, genes, TE (piRNA, siRNA), snoRNAs, tRFs-non-CCA or tRFs-CCA. The percentage of reads for each genotype is shown in pie-charts, which size reflects the depth of each bank (M = Millions of reads). **(B)** Nuclear and mitochondrial tRFs coverages were analyzed in *white-* control and *Rpp30* mutant ovaries using scaling factors (see section "MATERIALS AND METHODS"). Different tRFs are shown along the coverage profile from the beginning of the pre-tRNA molecule (TSS transcription start site) to the end of the extended edited reference genome (TES, transcription extended site). CCA edition point is shown with a red dot. 5′-tRFs and 3′-tRFs regions are zoomed in, for a better comparison between the genotypes. **(C)** General size distribution (15–29 nt) of normalized read counts corresponding to the different categories of tRFs in *white-* control and mutant ovaries. Color-codes on the right are the same as in **Figure 1B** for tRFs categories. **(D)** Logo for the last 15 nt tRFs sequences of *white-* control and mutant ovaries (all categories included, issued from *fasta* files containing all tRFs sequences).

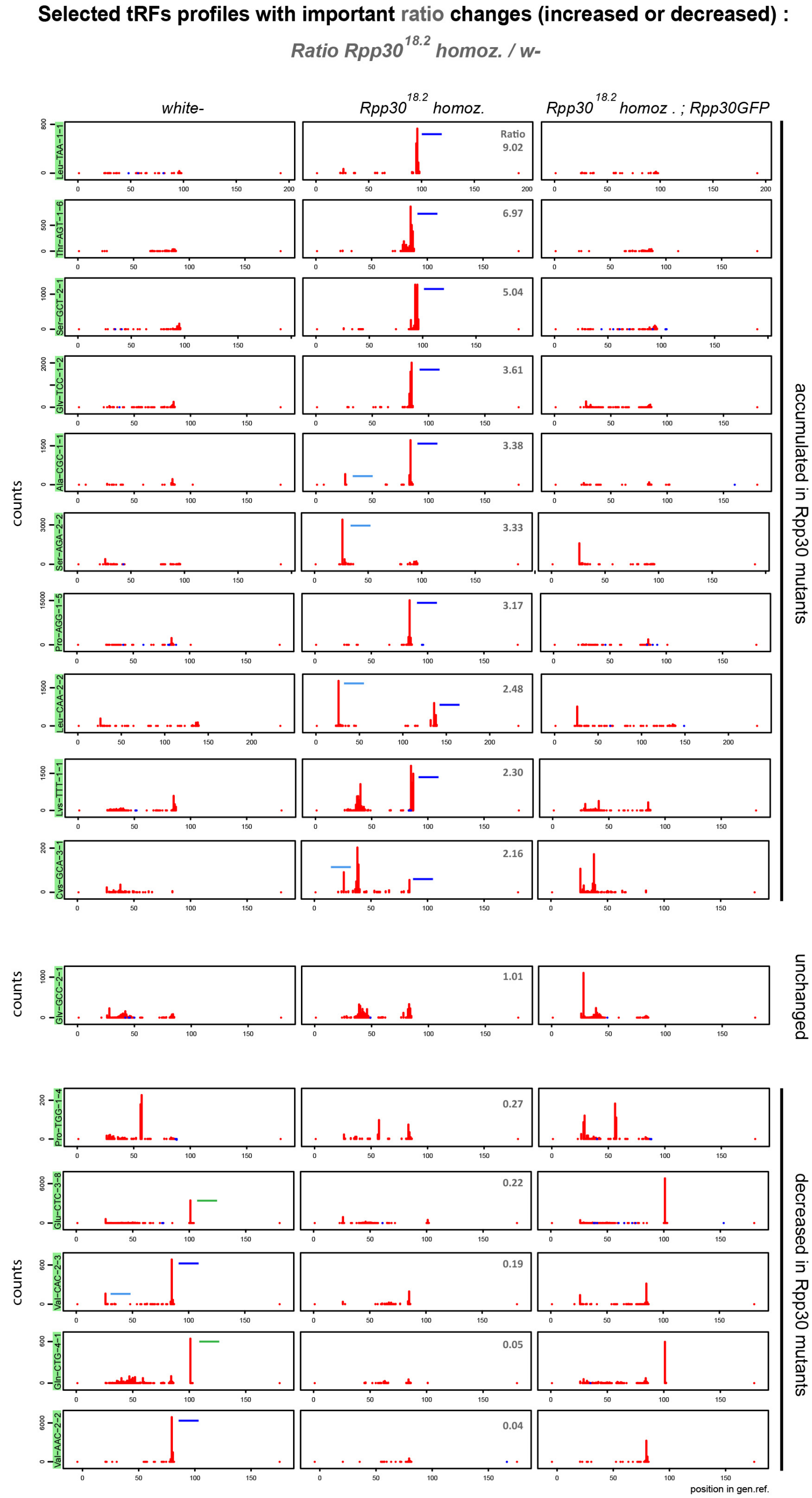


FIGURE 4 | Rpp30 mutation leads to an increase of 5′-tRFs, an increase of 3′-tRFs and a decrease of tRFs-1. 16 tRFs readmap profiles as examples of the most increased or decreased tRFs from the ratio *Rpp30*$^{18.2}$ *homoz./white-* (see in **Supplementary Figure 2B**) are shown for the different genotypes, using normalizing factors (see section "MATERIALS AND METHODS"). Since pre-tRNAs sequences are included in the tRNA reference genome, 5′-tRFs start at position 25 nt instead of position 0 nt. 3′-tRFs are located around the position 80 nt and tRFs-1 are located around position 100 nt (positions can vary depending on tRNA lengths and the presence of intron). Peaks determine the beginning of the reads sequences. tRFs are schematized in *white-* and *Rpp30*$^{18.2}$ homozygous for better comparison: 5′-tRFs in light blue, 3′-tRFs in dark blue and tRFs-1 in green. Ratio's values above 1 (upper pannels): tRFs increased in *Rpp30*$^{18.2}$ mutants. Ratio's values below 1 (lower panels): tRFs decreased in *Rpp30*$^{18.2}$ mutants.

tRFs biogenesis or stability, such as tRNA post-transcriptional modifications of tRNA molecules.

tRNA 2′-O-Methylation Defects Lead to a Decrease of tRFs-1 and an Increase of 3′-tRFs

Mutations of tRNAs 2′-O-methyltransferases (Nm MTases) *dTrm7_34* and *dTrm7_32* lead to *Drosophila* life span reduction, small RNA pathways dysfunction, increased sensitivity to RNA virus infections and tRFs-Phe accumulation (Angelova et al., 2020). In our cascade annotation analysis (non-normalized), non-CCA-tRFs decrease in *dTrm7_34** homozygous mutants when compared to control (**Figure 5A**, light green, 1.19% versus 0.52%), whereas tRFs-CCA slightly increase (**Figure 5A**, red, 0.09% versus 0.13%). Surprisingly, double mutants *dTrm7_34*, dTrm7_32** show profiles similar to control. By using normalization factors, the analysis of tRFs size distribution and of a Logo sequence revealed that 3′-tRFs of 18 nt increase in *dTrm7_34** homozygous mutants when compared to control (**Figure 5B**, C), which was rescued in double mutants. Finally, lowly expressed spanner tRFs were similar in *dTrm7_34** heterozygous and homozygous mutants when compared to control and slightly lower in double mutants (**Figure 5B**).

To obtain an overview of which tRFs classes were globally altered in Nm MTases mutants, we aligned all-tRFs together along a canonical nuclear or mitochondrial tRNA precursors. In heterozygous control ovaries (**Figure 5D**, gray), there is a majority of nuclear tRFs-1, similarly to control *white-* ovaries (**Figure 2B**). Of note, the size of heterozygous ovaries is bigger than *white-* ovaries, since they have early and older stages. Interestingly, *dTrm7_34** homozygous mutants and double mutants *dTrm7_34*, dTrm7_32** showed a decrease of tRFs-1 when compared to heterozygous control (**Figure 5D, orange** and **blue**), suggesting that these Nm MTases are involved in tRFs biogenesis and/or stability.

Since tRFs-1 reads signal is very high and the signal of 5′-tRFs and 3′-tRFs is lower, it was difficult to identify major changes in tRFs originating from mature tRNAs. By zooming into these regions, we observed that 5′-tRFs slightly decrease in *dTrm7_34* homozygous* mutants (**Figure 5D**, left panel, orange), whereas 3′-tRFs increase (**Figure 5D**, right panel, orange). In this analysis double mutants again show similar profiles to control, suggesting that *dTrm7_32* mutation somehow rescues *dTrm7_34* defects on 3′-tRFs accumulation. Interestingly, we recently reported that longer 5′-tRF-Phe (∼35 nt) were significantly increased in different combinations of *dTrm7_34* mutant alleles (Angelova et al., 2020).

Moreover, we observed mitochondrial tRFs in heterozygous ovaries similar to *white-* flies (**Figures 2B**, 5D, right panel, gray line), derived mostly from the first half of the molecule. Homozygous mutant for *dTrm7_34** ovaries are similar to heterozygous mutants, whereas double mutants *dTrm7_34*, dTrm7_32** show a global increase of mito-tRFs, suggesting that *dTrm7_32*, and not *dTrm7_34*, may be involved in mitochondrial-tRFs biogenesis and/or stability.

In summary, we have observed that defects of tRNA 2′-O-methylation affect tRFs populations in *Drosophila* ovaries. *dTrm7_34* and *dTrm7_32* mutations lead to a decrease of tRFs-1 and *dTrm7_34* mutation leads to an accumulation of 3′-tRFs and a slight decrease of 5′-tRFs.

tRNA Methylation Mutations Affect tRFs Derived From Different Isotypes of tRNAs

tRNA expression heatmaps using "extended tRNA CCA-edited reference genome" allowing the analysis of all types of tRFs showed that the most expressed tRFs in ovaries from heterozygous *dTrm7_34** mutants were those corresponding to tRNAs Glu-CTC, Pro-CGG and AGG, Val-TAC, Cys-GCA, Lys-TTT, Gly-TCC, Ala-CGC, His-GTG, Ser-GCT (**Supplementary Figure 4A**), similarly to *white-* control ovaries (**Supplementary Figure 2C**). In the ovaries from *dTrm7_34** homozygous mutants, we observed a decrease of the tRFs derived from Glu-CTC, Cys-GCA, Lys-TTT or Gly-TCC, whereas tRFs derived from Ser-GCT were increased when compared to control. These changes have been quantified by calculating the ratio between homozygous and heterozygous *dTrm7_34** read counts (**Supplementary Figure 4B**).

Considering the read counts ratio change between homozygous and heterozygous *dTrm7_34** ovaries (**Figure 6**, upper panels), we observe that 5′-tRFs are strongly decreased for several tRNAs, such as Glu-CTC, Gly-TCC, Cys-GCA. This effect is partially rescued in double mutants. In addition, we observe that tRFs-1 from tRNAs Gln-CTG-4-1 and Pro-AGG-2-1 are strongly reduced in the ovaries from both homozygous *dTrm7_34* and dTrm7_32*, dTrm7_34** compared to the control (**Figure 6**, upper panels). We detect no change in tRNA:Met-CAT-1-5 tRFs between control and mutants, where the observed tRFs population matches the anticodon region (**Figure 6**, middle panel). On the contrary, we clearly see an increase of 3′-tRFs in *dTrm7_34** homozygous mutants for several tRNAs: Pro-CGG-1-1, Thr-AGT-1-4, Gln-CTG-1-1, Arg-TCG-2-1, Ser-GCT-2-2 (**Figure 6**, lower panels). Surprisingly, those defects are rescued in double mutants *dTrm7_32*, dTrm7_34**. In addition, we find similar results analyzing profiles corresponding to highly expressed tRFs in heterozygous *dTrm7_34** ovaries for 5′-tRFs and tRFs-1 (**Supplementary Figure 5**). However, increase of 3′-tRFs are difficult to observe, indicating that the increased in *dTrm7_34** homozygous ovaries 3′-tRFs are not highly present in heterozygous ovaries.

We recently showed that dTrm7_34 and dTrm7_32 methylate tRNA-Leu, Trp, Phe (conserved targets in yeast and humans), as well as that dTrm7_32 methylates tRNA-Glu and Gln in *Drosophila* (Angelova et al., 2020). Indeed, tRFs derived from these specific tRNAs show different profiles between mutants and control conditions (**Supplementary Figure 6**). First, tRNA-Leu tRFs have different profiles regarding their anticodon sequence. Some 5′-tRFs in control ovaries are decreased in *dTrm7_34** mutants and remain decreased or are rescued in double mutants. tRFs-1 decrease in *dTrm7_34** homozygous mutants, whereas 3′-tRFs increase. Thus, tRNA-Leu tRFs follow the general tendency

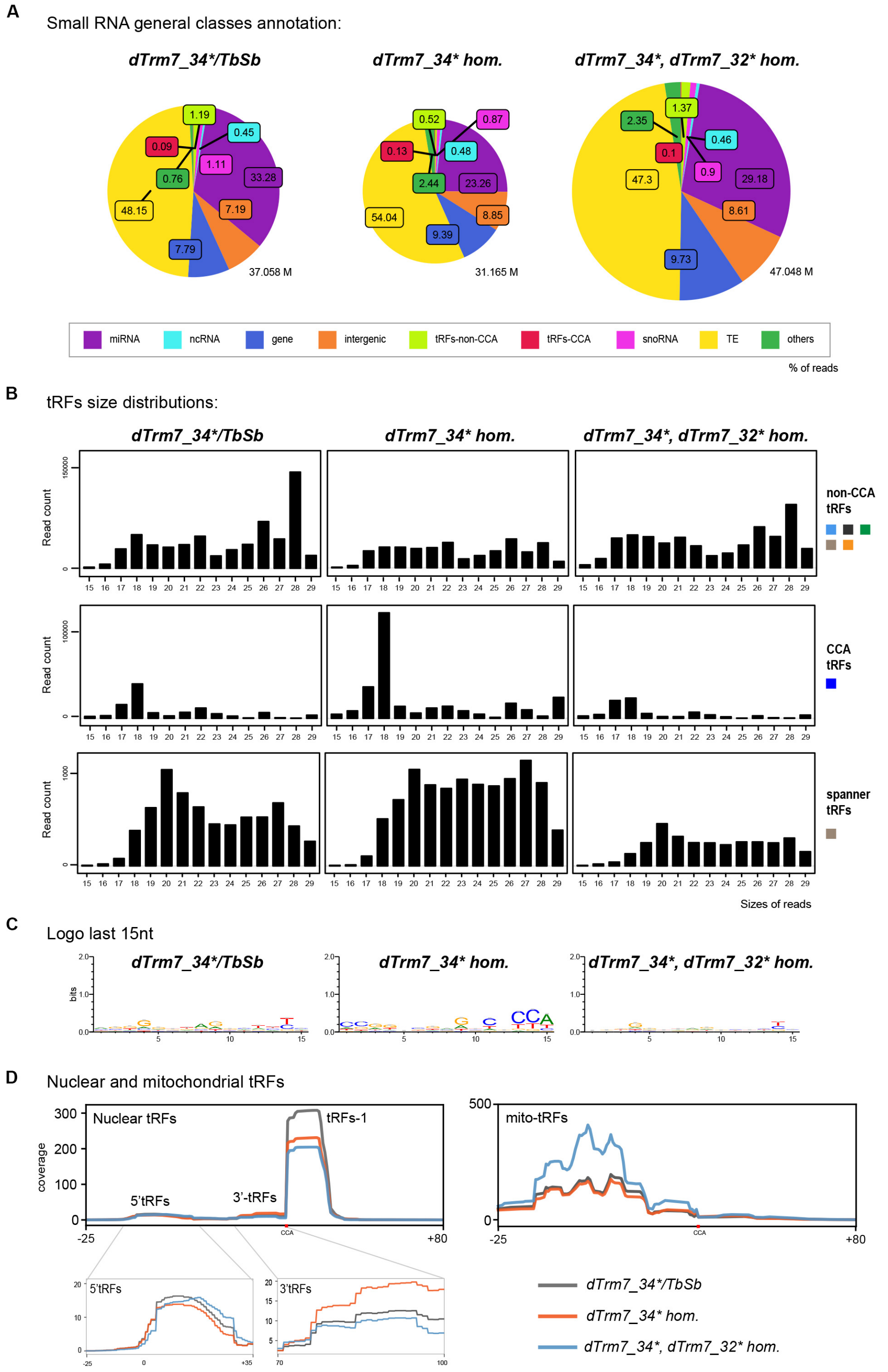


FIGURE 5 | tRNAs methylation defects alter nuclear and mitochondrial tRFs formation: **(A)** Small RNAs sequences from 15–29 nt were analyzed in different genotypes to distinguish different categories with the help of an annotation cascade tool in the following order: miRNA, ncRNA, intergenic, genes, TE (piRNA, siRNA), snoRNAs, tRFs-non-CCA or tRFs-CCA. The percentages of reads from *dTrm7_34** heterozygous, *dTrm7_34** homozygous and *dTrm7_34*, dTrm7_32** double mutant ovaries are shown in pie-charts. The pie-charts size reflects the depth of the bank (M = Millions of reads). **(B)** General size distribution (15–29 nt) of normalized read counts corresponding to the different categories of tRFs in different genotypes using scaling factors (see section "MATERIALS AND METHODS"). Color-codes for the tRFs categories on the right are described in **Figure 1B**. **(C)** Logo for the last 15 nt tRFs sequences of control and mutant ovaries (all categories included, issued from *fasta* files). **(D)** Nuclear and mitochondrial tRFs coverages were analyzed in different genotypes using scaling factors (see section "MATERIALS AND METHODS"). Different types of tRFs are shown along the coverage profile from the beginning of pre-tRNA (TSS transcription start site) to the end of the extended edited reference genome (TES, transcription extended site). CCA edition point is shown with a red dot. 5′-tRFs and 3′-tRFs regions are zoomed in for better comparison between the genotypes.

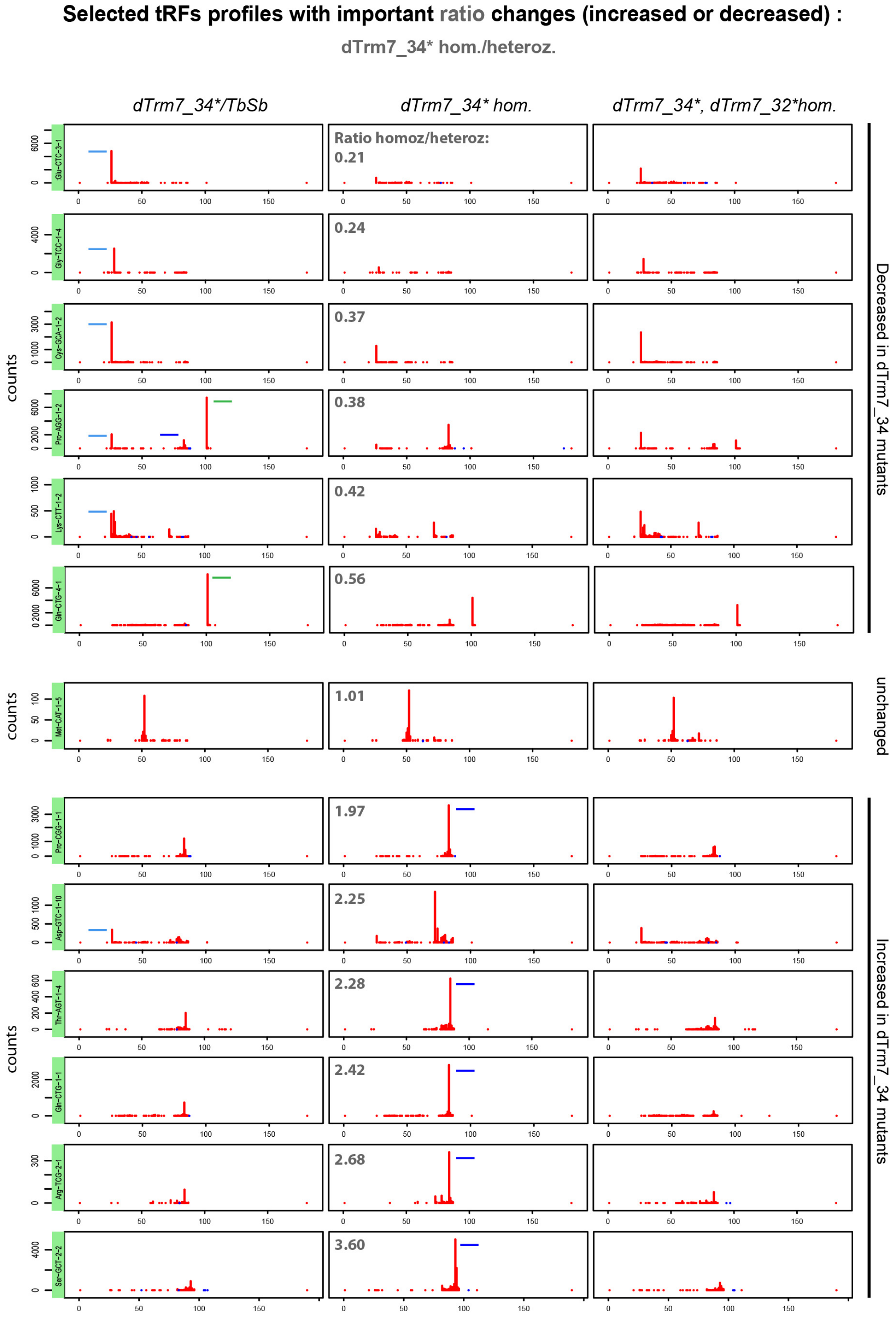


FIGURE 6 | tRFs expression is altered in tRNA methylation mutants: 13 tRFs readmap profiles as examples of the most increased or decreased tRFs from the ratio *dTrm7_34* homozygous/heterozygous* (see in **Supplementary Figure 4B**) are shown for the different genotypes, using normalizing factors (see section "MATERIALS AND METHODS"). Since pre-tRNAs sequences are included in the tRNA reference genome, 5′-tRFs start at position 25 nt instead of position 0 nt. 3′-tRFs are located around the position 80 nt and tRFs-1 are located around position 100 nt, depending on tRNA lengths and the presence of intron. Peaks determine the beginning of the reads sequences. tRFs are schematized in *dTrm7_34** homozygous and heterozygous mutants for better comparison: 5′-tRFs in light blue, 3′-tRFs in dark blue and tRFs-1 in green. Ratio's values above 1 (lower panels): tRFs increased in *dTrm7_34* homozygous* mutants. Ratio's values below 1 (upper panels): tRFs decreased in *dTrm7_34* homozygous* mutants.

observed in **Figure 5D**. In contrast, tRNA-Trp- and tRNA-Phe-derived 3′-tRFs increase in *dTrm7_34** homozygous mutants, while double mutants *dTrm7_34*, dTrm7_32** lose 3′-tRFs.

Overall, tRNA Nm methylation defects in the anticodon loop have a global impact on tRNA fragmentation, though to a lesser extent than tRNA processing defects (**Figure 7**). Indeed, tRFs-1 show a decrease in *dTrm7_34** homozygous mutants and 3′-tRFs are increased, whereas 5′-tRFs are slightly decreased. Intriguingly, double mutants *dTrm7_34*, dTrm7_32** have profiles similar to control, indicating that for at least some of the observed differentially expressed tRFs increased in *dTrm7_34** homozygotes, dTrm7_32-dependent Nm modification might have an effect on their biogenesis and/or stability. Finally, tRFs longer than 30 nt can't be properly detected in the analyzed libraries (size selection of 15–29 nt), so our analysis does not include the 35 nt long tRNA-Phe-derived 5′-tRFs characterized previously in *dTrm7_34** mutants (Angelova et al., 2020).

tRNA Processing and Methylation Avoid snoRNA Fragmentation

The increase of small RNAs derived from snoRNAs observed in *Rpp30* mutants (**Figure 3A**) led us to study this population in more detail. In *Drosophila*, snoRNAs > 120 nt are box H/ACA and play a role in pseudouridylation whereas snoRNAs < 120 nt are box C/D snoRNA and play a role in 2′-O-methylation (Huang, 2005; Angrisani et al., 2015; Falaleeva et al., 2017). Since RNase P has been shown to participate in snoRNAs maturation in some species (Coughlin et al., 2008) and since snoRNAs molecules can be cleaved to form snoRNA fragments (snoRFs) by enzymes that remains to be elucidated (Falaleeva and Stamm, 2012; Świątowy and Jagodziński, 2018), we studied a potential role of RNase P in snoRFs biogenesis (**Supplementary Figure 7**).

snoRFs size distributions shows that snoRFs are highly increased in *Rpp30*$^{18.2}$ homozygous mutants, with snoRFs mostly ranging between 15 and 23 nt (**Supplementary Figure 7A**). Since there are two main snoRNA populations (box H/ACA and C/D **Supplementary Figure 4B**), we analyzed them all together and separately to observe snoRFs coverages. Indeed, total snoRFs coverage shows that in control flies (*white-* and rescued *Rpp30*$^{18.2}$*; Rpp30GFP*) there is almost no snoRFs formation (**Supplementary Figure 7C**, upper panel, black and purple lines). However, snoRFs are highly increased in *Rpp30*$^{18.2}$ homozygous mutants (red line), mostly in 3′ of snoRNA molecules. Indeed, there is a strong accumulation of box C/D snoRFs mostly at 3′ of the snoRNA molecule, and an increase of box H/ACA 5′ and 3′ snoRFs in ovaries from *Rpp30* mutants compared to controls. The sequence specificities for box C/D or H/ACA can be observed by analyzing the Logo (**Supplementary Figure 7D**).

In methylation mutants, ovaries from heterozygous and homozygous *dTrm7_34** mutants show similar profiles, with snoRFs mostly ranging from 21-28 nt (**Supplementary Figure 7A**). Indeed, in comparison to *white-* where almost no snoRFs are detected (**Supplementary Figure 7C**), we observe that snoRFs accumulate in tRNA methylation mutant genetic backgrounds (heterozygous, homozygous and double mutants), mostly at the 3′ part of snoRNA molecules (**Supplementary Figures 7A**, C). These results suggest that dTrm7_34 and dTrm7_32 function(s) can be important in avoiding snoRFs fragmentation.

DISCUSSION

Our study presents an easy to share user friendly bioinformatic workflow for tRFs population analysis and its use on Illumina-generated small RNA libraries. As proof of principle we used libraries of control and mutant *Drosophila* for two key events of tRNA biology: tRNA processing and tRNA Nm methylation at the anticodon loop. We provide a new genome reference, comprising sequences upstream and downstream of mature tRNA genome sequences and bioinformatically added CCA tags that allow analysis of 3′-tRFs and 5′-tRFs, i-tRFs, tRFs-1, taRFs and spanners (**Figure 1A**).

Using mutant flies for the RNAse P subunit (*Rpp30*$^{18.2}$) we observed an important decrease in tRFs-1 (**Figure 7**). tRFs-1 are generated by RNase Z-mediated cleavage of pre-tRNAs. Interestingly, it has been described in *Drosophila* and other species that RNase P cleaves the 5′ trailer before RNase Z cleaves the 3′ trailer (Dubrovsky et al., 2004; Xie et al., 2013). In this way, an upstream defect on 5′ cleavage due to *Rpp30* mutation could affect RNase Z cleavage, thereby explaining why tRFs-1 decrease in *Rpp30* mutants. Moreover, *Rpp30*$^{18.2}$ mutants show an accumulation of 5′-tRFs. It is possible that a lack of 5′ leader cleavage affects tRNA secondary structure, promoting cleavage in the D-loop to form 5′-tRFs by Dicer or other endonucleases as already shown in mammals (Li et al., 2018). Finally, 3′-tRFs also increase in *Rpp30*$^{18.2}$ mutants. CCA is known to be added on mature tRNA, which suggests that Rpp30 mutation somehow affects tRNA cleavage after the CCA tRNA editing. Since 3′-tRFs are involved in TEs silencing control, increasing this tRFs population by promoting tRNA cleavage at the T-loop can be a way to control TEs when the main piRNA pathway is compromised. This observation is consistent with previous reports of tRFs functioning as a versatile and adaptive source for genome integrity protection (Martinez et al., 2017; Schorn et al., 2017). Also, it is important to mention that *Rpp30* mutants accumulate short tRFs (15–17 nt) which origin is difficult to know: tRNA-space versus non-tRNA space. Indeed, when 15–16 nt are excluded from the analysis, while control profiles do not change, the tRFs increase is less dramatic (**Supplementary Figure 1E**), suggesting that they could partly correspond to non-tRNA space. For example, they could originate from TEs overexpression and fragmentation, but this remains to be elucidated.

Besides tRFs, we observed that snoRNAs fragments (snoRFs) accumulate in *Rpp30* homozygous mutant ovaries. In this sense, it has been shown that snoRNAs can be a target of RNase P in some species during snoRNA maturation (Coughlin et al., 2008; Marvin et al., 2011). We know now that snoRNAs molecules can be cleaved into snoRNA fragments (snoRFs) but the enzyme(s) responsible for their cleavage remain(s) poorly characterized (Świątowy and Jagodziński, 2018). snoRFs are aberrantly present in several pathologies such as cancer and neurodegenerative

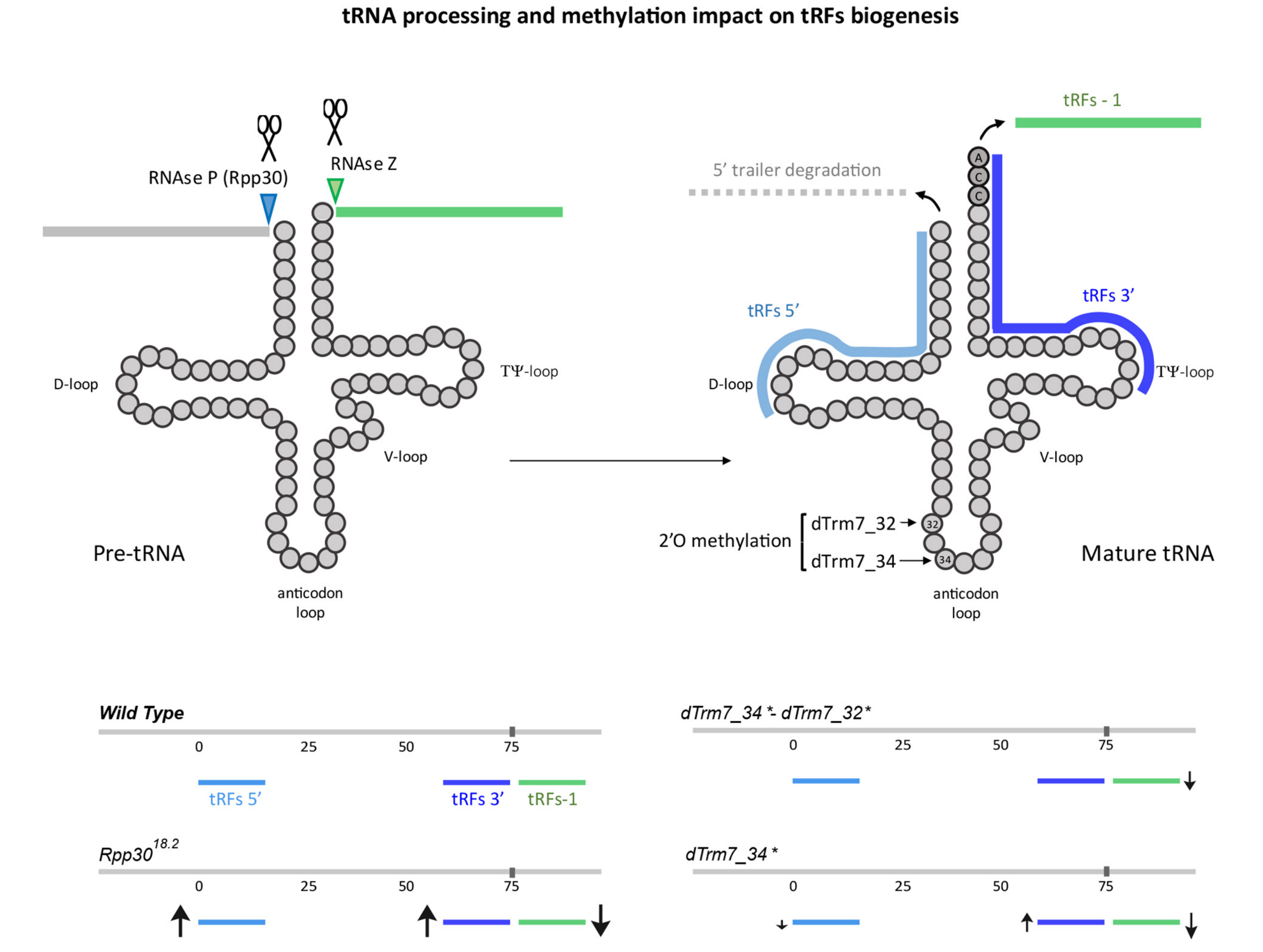


FIGURE 7 | tRNA processing and methylation defects impact on tRFs biogenesis: The main steps of tRNA processing are depicted. Cleavage sites for ribozymes RNase P and Z are indicated on a pre-tRNA molecule. Cleavage of the 3′ trailer forms tRFs-1 (green). Upon cleavage of the leader and trailer sequences and CCA addition (dark gray), yielding mature tRNAs, they can be cleaved at the D-loop, forming the tRFs 5′ (light blue) and at the T-loop, forming 3′-tRFs. 2′O-methylation sites for dTrm7_32 and dTrm7_34 are shown at the anticodon loop. Increase or decrease of different tRFs populations in mutants for tRNA processing or tRNA methylation are schematized with arrows of different sizes (↑, increased, ↓, decreased).

diseases (Falaleeva and Stamm, 2012; Patterson et al., 2017; Romano et al., 2017; Światowy and Jagodziński, 2018). It is thus possible that RNase P limits snoRNA fragmentation to preserve homeostasis by an uncharacterized mechanism. Interestingly, mice mutant for RNase Z (the other major tRNA processing enzyme) showed an increase in snoRNAs expression. This phenomenon was proposed to compensate translation defects produced by the lack of correct 3′ tRNA processing (Siira et al., 2018). However, a role of RNase Z in snoRFs formation has not been described.

As introduced previously, in *Drosophila* some methylation marks protect tRNAs from cleavage and aberrant tRFs populations accumulate in mutants for different methyltransferases, such as Dnmt2 (catalyzes m^{5}C methylation) and dTrm7_32 and dTrm7_34 (catalyze 2′-O-methylation) (Schaefer et al., 2010; Durdevic et al., 2013b; Genenncher et al., 2018; Angelova et al., 2020). In addition, It has recently been shown in mice that loss of NSUN2 altered tRFs profiles in response to stress, impairing protein synthesis (Gkatza et al., 2019). Our analysis of the tRFs populations in ovaries mutant for dTrm7_34 and dTrm7_32, two Nm MTases of the anticodon loop of some tRNAs, showed that *dtrm7_34* mutants have different tRFs profiles when compared to *Rpp30*$^{18.2}$ mutants (**Figure 7**): tRFs-1 are decreased compared to control, 5′-tRFs are slightly decreased, whereas 3′-tRFs are increased. dTrm7_34 has been shown to methylate tRNAs at the wobble position 34 of the anticodon region and its mutation leads to an accumulation of tRNA halves fragments (around 35 nt length) (Angelova et al., 2020). Thus, an accumulation of longer tRFs could impede a cleavage in the D-loop, explaining a decrease in 5′-tRFs. However, in this study we cannot detect tRNA halves since our datasets contain RNAs of 15–29 nt only. Moreover, 3′-tRFs increase in *dTrm7_34** homozygous mutants, suggesting that tRNA Nm methylation at position 34 somehow limits

T-loop cleavage. The other anticodon Nm methyltransferase, dTrm7_32, has been shown to methylate position 32 of its substrate tRNAs (Angelova et al., 2020). Interestingly, double mutants *dTrm7_34*, dTrm7_32** show a different tRFs profile when compared to *dTrm7_34** single mutant. This result suggests that a lack of methylation in the anticodon loop region can somehow favorize the production and/or stabilize some tRFs, as proposed recently for tRNA halves in Nm mutants (Angelova et al., 2020).

Finally, our study detected an increase of mitochondrial derived tRFs in $Rpp30^{18.2}$ mutants, as well as in double mutants *dTrm7_34*, dTrm7_32** when compared to control, whereas control conditions show very low levels of mito-tRFs. This observation indicates that tRNA processing and tRNA Nm methylation pathways of the anticodon loop limit aberrant fragmentation of mitochondrial tRNAs. Mito-tRNAs are polycistronic sequences cleaved by conserved mitochondrial RNase P and Z complexes in several species (Jarrous and Gopalan, 2010; Rossmanith, 2012). Intriguingly, a recent study reported an interplay between RNase P complex and mito-tRNA methylation enzymes in human cells. Indeed, mito-RNAse P was shown to recognize, cleave and methylate some mitochondrial tRNAs *in vitro*, and its activity was enhanced by interaction with a tRNA methylation cofactor (Karasik et al., 2019). Mito-RNAse P and Z dysfunctions have also been linked to several human mitochondrial diseases, as myopathies and neurodevelopmental disorders (Barchiesi and Vascotto, 2019; Saoura et al., 2019). A description of mitochondrial tRFs biogenesis could thus help to better understand the molecular mechanisms underlying these pathologies. In line with neurodegenerative diseases implication, tRFs have been shown to be present in the brain of different species, and their populations were shown to vary during aging in *Drosophila* (Karaiskos et al., 2015; Karaiskos and Grigoriev, 2016; Angelova et al., 2018).

High throughput Illumina sequencing of small RNA libraries could introduce biases in tRFs detection, since tRNAs are highly modified molecules and very few techniques are able to properly describe these modifications in a transcriptome-wide way, such as ARM-seq or Circ-RNA-seq tRNA (Cozen et al., 2015; Zhang et al., 2015). For example, in *white-* control ovaries, tRFs-1 are the most highly present, followed by 3′-tRFs and 5′-tRFs. This tRF distribution in the sample could be due to the method of library preparation or sequencing, since with standard small RNA-Seq protocols, tRFs-1 could be preferentially sequenced as they are poorly modified post-transcriptionally. In addition, since the reverse transcription occurs from the 3′-end of the tRNA sequence, because of tRNA modifications libraries could be biased toward detection of reads mapping to the 3′-end of tRNA sequences (Torres et al., 2019). However, some studies have reported that tRNA modifications only have a limited impact on data mining when studying tRFs in The Cancer Genome Atlas (Telonis et al., 2019). Importantly, a huge number of datasets are already available with valuable information to extract. By analyzing different mutants from distinct pathways we should be able to increase our knowledge on tRFs biogenesis and/or stability, as well as on the potential interactions between the diverse mechanisms impacting tRFs biology. For example, it has been recently shown that snoRNAs can 2′-O-methylate tRNA-CAT at position 34 in mammalian cells, similarly to dTrm7_34 (Vitali and Kiss, 2019; Angelova et al., 2020). Conversely, tRNA methylation could have an impact on snoRNAs biogenesis, as observed in this study. Thus, our new workflow can help to analyze past, present and future small RNA sequences obtained by different means. It will be interesting to obtain a tRF cartography in different tissues, organs and species; to determine tRFs targets and biogenesis factors; as well as to elucidate tRFs functions in gene expression regulation. It will also be interesting to compare datasets obtained from classical Illumina sequencing with other techniques such as ARM-seq, which provides a read out of some modifications and may reveal additional tRFs populations. Our study thus has the potential to participate in the discovery of novel nuclear or mito-tRFs that could help advance the understanding of the etiology of a wide range of human pathologies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the European Nucleotide Archive (ENA) http://www.ebi.ac.uk/ena, accession numbers: PRJEB10569 (Rpp30 mutants), PRJEB35301 and PRJEB35713 (Nm mutants).

AUTHOR CONTRIBUTIONS

AM-H and MA performed experiments. AM-H and CA designed and performed bioinformatic data analyses. AM-H wrote the manuscript. J-RH participated in data analysis and manuscript writing. MA, CA, and CC participated in manuscript writing. MG did the comparative table of tRFs analysis methods. All authors contributed to the article and approved the submitted version.

FUNDING

J-RH lab was supported by CNRS, Inserm, Collège de France, FRM (Equipe FRM DEQ20160334884), ANR (ANR-13-BSV2-0007-02 PlasTiSiPi; ANR-15-CE13-0001-01, AbsCyStem), and Bettencourt-Schueller Foundations. CC received financial support from CNRS and Sorbonne Université, Fondation Maladies Rares (FMR), IBPS AI 2018-2020, and the COST action EPITRAN CA16120. CC is a member of the COST action EPITRAN CA16120. CA received financial support from the ANR (ANR-13-BSV2-0007-02 PlasTiSiPi), the Institut Français de Bioinformatique (IFB BioInfo Tools 2017), and the IBPS (AI 2018-2020). MG was payed by Ecole Normale Supérieure (ENS), PSL Research University.

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Horizon 2020
European Union Funding
for Research & Innovation

ACKNOWLEDGMENTS

We thank Ana Maria Vallés (Collège de France, Paris) for helping with manuscript reading and the Galaxy Developers Community for its continuous support in tool development. This article/publication is based upon work from EPITRAN COST Action supported by COST (European Cooperation in Science and Technology).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.518949/full#supplementary-material

Supplementary Figure 1 | tRFs could originate from tRNA space and non-tRNA space: **(A)** Representation of the different genomic loci that can give rise to tRFs: tRNAs, tRNA-lookalikes, truncated tRNAs or repeated elements. **(B)** tRFs size distribution of reads matching "tRNA space" (gray) and "non-tRNA space" (blue). **(C)** Fold change of tRF reads matching the "tRNA-space" versus tRFs reads matching the "non-tRNA space." **(D)** tRFs matching the "non-tRNA space" corresponding to tRFs-CCA (dark pink) and tRFs-non-CCA (light pink) in the represented genotypes. **(E)** comparison between Bam Coverages obtained with 15–29 nt (left, same as **Figure 3B**) and 17–29 nt (right) in different genotypes.

Supplementary Figure 2 | tRFs expression is altered in Rpp30 mutants: **(A)** Unique mature-tRNA CCA-tagged sequences were used as a reference genome to identify tRNAs that are sources of tRFs. tRFs reads were counted in a hit table, normalized by DESeq normalization Geometrical method and used to generate a heatmap for the different genotypes (1-3) reflecting all nuclear and mitochondrial tRNA genes in *Drosophila*'s genome. The read counts were sorted from maximum to minimum values for the *white-* column (left) for a better comparison with the other genotypes. Expression levels are reflected with a color-code going from blue (lowest levels), through white (middle levels), to red (highest levels) and ranging from 0 to 15 log2 counts. **(B)** Ratio of the read counts *Rpp30*$^{18.2}$ homozygous/*white-* presented in **(A)** Minimal values are in dark blue, middle values are in white, and maximal values are in red. **(C)** Heatmap generated as in **(A)**. by using "tRNA-extended CCA edited" sequences in order to analyze all types of tRFs: from precursor and mature tRNAs.

Supplementary Figure 3 | Rpp30 mutations affects tRFs biogenesis: 11 tRFs readmap profiles representing the most expressed tRFs in *white-* (see **Supplementary Figure 2C**) are shown for the different genotypes. The readmaps were obtained using normalizing factors (see Methods). Since pre-tRNAs sequences are included in the tRNA reference genome, 5′-tRFs are located at position 25 nt instead of position 0 nt. 3′-tRFs are located around position 75 nt and tRFs-1 are located around position 100 nt, depending on the length of the tRNAs and the presence of intron. Peaks determine the beginning of the reads sequences. tRFs categories are schematized in *white-* and *Rpp30*$^{18.2}$ homozygous for better comparison: 5′-tRFs in light blue, 3′-tRFs in dark blue and tRFs-1 in green.

Supplementary Figure 4 | tRFs biogenesis is altered in tRNA methylation mutants: **(A)** tRFs were counted in a hit table using tRNA-extended CCA edited as a reference genome. The table was normalized by DESeq normalization Geometrical method and used to generate a heatmap for the different genotypes (1-3) reflecting all nuclear and mitochondrial tRNA genes in *Drosophila*'s genome. The read counts were sorted from maximum to minimum values for the *white-* column (left) for better comparison with the other genotypes. Expression levels are reflected with a color-code going from blue (lowest levels), through white (middle levels), to red (highest levels). **(B)** Read counts ratio *dTrm7_34* homoz./heteroz.* calculated from **(A)** Minimal values are in dark blue, middle values are in white, and maximal values are in red, ranging from 0 to 15 log2 counts.

Supplementary Figure 5 | tRNA methylation defects alter tRFs populations: 12 tRFs normalized readmap profiles representing the most decreased or increased tRFs from the ratio *dTrm7_34** homozygous/heterozygous (see **Supplementary Figure 4B**) are shown for the different genotypes. Since pre-tRNAs sequences are included in the tRNA-reference, 5′-tRFs are located at position 25 nt instead of position 0 nt. 3′-tRFs are located around position 75 nt and tRFs-1 are located around position 100 nt, depending on the length of the tRNAs and the presence of an intron*. Peaks determine the beginning of the reads sequences. The tRFs categories are schematized in *dTrm7_34*/Tb,Sb* heterozygous for better comparison: 5′-tRFs in light blue, 3′-tRFs in dark blue and tRFs-1 in green. Ratio's values above 1: tRFs increased in *dTrm7_34** homozygous mutants. Ratio's values below 1: tRFs decreased in *dTrm7_34** homozygous mutants.

Supplementary Figure 6 | Lack of dTrm7_32 and dTrm7_34 affects the abundance of tRFs, derived from their substrate tRNAs: 13 tRFs normalized readmap profiles as examples of tRNA substrates of dTrm7_34 and dTrm7_32 are shown for the different genotypes. Since pre-tRNAs sequences are included in the tRNA-reference, 5′-tRFs are located at position 25 nt instead of position 0 nt. 3′-tRFs are located around position 75 nt and tRFs-1 are located around position 100 nt, depending on the length of the tRNA and the presence of an intron. Peaks determine the beginning of the reads sequences. tRFs are schematized in *dTrm7_34*/Tb,Sb* heterozygous and homozygous mutants for better comparison: 5′-tRFs in light blue, 3′-tRFs in dark blue and tRFs-1 in green.

Supplementary Figure 7 | tRNA processing and tRNA methylation affects snoRNA fragments (snoRFs) profiles. **(A)** General size distribution (15–29 nt) of normalized snoRFs read counts is shown for the different genotypes of tRNA processing and tRNA methylation mutants. **(B)** Violin plot reflecting snoRNAs populations found in *Drosophila melanogaster* genome. snoRNAs of more than 120 nt belong to *box H/ACA* class whereas snoRNAs of less than 120 nt belong to *box C/D* class. **(C)** Shown are snoRFs coverage profiles for the indicated genotypes (scaling factors used, see Methods). TSS: Transcription Start Site. TES, Transcription End Site. **(D)** Logo for the most representative sequences found in the last 15 nt of snoRFs for *Rpp30* mutants (issued from *fasta* files containing all sequences).

Supplementary Figure 8 | Workflow related to miRNA normalization, scale factors and global cascade annotation. Representation of the Workflows used to perform miRNA normalization **(A)** and the Scale Factors **(B)**. Shown are Scale factors for all the genotypes used in this study. For simplicity, only *white-*, *Rpp30*$^{18.2}$*homoz., Rpp30*$^{18.2}$*;ubiRpp30GFP, dTrm7_34*/Tb,Sb** heterozygous, *dTrm7_34** homozygous and *dTrm7_34*, dTrm7_32** double mutants have been used for the main figures. **(C)** Workflow used to generate the cascade annotations represented in **Figures 2**, **3**, **5**.

Supplementary Figure 9 | Workflow related to Bam Coverages and tRNA expression heatmaps. Scheme of the Workflows used for generation of tRFs Bam Coverages represented in **Figures 2**, **3**, **5**. **(A)**, tRFs Heatmaps represented in **Supplementary Figures 2**, **4 (B)**, and snoRFs Bam Coverage represented in **Supplementary Figure 7 (C)**.

Supplementary Table 1 | Comparison of some existing tRFs bioinformatic analysis tools. Comparison is only based on some selected bibliographic resources available in Pubmed, April, 2020. y: yes, n: no. "-" indicates that the information is not relevant or that the item doesn't exist.

REFERENCES

Akiyama, Y., Kharel, P., Abe, T., Anderson, P., and Ivanov, P. (2020). Isolation and initial structure-functional characterization of endogenous tRNA-derived stress-induced RNAs. *RNA Biol.* 17, 1116–1124. doi: 10.1080/15476286.2020.1732702

Angelova, M. T., Dimitrova, D. G., Da Silva, B., Marchand, V., Jacquier, C., Achour, C., et al. (2020). tRNA 2′-O-methylation by a duo of TRM7/FTSJ1 proteins modulates small RNA silencing in Drosophila. *Nucl. Acids Res.* 48, 2050–2072.

Angelova, M. T., Dimitrova, D. G., Dinges, N., Lence, T., Worpenberg, L., Carré, C., et al. (2018). The emerging field of epitranscriptomics in neurodevelopmental and neuronal disorders. *Front. Bioeng. Biotech.* 6:46. doi: 10.3389/fbioe.2018.00046

Angrisani, A., Tafer, H., Stadler, P. F., and Furia, M. (2015). Developmentally regulated expression and expression strategies of Drosophila snoRNAs. *Insect. Biochem. Mol. Biol.* 61, 69–78. doi: 10.1016/j.ibmb.2015.01.013

Balatti, V., Nigita, G., Veneziano, D., Drusco, A., Stein, G. S., Messier, T. L., et al. (2017). tsRNA signatures in cancer. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8071–8076.

Barchiesi, A., and Vascotto, C. (2019). Transcription, Processing, and Decay of Mitochondrial RNA in Health and Disease. *Int. J. Mol. Sci.* 20:2221. doi: 10.3390/ijms20092221

Boskovic, A., Bing, X. Y., Kaymak, E., and Rando, O. J. (2019). Control of noncoding RNA production and histone levels by a 5′ tRNA fragment. *Genes Dev.* 34, 118–131. doi: 10.1101/gad.332783.119

Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., Brown, J. W. S., et al. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* 15, 2147–2160. doi: 10.1261/rna.1738409

Coughlin, D. J., Pleiss, J. A., Walker, S. C., Whitworth, G. B., and Engelke, D. R. (2008). Genome-wide search for yeast RNase P substrates reveals role in maturation of intron-encoded box C/D small nucleolar RNAs. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12218–12223. doi: 10.1073/pnas.0801906105

Cozen, A. E., Quartley, E., Holmes, A. D., Hrabeta-Robinson, E., Phizicky, E. M., and Lowe, T. M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods* 12, 879–884. doi: 10.1038/nmeth.3508

Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E. L., Fabry, M. H., Kneuss, E., et al. (2018). piRNA-guided genome defense: from biogenesis to silencing. *Annu. Rev. Genet.* 52, 131–157. doi: 10.1146/annurev-genet-120417-031441

Dai, Q., Zheng, G., Schwartz, M. H., Clark, W. C., and Pan, T. (2017). Selective enzymatic demethylation of N 2,N 2 -Dimethylguanosine in RNA and its application in high-throughput tRNA sequencing. *Angew. Chem. Int. Ed.* 56, 5017–5020. doi: 10.1002/anie.201700537

Delaunay, S., and Frye, M. (2019). RNA modifications regulating cell fate in cancer. *Nat. Cell Biol.* 21, 552–559. doi: 10.1038/s41556-019-0319-0

Dimitrova, D. G., Teysset, L., and Carré, C. (2019). RNA 2′-O-Methylation (Nm) modification in human diseases. *Genes* 10:117. doi: 10.3390/genes10020117

Dubrovsky, E. B., Dubrovskaya, V. A., Levinger, L., Schiffer, S., and Marchfelder, A. (2004). Drosophila RNase Z processes mitochondrial and nuclear pre-tRNA 3′ ends in vivo. *Nucl. Acids Res.* 32, 255–262. doi: 10.1093/nar/gkh182

Durdevic, Z., Hanna, K., Gold, B., Pollex, T., Cherry, S., Lyko, F., et al. (2013a). Efficient RNA virus control in Drosophila requires the RNA methyltransferase Dnmt2. *EMBO Rep.* 14, 269–275. doi: 10.1038/embor.2013.3

Durdevic, Z., Mobin, M. B., Hanna, K., Lyko, F., and Schaefer, M. (2013b). The RNA Methyltransferase Dnmt2 is required for efficient Dicer-2-dependent siRNA pathway activity in Drosophila. *CellReports* 4, 931–937. doi: 10.1016/j.celrep.2013.07.046

Falaleeva, M., and Stamm, S. (2012). Processing of snoRNAs as a new source of regulatory non-coding RNAs. *Bioessays* 35, 46–54. doi: 10.1002/bies.201200117

Falaleeva, M., Welden, J. R., Duncan, M. J., and Stamm, S. (2017). C/D-box snoRNAs form methylating and non-methylating ribonucleoprotein complexes: old dogs show new tricks. *Bioessays* 39:1600264. doi: 10.1002/bies.201600264

Fu, H., Feng, J., Liu, Q., Sun, F., Tie, Y., Zhu, J., et al. (2008). Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett.* 583, 437–442. doi: 10.1016/j.febslet.2008.12.043

Genenncher, B., Durdevic, Z., Hanna, K., Zinkl, D., Mobin, M. B., Senturk, N., et al. (2018). Mutations in Cytosine-5 tRNA Methyltransferases impact mobile element expression and genome stability at specific DNA repeats. *Cell Rep.* 22, 1861–1874. doi: 10.1016/j.celrep.2018.01.061

Gkatza, N. A., Castro, C., Harvey, R. F., Heiß, M., Popis, M. C., Blanco, S., et al. (2019). Cytosine-5 RNA methylation links protein synthesis to cell metabolism. *PLoS Biol.* 17:e3000297. doi: 10.1371/journal.pbio.3000297

Goodarzi, H., Liu, X., Nguyen, H. C. B., Zhang, S., Fish, L., and Tavazoie, S. F. (2015). Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell* 161, 790–802. doi: 10.1016/j.cell.2015.02.053

Guan, L., Karaiskos, S., and Grigoriev, A. (2019). Inferring targeting modes of Argonaute-loaded tRNA fragments. *RNA Biol.* 17, 1070–1080. doi: 10.1080/15476286.2019.1676633

Guy, M. P., Shaw, M., Weiner, C. L., Hobson, L., Stark, Z., Rose, K., et al. (2015). Defects in tRNA Anticodon Loop 2′-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability due to Mutations inFTSJ1. *Hum. Mutat.* 36, 1176–1187. doi: 10.1002/humu.22897

Guzzi, N., and Bellodi, C. (2020). Novel insights into the emerging roles of tRNA-derived fragments in mammalian development. *RNA Biol* 17, 1214–1222. doi: 10.1080/15476286.2020.1732694

Haeusler, R. A., and Engelke, D. R. (2006). Spatial organization of transcription by RNA polymerase III. *Nucl. Acids Res.* 34, 4826–4836. doi: 10.1093/nar/gkl656

Hanada, T., Weitzer, S., Mair, B., Bernreuther, C., Wainger, B. J., Ichida, J., et al. (2013). CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature* 495, 474–480. doi: 10.1038/nature11923

Honda, S., Loher, P., Morichika, K., Shigematsu, M., Kawamura, T., Kirino, Y., et al. (2017). Increasing cell density globally enhances the biogenesis of Piwi-interacting RNAs in Bombyx mori germ cells. *Sci Rep.* 7:4110.

Huang, Z.-P. (2005). Genome-wide analyses of two families of snoRNA genes from Drosophila melanogaster, demonstrating the extensive utilization of introns for coding of snoRNAs. *RNA* 11, 1303–1316. doi: 10.1261/rna.2380905

Jarrous, N. (2017). Roles of RNase P and Its Subunits. *Trends Genet.* 33, 594–603. doi: 10.1016/j.tig.2017.06.006

Jarrous, N., and Gopalan, V. (2010). Archaeal/Eukaryal RNase P: subunits, functions and RNA diversification. *Nucl. Acids Res.* 38, 7885–7894. doi: 10.1093/nar/gkq701

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* 16, 421–433. doi: 10.1038/nrg3965

Karaiskos, S., and Grigoriev, A. (2016). Dynamics of tRNA fragments and their targets in aging mammalian brain. *F1000Res.* 5:2758. doi: 10.12688/f1000research.10116.1

Karaiskos, S., Naqvi, A. S., Swanson, K. E., and Grigoriev, A. (2015). Age-driven modulation of tRNA-derived fragments in Drosophila and their potential targets. *Biol. Direct* 10:51.

Karasik, A., Fierke, C. A., and Koutmos, M. (2019). Interplay between substrate recognition, 5′ end tRNA processing and methylation activity of human mitochondrial RNase P. *RNA* 25, 1646–1660. doi: 10.1261/rna.069310.118

Keam, S. P., Young, P. E., McCorkindale, A. L., Dang, T. H. Y., Clancy, J. L., Humphreys, D. T., et al. (2014). The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucl. Acids Res.* 42, 8984–8995. doi: 10.1093/nar/gku620

Kim, H. K., Xu, J., Chu, K., Park, H., Jang, H., Li, P., et al. (2019). A tRNA-derived small RNA regulates ribosomal protein S28 protein levels after translation initiation in humans and mice. *Cell Rep.* 29, 3816.e4–3824.e4.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 10, 126–139. doi: 10.1038/nrm2632

Kumar, P., Anaya, J., Mudunuri, S. B., and Dutta, A. (2014a). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol.* 12:78. doi: 10.1186/s12915-014-0078-0

Kumar, P., Mudunuri, S. B., Anaya, J., and Dutta, A. (2014b). tRFdb: a database for transfer RNA fragments. *Nucl. Acids Res.* 43, D141–D145.

Kumar, P., Mudunuri, S.B., Anaya, J., and Dutta, A. (2015). tRFdb: a database for transfer RNA fragments. *Nucl. Acids Res.* 43, D141–D145. doi: 10.1093/nar/gku1138

Kumar, P., Kuscu, C., and Dutta, A. (2016). Biogenesis and function of transfer RNA-related fragments (tRFs). *Trends Biochem. Sci.* 41, 679–689. doi: 10.1016/j.tibs.2016.05.004

Kuscu, C., Kumar, P., Kiran, M., Su, Z., Malik, A., and Dutta, A. (2018). tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA* 24, 1093–1105. doi: 10.1261/rna.066126.118

Li, L., Gu, W., Liang, C., Liu, Q., Mello, C. C., and Liu, Y. (2012). The translin–TRAX complex (C3PO) is a ribonuclease in tRNA processing. *Nat. Struct. Mol. Biol.* 19, 824–830. doi: 10.1038/nsmb.2337

Li, S., Xu, Z., and Sheng, J. (2018). tRNA-derived small RNA: a novel regulatory small non-coding RNA. *Genes* 9:246. doi: 10.3390/genes9050246

Li, Z., Ender, C., Meister, G., Moore, P. S., Chang, Y., and John, B. (2012). Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucl. Acids Res.* 40, 6787–6799. doi: 10.1093/nar/gks307

Liu, S., Chen, Y., Ren, Y., Zhou, J., Ren, J., Lee, I., et al. (2018). A tRNA-derived RNA fragment plays an important role in the mechanism of Arsenite -induced cellular responses. *Sci. Rep.* 8:16838. doi: 10.1038/s41598-018-34899-2

Loher, P., Telonis, A. G., and Rigoutsos, I. (2017a). *Accurate Profiling and Quantification of tRNA Fragments from RNA-Seq Data: A Vade Mecum for MINTmap in Methods in Molecular Biology*. New York, NY: Springer, 237–255.

Loher, P., Telonis, A. G., and Rigoutsos, I. (2017b). MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci. Rep.* 7:41184. doi: 10.1038/srep41184

Lyons, S. M., Fay, M. M., and Ivanov, P. (2018). The role of RNA modifications in the regulation of tRNA cleavage. *FEBS Lett.* 592, 2828–2844. doi: 10.1002/1873-3468.13205

Martinez, G. (2017). tRNA-derived small RNAs: new players in genome protection against retrotransposons. *RNA Biol.* 15, 170–175. doi: 10.1080/15476286.2017.1403000

Martinez, G., Choudury, S. G., and Slotkin, R. K. (2017). tRNA-derived small RNAs target transposable element transcripts. *Nucl. Acids Res.* 45, 5142–5152. doi: 10.1093/nar/gkx103

Marvin, M. C., Clauder-Munster, S., Walker, S. C., Sarkeshik, A., Yates, J. R., Steinmetz, L. M., et al. (2011). Accumulation of noncoding RNA due to an RNase P defect in Saccharomyces cerevisiae. *RNA* 17, 1441–1450. doi: 10.1261/rna.2737511

Maute, R. L., Schneider, C., Sumazin, P., Holmes, A., Califano, A., Basso, K., et al. (2013). tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1404–1409. doi: 10.1073/pnas.1206761110

Mo, D., Jiang, P., Yang, Y., Mao, X., Tan, X., Tang, X., et al. (2019). A tRNA fragment, 5′-tiRNAVal, suppresses the Wnt/β-catenin signaling pathway by targeting FZD3 in breast cancer. *Cancer Lett.* 457, 60–73. doi: 10.1016/j.canlet.2019.05.007

Mollà-Herman, A., Angelova, M., Carré, C., Antoniewski, C., and Huynh, J.-R. (2019). tRNA fragments (tRFs) populations analysis in mutants affecting tRNAs processing and tRNA methylation. *Biorxiv[Preprint]*. doi: 10.1101/869891

Molla-Herman, A., Vallés, A. M., Ganem-Elbaz, C., Antoniewski, C., and Huynh, J.-R. (2015). tRNA processing defects induce replication stress and Chk2-dependent disruption of piRNA transcription. *EMBO J.* 34, 3009–3027. doi: 10.15252/embj.201591006

Motorin, Y., and Helm, M. (2019). Methods for RNA modification mapping using deep sequencing: established and new emerging technologies. *Genes* 10:35. doi: 10.3390/genes10010035

Ontiveros, R. J., Stoute, J., and Liu, K. F. (2019). The chemical diversity of RNA modifications. *Biochem. J.* 476, 1227–1245. doi: 10.1042/bcj20180445

Park, O. H., Ha, H., Lee, Y., Boo, S. H., Kwon, D. H., Song, H. K., et al. (2019). Endoribonucleolytic cleavage of m6A-containing RNAs by RNase P/MRP complex. *Mol. Cell* 74, 494.e–507.e.

Patterson, D. G., Roberts, J. T., King, V. M., Houserova, D., Barnhill, E. C., Crucello, A., et al. (2017). Human snoRNA-93 is processed into a microRNA-like RNA that promotes breast cancer cell invasion. *NPJ Breast Cancer* 3:25. doi: 10.1038/s41523-017-0032-8

Pintard, L. (2002). Trm7p catalyses the formation of two 2′-O-methylriboses in yeast tRNA anticodon loop. *EMBO J.* 21, 1811–1820. doi: 10.1093/emboj/21.7.1811

Pliatsika, V., Loher, P., Telonis, A. G., and Rigoutsos, I. (2016). MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics* 32, 2481–2489. doi: 10.1093/bioinformatics/btw194

Romano, G., Veneziano, D., Acunzo, M., and Croce, C. M. (2017). Small non-coding RNA and cancer. *Carcinogenesis* 38, 485–491.

Rossmanith, W. (2012). Of P and Z: mitochondrial tRNA processing enzymes. *Biochim. Biophys. Acta Gene Regul. Mech.* 1819, 1017–1026. doi: 10.1016/j.bbagrm.2011.11.003

Saoura, M., Powell, C. A., Kopajtich, R., Alahmad, A., Al-Balool, H. H., Albash, B., et al. (2019). Mutations in ELAC2 associated with hypertrophic cardiomyopathy impair mitochondrial tRNA 3′-end processing. *Hum. Mutat.* 40, 1731–1748. doi: 10.1002/humu.23777

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., et al. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.* 24, 1590–1595. doi: 10.1101/gad.586710

Schorn, A. J., Gutbrod, M. J., LeBlanc, C., and Martienssen, R. (2017). LTR-retrotransposon control by tRNA-derived Small RNAs. *Cell* 170, 61.e11–71.e11.

Schorn, A. J., and Martienssen, R. (2018). Tie-Break: host and Retrotransposons Play tRNA. *Trends Cell Biol.* 28, 793–806. doi: 10.1016/j.tcb.2018.05.006

Selitsky, S. R., and Sethupathy, P. (2015). tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. *BMC Bioinformatics* 16:354. doi: 10.1186/s12859-015-0800-0

Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., Bing, X. Y., et al. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* 351, 391–396. doi: 10.1126/science.aad6780

Sharma, U., Sun, F., Conine, C. C., Reichholf, B., Kukreja, S., Herzog, V. A., et al. (2018). Small RNAs are trafficked from the epididymis to developing mammalian sperm. *Dev. Cell* 46, 481.e–494.e.

Shen, Y., Yu, X., Zhu, L., Li, T., Yan, Z., and Guo, J. (2018). Transfer RNA-derived fragments and tRNA halves: biogenesis, biological functions and their roles in diseases. *J. Mol. Med.* 96, 1167–1176. doi: 10.1007/s00109-018-1693-y

Siira, S. J., Rossetti, G., Richman, T. R., Perks, K., Ermer, J. A., Kuznetsova, I., et al. (2018). Concerted regulation of mitochondrial and nuclear non-coding RNAs by a dual-targeted RNase Z. *EMBO Rep* 19:e46198. doi: 10.15252/embr.201846198

Soares, A. R., and Santos, M. (2017). Discovery and function of transfer RNA-derived fragments and their role in disease. *Wiley Interdiscip. Rev. RNA* 8, doi: 10.1002/wrna.1423

Sobala, A., and Hutvagner, G. (2011). Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdiscip. Rev. RNA* 2, 853–862. doi: 10.1002/wrna.96

Sokołowski, M., Klassen, R., Bruch, A., Schaffrath, R., and Glatt, S. (2018). Cooperativity between different tRNA modifications and their modification pathways. *Biochim. Biophys. Acta Gene Regul. Mech.* 1861, 409–418. doi: 10.1016/j.bbagrm.2017.12.003

Su, Z., Kuscu, C., Malik, A., Shibata, E., and Dutta, A. (2019). Angiogenin generates specific stress-induced tRNA halves and is not involved in tRF-3–mediated gene silencing. *J. Biol. Chem.* 294, 16930–16941. doi: 10.1074/jbc.ra119.009272

Światowy, W., and Jagodziński, P. P. (2018). Molecules derived from tRNA and snoRNA: entering the degradome pool. *Biomed. Pharmacother.* 108, 36–42. doi: 10.1016/j.biopha.2018.09.017

Telonis, A. G., Kirino, Y., and Rigoutsos, I. (2015a). Mitochondrial tRNA-lookalikes in nuclear chromosomes: could they be functional? *RNA Biol.* 12, 375–380. doi: 10.1080/15476286.2015.1017239

Telonis, A. G., Loher, P., Honda, S., Jing, Y., Palazzo, J., Kirino, Y., et al. (2015b). Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget* 6, 24797–24822. doi: 10.18632/oncotarget.4695

Telonis, A. G., Loher, P., Kirino, Y., and Rigoutsos, I. (2014). Nuclear and mitochondrial tRNA-lookalikes in the human genome. *Front. Genet.* 5:344. doi: 10.3389/fgene.2014.00344

Telonis, A. G., Loher, P., Kirino, Y., and Rigoutsos, I. (2016). Consequential considerations when mapping tRNA fragments. *BMC Bioinform.* 17:123. doi: 10.1186/s12859-016-0921-0

Telonis, A. G., Loher, P., Magee, R., Pliatsika, V., Londin, E., Kirino, Y., et al. (2019). tRNA fragments show intertwining with mRNAs of specific repeat content and have links to disparities. *Cancer Res.* 79, 3034–3049. doi: 10.1158/0008-5472.can-19-0789

Thompson, D. M., Lu, C., Green, P. J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA* 14, 2095–2103. doi: 10.1261/rna.1232808

Torres, A. G., Reina, O., Stephan-Otto Attolini, C., and Ribas de Pouplana, L. (2019). Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proc. Natl. Acad. Sci. U.S.A.* 116, 8451–8456. doi: 10.1073/pnas.1821120116

Vitali, P., and Kiss, T. (2019). Cooperative 2′-O-methylation of the wobble cytidine of human elongator tRNAMet(CAT) by a nucleolar and a Cajal body-specific box C/D RNP. *Genes Dev.* 33, 741–746. doi: 10.1101/gad.326363.119

Wellner, K., Betat, H., and Mörl, M. (2018). A tRNA's fate is decided at its 3′ end: collaborative actions of CCA-adding enzyme and RNases involved in tRNA processing and degradation. *Biochim. Biophys. Acta Gene Regul. Mech.* 1861, 433–441. doi: 10.1016/j.bbagrm.2018.01.012

Willis, I. M., and Moir, R. D. (2018). Signaling to and from the RNA polymerase III transcription and processing machinery. *Annu. Rev. Biochem.* 87, 75–100. doi: 10.1146/annurev-biochem-062917-012624

Xie, X., Dubrovskaya, V., Yacoub, N., Walska, J., Gleason, T., Reid, K., et al. (2013). Developmental roles of Drosophila tRNA processing endonuclease RNase ZL as revealed with a conditional rescue system. *Dev. Biol.* 381, 324–340. doi: 10.1016/j.ydbio.2013.07.005

Yamanaka, S., and Siomi, H. (2015). Misprocessed tRNA response targets pi RNA clusters. *EMBO J.* 34, 2988–2989. doi: 10.15252/embj.201593322

Zhang, W., Chang, J.-W., Lin, L., Minn, K., Wu, B., Chien, J., et al. (2015). Network-based isoform quantification with RNA-Seq data for cancer transcriptome analysis. *PLoS Comput. Biol.* 11:e1004465. doi: 10.1371/journal.pcbi.1004465

Zhu, L., Liu, X., Pu, W., and Peng, Y. (2018). tRNA-derived small non-coding RNAs in human disease. *Cancer Lett.* 419, 1–7. doi: 10.1016/j.canlet.2018.01.015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article’s readership