Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

MetaTX: deciphering the distribution of mRNA-related features in the presence of isoform ambiguity, with applications in epitranscriptome analysis

Yue Wang[1,4], Kunqi Chen[2,5], Zhen Wei[2,5], Frans Coenen[4], Jionglong Su[1] and Jia Meng[2,3,6,*]

[1]Department of Mathematical Sciences, [2]Department of Biological Sciences, [3]AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; [4]Department of Computer Science, [5]Institute of Ageing & Chronic Disease, [6]Institute of Integrative Biology, University of Liverpool, L69 7ZB, Liverpool, United Kingdom;

*To whom correspondence should be addressed: jia.meng@xjtlu.edu.cn (JM)

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The distribution of biological features strongly indicates their functional relevance. Compared to DNA-related features, deciphering the distribution of mRNA-related features is non-trivial due to the existence of isoform ambiguity and compositional diversity of mRNAs.

Results: We propose here a rigorous statistical framework, MetaTX, for deciphering the distribution of mRNA-related features. Through a standardized mRNA model, MetaTX firstly unifies various mRNA transcripts of diverse compositions, and then corrects the isoform ambiguity by incorporating the overall distribution pattern of the features through an EM algorithm. MetaTX was tested on both simulated and real data. Results suggested that MetaTX substantially outperformed existing direct methods on simulated datasets, and that a more informative distribution pattern was produced for all the three datasets tested, which contain N^6-Methyladenosine sites generated by different technologies. MetaTX should make a useful tool for studying the distribution and functions of mRNA-related biological features, especially for mRNA modifications such as N^6-Methyladenosine.

Availability: The MetaTX R package is freely available at GitHub: https://github.com/yue-wang-biomath/MetaTX.1.0.

1 Introduction

Recent development of high-throughput sequencing technologies has enabled the transcriptome-wide profiling of RNA modification sites (Dominissini, *et al*., 2013; Dominissini, *et al*., 2012; Meyer, *et al*., 2012; Schaefer, *et al*., 2009). To date, more than 170 different types of RNA modification have been identified in all three kingdoms of life, many of which have been found to play important roles in various biological processes. For example, N^6-Methyladenosine (m^6A) can regulate the stability and translation efficiency of mRNA (Mauer, *et al*., 2017; Slobodin, *et al*., 2017; Wang, *et al*., 2015; Wang, *et al*., 2014), and affect the circadian clock, cell differentiation, neuron production, alternative splicing and RNA-protein interaction (Fustin, *et al*., 2013; Geula, *et al*., 2015; Liu, *et al*., 2015; Pendleton, *et al*., 2017).

One basic way to characterize a biological feature is to see how it is distributed with respect to a gene, which may be shown in the form of a metagene plot (Beauparlant, et al., 2016), also referred to as a meta-gene (Shin, et al., 2009) or aggregation plot (Kundaje, et al., 2012). The distribution of a biological feature strongly indicates the potential functional relevance of the feature of interests, although such association may not be direct or causal. For example, the enrichment of histone modification H3K4me3 near to transcription start sites is clearly linked to its transcription initialization function (Barski, et al., 2007). However, compared to DNA-related features (such as histone modification and DNA methylation), deciphering the distribution of mRNA-related features (such as mRNA modifications) is non-trivial due to the following reasons:

- *Isoform Ambiguity*. Although actually located in the heterogeneous transcriptome, mRNA-related biological features are often denoted only by genome-based coordinates in bioinformatics databases. The isoform-specific belongings of mRNA related features may be unavailable in the presence of multiple isoform transcription of the same gene due to technical limitations. For example, most of the existing epitranscriptome profiling approaches, such as MeRIP-seq and

miCLIP, suffer from the isoform ambiguity issue, and instead report only the genome-based coordinates of RNA modifications. When an RNA modification site overlaps with multiple transcripts according to the transcriptome annotation, it is not clear which transcript is associated with the site. This creates obvious difficulty when characterizing the distribution of the feature of interest.

- *More complex landscape of mRNAs.* The distribution analysis for DNA-related features is usually based on two landmarks only, i.e., the transcription start and transcription end positions, while the analysis of mRNA-related features should involve four landmarks of the mRNAs molecule, i.e., 5'end, start codon, stop codon and 3'end of mRNAs, making it more complex than the case of DNA-related features.
- *Compositional diversity of mRNAs.* Additionally, it is important to note the existence of compositional diversity among different mRNAs. For example, some mRNAs may have short CDS but a super long 3'UTR, while some others may have very short 3'UTR, or even no 3'UTR at all for some cases, according to the transcriptome annotation database. The vast compositional diversity among different mRNAs makes it difficult to compare across multiple mRNAs (or genes), and may bring concerns about the validity of the overall distribution pattern necessary for characterizing mRNA-related features.

The reasons provided above (see Figure 1), together compound the difficulty and complexity of distribution characterization for mRNA-related features.

To date, a number of software tools have been developed for deciphering the distribution of mRNA-related features. Guitar was the first method dedicated to sketching the transcriptomic view of RNA-related genomic features, and provided an open source R/Bioconductor package (Cui, *et al*., 2016). The Perl/R pipeline MetaPlotR was invented for plotting metagenes of various modified sites (Olarerin-George and Jaffrey, 2017). A Shiny web framework-based web server txCoords was invented for transcriptomic peak re-mapping (Yan, *et al*., 2017). The epitranscriptome database MeTDB (Liu, *et al*., 2018) provided a web-based graphical user interface for the Guitar R package (Cui, *et al*., 2016). The RNA modification annotation database, RNAmod, also supported metagene plot functionality along with various annotations of diverse mRNA modifications in different species (Liu and Gregory, 2019). Despite the efforts that have been made, only simple heuristic strategies have been taken by the above approaches to resolve the aforementioned difficulties, e.g., retaining only the longest transcript of a gene in the analysis to avoid isoform ambiguity, keeping only the mRNAs with both 5'UTR and 3UTR longer than 100nt to ensure they are relatively comparable. None of the above approaches quantitatively formulated the problem of concern, and a general and rigorous solution has yet to be available.

We propose here a rigorous statistical framework, MetaTX, for deciphering the distribution of mRNA-related features in the presence of isoform ambiguity and compositional diversity of mRNAs. Through a standardized mRNA model, MetaTX first unifies various mRNA transcripts, some of which may have vastly different compositions, and then corrects the isoform ambiguity by incorporating the overall distribution pattern of the feature through an EM algorithm via a latent variable. MetaTX was tested on both simulation data and real RNA N^6-methyladenosine data generated from three different epitranscriptome profiling approaches (Chen, *et al*., 2015; Olarerin-George and Jaffrey, 2017; Schwartz, *et al*., 2014). Results suggested that MetaTX consistently exhibited an improved performance with higher accuracy on simulated data compared to the Guitar (Cui, *et al*., 2016) and MetaPlotR (Olarerin-George and Jaffrey, 2017), which did not consider biases from isoform ambiguity, and reported more prominent distribution patterns for all the three datasets tested. MetaTX is available as an open source R package, and is a useful tool for studying the distribution and functions of mRNA-related biological features, especially for mRNA modifications, such as N^6-Methyladenosine and Pseudouridylation.

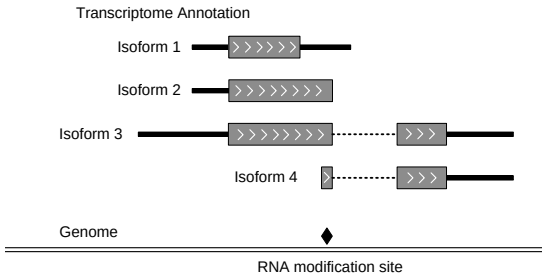

Fig. 1. Isoform ambiguity and compositional diversity of mRNAs. Although physically located on the RNAs, many mRNA-related features are only recorded by genome-based coordinates, the transcript-level to which they belong remains unclear due to technical limitations. In the above example, the RNA modification site is denoted by genome-based coordinate, and overlaps with 4 isoform transcripts of the same gene. It may be associated with the 3'UTR of isoform 1, near the stop codon on the CDS of isoform 2, etc., which may cause problems when characterizing the distribution of this mRNA-related feature. Note that, isoform 1 has longer 3'UTR, isoform 2 has no 3'UTR, and isoform 3 has longer 5' UTR, while isoform 4 has no 5' UTR at all. The compositional difference may make it difficult to compare across multiple mRNAs of the same or different genes.

2 Methods

In this section, we first introduce the standard mRNA model, through which mRNAs of diverse compositions may be unified, and then propose our overall formulation for the distribution analysis of mRNA-related features. An EM solution is then provided to resolve the isoform ambiguity problem via the overall distribution pattern inferred.

2.1 Coordinate Standardization

To unify the mRNAs with diverse composition, we considered a standard mRNA model, in which the three main components of mRNA (5'UTR, CDS and 3'UTR) were each divided into the same number of bins of equal width, for every individual mRNA. Figure 2 illustrates the process of coordinate standardization, and we refer to each bin of mRNA by its coordinate on the standardized mRNA model. Conceivably, as the corresponding coordinates on different mRNAs are located on biologically similar regions, they are likely to be associated with the same type of biological features, or regulated by the same type of signal (such as the same type of RNA modification). The coordinates of different mRNA were then made comparable. It is worth noting that, although not explicitly stated, many existing approaches will have assumed a similar standardized mRNA model in their analysis.

In practice, we also considered the flanking regions of the mRNAs, including 1kb promoter regions before the 5'end and 1kb tail region after the 3'end. These two regions are also independently divided into the same number of bins with equal width. Although theoretically there should be no mRNA-related features originating from these two regions, there are always quite a few mRNA-related features that fall into these regions due to incomplete transcriptome annotation, isoform ambiguity, noise, etc.

These two regions are used as the negative control regions in our analysis; the signal within these two regions directly reflects analysis bias.

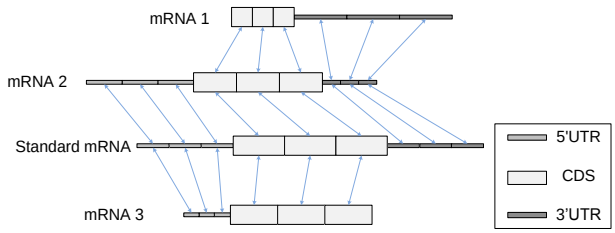

Fig. 2. Coordinate Standardization. The three main components of mRNA, i.e., 5'UTR, CDS and 3'UTR, were each divided into the same number of bins of equal width, respectively, for every individual mRNA. As the corresponding bins (referred to as coordinates) on different mRNAs are located on biologically similar region, they are likely to be associated with the same type of biological features. The coordinates of mRNAs, which are of diverse composition, were then made comparable.

2.2 The MetaTX Model

2.2.1 Basic formulation

We refer mRNA features to biological features that reside on mRNAs, such as, m^{6}A RNA methylation sites, microRNA binding site, etc. Of interests here is the distribution of mRNA-related features over the standard mRNA model, as is shown in a metagene plot. We denote the total number of mRNA-related features by *S*. As previously stated, we independently divided the 5'UTR, CDS and 3'UTR of mRNAs into a number of bins with equal width, respectively. We assume there are a total of *C* bins (or coordinates) on each mRNA. The parameter *C* is essentially the resolution of our distribution analysis. A greater *C* may result in a distribution analysis with high-resolution, but also increases the computation load.

With the coordinates of mRNA standardized, it is now possible to calculate the standardized coordinate of each mRNA-related feature in the mRNA that overlapped with it. Let *T* represent the total number of transcripts. We denote the overlap between features and mRNAs with a three dimensional matrix $O := \{o_{s,t,c} \mid s = 1,....,S, t = 1,...,T, c = 1,....,C\}$, with $o_{s,t,c} = 1$ indicating the *s*-th feature overlap with the *c*-th coordinate on the *t*-th mRNAs on the genome, suggesting a possible association between the two; and $o_{s,t,c} = 0$ otherwise. It should be noted that a feature may overlap with more than one mRNAs due to isoform ambiguity, and $o_{s,t,c} = 1$ does not necessarily mean that the *s*-th feature is actually associated with the *t*-th mRNA.

Additionally, we denote the width of the *c*-th coordinate on the *t*-th transcript by $w_{t,c}$ with $t = \{1,...,T\}$ and $c = \{1,...,C\}$. This parameter is important in penalizing the varying width of the corresponding bins on different mRNAs. When the corresponding component is not available, e.g., an mRNA without 3'UTR, the width of the corresponding bin (coordinate) should be set to 0.

2.2.2 Maximum a Likelihood Estimation (MLE)

We define a parameter set $\Theta := \{\alpha_c \mid c = 1,....,C\}$, where α_c is the probability that a site is located on the *c*-th coordinate of mRNA, and with $\sum_{c=1}^{C} \alpha_c = 1$. Due to the alternative splicing, a site may be located on several distinct transcript isoforms. So we denote the probability that the *s*-th site in the observed data O resides on the *t*-th transcript by a variable $\mu_{s,t}$. We should have $\sum_{t=1}^{T} \mu_{s,t} = 1$.

The probability of observing all overlapping events of a feature is:

$$P(\{o_{s,t,c} \mid t = 1,...,T, c = 1,...,C\}) = \sum_{t=1}^{T}\sum_{c=1}^{C} o_{s,t,c}\mu_{s,t}\alpha_c \quad (1)$$

With the assumption of each feature being observed independently, the likelihood of all our observed overlapping events, between the features and transcripts, can then be represented by:

$$L(O \mid \Theta) = \prod_{s=1}^{S}\sum_{t=1}^{T}\sum_{c=1}^{C} o_{s,t,c}\mu_{s,t}\alpha_c \quad (2)$$

Furthermore, the estimated parameters are denoted by:

$$\hat{\Theta} = \arg\max_{\theta} \log L(\Theta \mid O) \quad (3)$$

2.2.3 Weight Assignment

Due to isoform ambiguity, a given mRNA-feature may overlap with multiple transcripts, and the significance of different overlapping events may not be the same. There exist two ways to resolve the isoform ambiguity problem. One way is to pick the highest expressed isoform when gene expression data is available. Considering the limited detectability (or sensibility) of biotechnology, it is often safe to assume the observed phenomenon is associated with the most abundant molecule rather than one with a lower abundance; Alternatively, an equally popular but more convenient method, is to consider only the primary transcript, which is usually regarded as the longest one (Olarerin-George and Jaffrey, 2017).

Since most genes have multiple isoform transcripts, rather than considering only one isoform transcript with all other transcripts discarded, we seek to consider all transcripts simultaneously. This should lead to a more reliable result reflecting more general characteristics of the entire feature set of interest. Specifically, we introduced a weight ω_t to penalize the overlapping event observed with a specific transcript, which directly reflects the relative importance of the transcript in the problem of concern. The probability of observing all overlapping events of a feature then becomes:

$$P(\{o_{s,t,c} \mid t = 1,...,T, c = 1,...,C\}) = \left(\sum_{t=1}^{T}\sum_{c=1}^{C} o_{s,t,c}\mu_{s,t}\alpha_c\omega_t\right) \Big/ W_s \quad (4)$$

where W_s is a normalizing constant equals to $\sum_{t=1}^{T}\sum_{c=1}^{C} o_{s,t,c}\omega_t$. So the likelihood of our observation can be represented by:

$$L(\Theta \mid O) \propto \prod_{s=1}^{S}\sum_{t=1}^{T}\sum_{c=1}^{C} o_{s,t,c}\mu_{s,t}\omega_t\alpha_c \quad (5)$$

As longer transcripts can provide higher relative resolution with respect to the location on a standard gene model (see **Supplementary Figure S2**), without loss of generality, MetaTX implements the following default setting of ω_t:

$$\omega_t = \left(\sum_{c=1}^{C} w_{t,c}\right)^{\lambda} \quad (6)$$

where $\lambda = 2$ represents the degree of penalization for shorter transcripts to favor the primary transcript. Keeping only the longest transcript for the analysis, as often done in existing studies, would equate to setting $\lambda = \infty$. When λ is set to 0, all transcripts are considered equally without preference. Additionally, it is also possible to customize ω_t with other information such as the expression of isoform transcripts or a combination of both the expression and transcript length.

2.3 The MetaTX Model

The MLE problem in (5) may be solved using the Expectation-Maximization framework (Dempster, *et al*., 1977). We introduced the latent variables $\{\gamma_{s,t,c} \mid s=1,....,S, t=1,...,T, c=1,....,C\}$, where $\gamma_{s,t,c}=1$ means the *s*-th feature is physically located at the *c*-th coordinate in the *t*-th transcript; and $\gamma_{s,t,c}=0$ otherwise. Note that $\gamma_{s,t,c}$ is different from $o_{s,t,c}$, i.e., when $\gamma_{s,t,c}=1$, we should have $o_{s,t,c}=1$; but the other way around may not be true. Let $p_{s,t,c}=P(\gamma_{s,t,c}=1|O;\Theta)$ be the conditional probability of the *s*-th feature located at the *c*-th bin of the *t*-th transcript conditioned on parameters Θ and the observed data O. For each feature *s*, we should have $\sum_{t=1}^{T}\sum_{c=1}^{C} p_{s,t,c}=1$. Additionally, $\mu_{s,t}$ is equivalent to $\sum_{c=1}^{C} p_{s,t,c}$. Particularly, $\mu_{s,t}$ takes the same value of $p_{s,t,c}$ where $o_{s,t,c}=1$.

We define the set of unknown parameters to be estimated as $\Omega := \{\alpha_c, p_{s,t,c} \mid s=1,...,S, t=1,...,T, c=1,...C\}$. To estimate the unknown parameters, an EM algorithm is implemented. In the E step of the EM framework, we update the latent variable by taking its expected value:

$$p_{s,t,c}=\frac{o_{s,t,c}\omega_t\alpha_c}{\sum_{i=1}^{T}\sum_{j=1}^{C} o_{s,i,j}\omega_i\alpha_j} \tag{7}$$

Next, we define $F=\prod_{s=1}^{S}\sum_{t=1}^{T}\sum_{c=1}^{C} p_{s,t,c}\omega_t\alpha_c$, which equates to, and is a simplified result of, the right hand side of (5). Then a lower bound of F may be found with Jensen's inequality:

$$\log\left(\sum_{c=1}^{C}\sum_{t=1}^{T} p_{s,t,c}\omega_t\alpha_c\right) \geq \sum_{c=1}^{C}\left(\sum_{t=1}^{T} p_{s,t,c}\omega_t\right)\log(\alpha_c) \tag{8}$$

Taking the right side of (8) as the exponent and the natural constant as the base, leads to:

$$\sum_{t=1}^{T}\sum_{c=1}^{C} p_{s,t,c}\omega_t\alpha_c \geq \prod_{c=1}^{C}\alpha_c^{\sum_{t=1}^{T} p_{s,t,c}\omega_t} \tag{9}$$

Finally, we obtained the lower bound of *F* according to the inequality in (9), denoted by:

$$f=\prod_{s=1}^{S}\prod_{c=1}^{C}\alpha_c^{\sum_{t=1}^{T} p_{s,t,c}\omega_t} \tag{10}$$

According to the Lagrange multiplier method, to maximize $f(\{\alpha_c\}_{c=1}^{C})$ with the constraint $\sum_{c=1}^{C}\alpha_c=1$, is equivalent to solving the following two equations:

$$\begin{aligned} g(\{\alpha_c\}_{c=1}^{C}) &= \sum_{c=1}^{C}\alpha_c - 1 = 0 \\ \nabla f(\{\alpha_c\}_{c=1}^{C}) &= \lambda\nabla g(\{\alpha_c\}_{c=1}^{C}) \end{aligned} \tag{11}$$

Solving (11), $f(\{\alpha_c\}_{c=1}^{C})$ attains a maximum when $\{\alpha_c\}_{c=1}^{C}$ take values below:

$$\alpha_c=\frac{\sum_{i=1}^{S}\sum_{j=1}^{T} p_{i,j,c}\omega_j}{\sum_{i=1}^{S}\sum_{j=1}^{T}\sum_{l=1}^{C} p_{i,j,l}\omega_j} \tag{12}$$

The EM algorithm of our model may then be summarized as follows:

- Given $\{\alpha_c\}_{c=1}^{C}$ estimate $\{p_{s,t,c} \mid s=1,....,S, t=1,...,T, c=1,....,C\}$ by E step (7).
- Given $\{p_{s,t,c} \mid s=1,....,S, t=1,...,T, c=1,....,C\}$ estimate $\{\alpha_c\}_{c=1}^{C}$ by M step (12).
- Iteratively perform the above two steps until convergence. The initial value of each α_c is set to 1/*C*.

2.4 Absolutely abundance of mRNA features

Combined with the actual width $w_{t,c}$ of each mRNA component, we may obtain the absolute abundance of mRNA-features for each coordinate on the standardized mRNA model, with the unit number of features per nt sequence of the mRNA transcript, as follows:

$$d_c = S\alpha_c \Big/ \sum_{t=1}^{T} w_{t,c} \tag{13}$$

The average abundance of the features on the entire mRNA may then be calculated in a similar way by:

$$d_a = S \Big/ \left(\sum_{c=1}^{C}\sum_{t=1}^{T} w_{t,c}\right) \tag{14}$$

which may be used as a standard to search for the regions (or coordinates) enriched with the feature of interest.

It is important to note that mRNA (rather than DNA) is used as the background during the calculation of feature abundance. The shared exons of multiple isoform mRNAs were counted multiple times for each individual transcript and with all the introns removed. The absolute and average density d_c and d_a estimated from our model is thus likely to be different from those returned from existing genome-based methods. With the help of *ggplot2* (Ginestet, 2011) and other tools, our MetaTX R package provides a visualization of the distribution of mRNA-features alongside the standard mRNA model. Inclusion of the promoter and tail regions are also optional.

3 Results

3.1 Testing on simulated data

We firstly validated the proposed method on simulated datasets, which contained the 5'UTR, CDS, and 3'UTR regions respectively. When generating the simulated datasets using 1,000 transcripts randomly selected from the UCSC gene database, 10 sites were then randomly picked from each transcript within the relative mRNA component. As a result, there were a total of 10,000 sites, chosen from each mRNA component. After remapping, it may be expected that these sites are arranged evenly within the corresponding mRNA component, but not in other regions.

We then drew the distribution of the three simulated datasets, corresponding to 5'UTR, CDS and 3'UTR, via the Guitar, MetaPlotR, filter method and MetaTX. Note none of the other methods except MetaTX consider biases from isoform ambiguity problem. Guitar counts the mRNA-features multiple times for all transcripts when isoform ambiguity exists; while MetaPlotR by default retrains only the primary (or longest) isoform transcript of a gene to avoid such ambiguity. The filter method discards short mRNA components (less than 100 nt) to keep only the information located on long components, which should more informative. We also included the promoter and tail regions as negative control regions, which do not correspond to mRNA regions and thus in theory should not contain signals from mRNA-related features. As shown in Figure 3 (a), (b), and (c), stronger bias was observed in the results of the direct estimation

method and the filter method. After correcting this isoform ambiguity via the MetaTX model, the accuracy of estimates increased notably.

To quantitatively measure the accuracy of our model, we calculated the consistency between the estimated distribution and the ground truth distribution with the Kullback-Leibler (KL) divergence (**Table 1**). Results suggested that MetaTX substantially outperformed the competing methods (Guitar and MetaPlotR), and the filter strategy can slightly boost the quality of results by removing less informative datasets.

Table 1 Performance evaluation of MetaTX and competing methods

Methods	The 3 Tests (and KL Divergence)		
	3'UTR	CDS	5'UTR
MetaTX	0.07	0.12	0.06
Guitar	0.29	1.02	0.22
MetaPlotR	0.44	1.09	0.32
Filtered	0.25	1.01	0.26

3.2 Testing on real data

Next we analyzed N^6-methyladenosine (m^6A) datasets derived from different high-throughput sequencing approaches, including an miCLIP-seq dataset (Linder, *et al*., 2015; Olarerin-George and Jaffrey, 2017), a PA-m^6A-seq dataset (Chen, *et al*., 2015) and an m^6A-seq dataset (Schwartz, *et al*., 2014). N^6-methyladenosine is the most abundant RNA modification on mRNA, and has been regarded to be enriched near the stop codon of mRNAs (Dominissini, et al., 2012; Meyer, et al., 2012).

As is shown in Figure 3 (d), (e) and (g), all three methods reported an enrichment of m^6A RNA methylation near the stop codon of mRNAs. Although the ground truth distribution of m^6A on mRNAs is unavailable, it is evident that MetaTX reported a more prominent and reasonable pattern for all the 3 datasets tested, which is reflected by the reduced signal at the negative control regions (promoter and tail DNA regions). The negative control regions do not correspond to mRNA transcripts and thus shouldn't carry m^6A signals. It is worth noting that the promoter and tail regions were defined in a transcript-specific manner, i.e., each transcript has a different set of promoter and tail. Including all isoforms does not necessarily reduce the number of sites being assigned to the negative control region. Because the MetaTX model considers all isoform transcripts including the very short ones, it actually has larger proportion of negative control regions compared to the method which considers only the longest transcript.

Meanwhile, consistent of our knowledge (Dominissini, et al., 2012; Meyer, et al., 2012), a strong enrichment pattern was also observed around the stop codon compared to the methods without correction of the isoform ambiguity. The correction seems particularly effective for the PA-m^6A-seq dataset shown in Figure 3 (e). Importantly, MetaTX method achieved very stable performance even when we discard all the features without isoform ambiguity and keep only those overlap with multiple isoform transcripts, suggesting the capability of MetaTX in dealing with features with high degree of isoform ambiguity (see **Supplementary Figure S3**). In another case study, we showed a different RNA modification mark, m^5C RNA methylation, is enriched at the 5'UTRs in human (**Supplementary Figure S4)**.

3.3 MetaTX package

An R package implementing the proposed MetaTX model was developed for estimating and visualizing the distribution density of mRNA-related features (d_c) along the standard mRNA model. MetaTX requires genome-based locations of mRNA-related features and the associated transcriptome annotations in TxDb format (Lawrence, et al., 2013). The standardized coordinates of mRNA-related features are calculated for the 5'UTR, CDS and 3'UTR regions, with each region is divided into a user-defined number of bins (Default 10) with equal length. Including the promoter or the tail DNA regions is optional.

The MetaTX package has a few useful functions. The `remapCoord` function calculates which bin and which component a particular feature overlaps with according to the genome-based coordinates (the O matrix). The `metaTXplot` function returns a density plot of the input feature set on a standard mRNA model, which supports customized relative length of different mRNA components during visualization, e.g., a shorter 5'UTR and longer CDS (see **Supplementary Figure S5** and **Supplementary Figure S6**). The package also provides an `isoformProb` function that can return the probabilities of a particular feature being located on different isoforms ($p_{s,t,c}$). The newly developed MetaTX R package is freely available at the GitHub repository(https://github.com/yue-wang-biomath/MetaTX.1.0) with examples and detailed documentation.

4 Conclusions

The distribution of a biological feature strongly indicates its functions, and is often the first step when characterizing its functional relevance. To date, a number of studies have been proposed for distribution analysis of mRNA-related features, but none of them provided a quantitative formulation for the problem of concern.

We proposed here the first rigorous statistical model, MetaTX, together with its EM solution for estimating the distribution of mRNA-related features in the presence of isoform ambiguity and differential composition among mRNAs. MetaTX was tested on both simulated data and RNA N^6-methyladenosine data derived from different high-throughput sequencing approaches, and demonstrated stable performance with more prominent and reasonable distribution patterns for all the datasets tested. An open source R package was developed for estimating and sketching a global view of mRNA-related features along the standard mRNA model. We believe that MetaTX should make a useful tool for studying the distribution and functions of mRNA-related biological features, especially for mRNA modifications such as N^6-methyladenosine.

From modeling perspective, MetaTX model is substantially different from existing approaches for resolving the ambiguity in RNA data, which usually relies on the fact that the read distribution is uniform on a transcript. On the contrary, MetaTX model relied on the non-uniform distribution of mRNA-related features on the entire transcripts, i.e, the tendency of the features to be enriched or depleted at different transcript coordinates.

For the work reported here we have not discussed: the possibility of an mRNA feature overlapping with multiple bins of the same transcript, the filtering of highly noisy features mapped to too many transcripts to improve data quality, the possibility of multiple features located on different transcripts being mapped to the same genome-based coordinate, or the possibility of incomplete or even incorrect transcriptome annotation. All of these will have a profound impact on the analysis results obtained using the MetaTX method. Meanwhile, the EM algorithms do not guarantee to converge to the global optimum, especially on a small dataset.

It is worth noting that there may be other interesting landmarks on transcripts. For example, it was reported previously that 70% of m^6A sites were found around the 3'UTR's last exons (Ke, et al., 2015), making the last exon also of interests. However, because the starting position of the last exon can appear before or after the stop codon of the corresponding

transcripts, the last codon cannot be integrated into the current MetaTX model. Nevertheless, it should be fairly straightforward to capture the distribution patterns with respect to new landmarks by altering the transcriptome annotations, e.g., for the MetaTX R package to acknowledge the last exon as the new landmark of interests, we may simply alter the transcriptome annotation by labeling the last exon-exon junctions as the 'pseudo stop codons'.

We discussed in this work the application of MetaTX model to base-resolution features only. It should be fairly straightforward to extend the model to cover wider features such as microRNA binding sites. However, more complicated weighting strategy will be necessary for dealing cases, e.g., an mRNA feature of 250nt width has 100nt overlap with one transcript and 150nt overlaps with another transcripts. Furthermore, although the MetaTX model is fairly efficient and can handle transcriptome-wide feature sets, it cannot be applied directly to millions of features due to the computational complicity of the model. Additional work will be needed to allow the model correct raw sequence alignments generated from high throughput RNA sequencing experiment.

Additionally, although our model was originally aimed at deciphering the overall distribution pattern of mRNA-related features, it is noteworthy that it has explicitly resolved the isoform-specific belongings of mRNA-related features through the calculation of the parameter $p_{s,t,c}$, i.e., a by-product of the MetaTX model is the probabilities of a particular feature being located on different isoforms; however, the accuracy and reliability of these estimates remains to be examined and tested.

Funding

National Natural Science Foundation of China [31671373]; XJTLU Key Program Special Fund [KSF-T-01]. This work is partially supported by the AI University Research Centre through XJTLU Key Programme Special Fund (KSF-P-02).
Conflict of Interest: none declared.

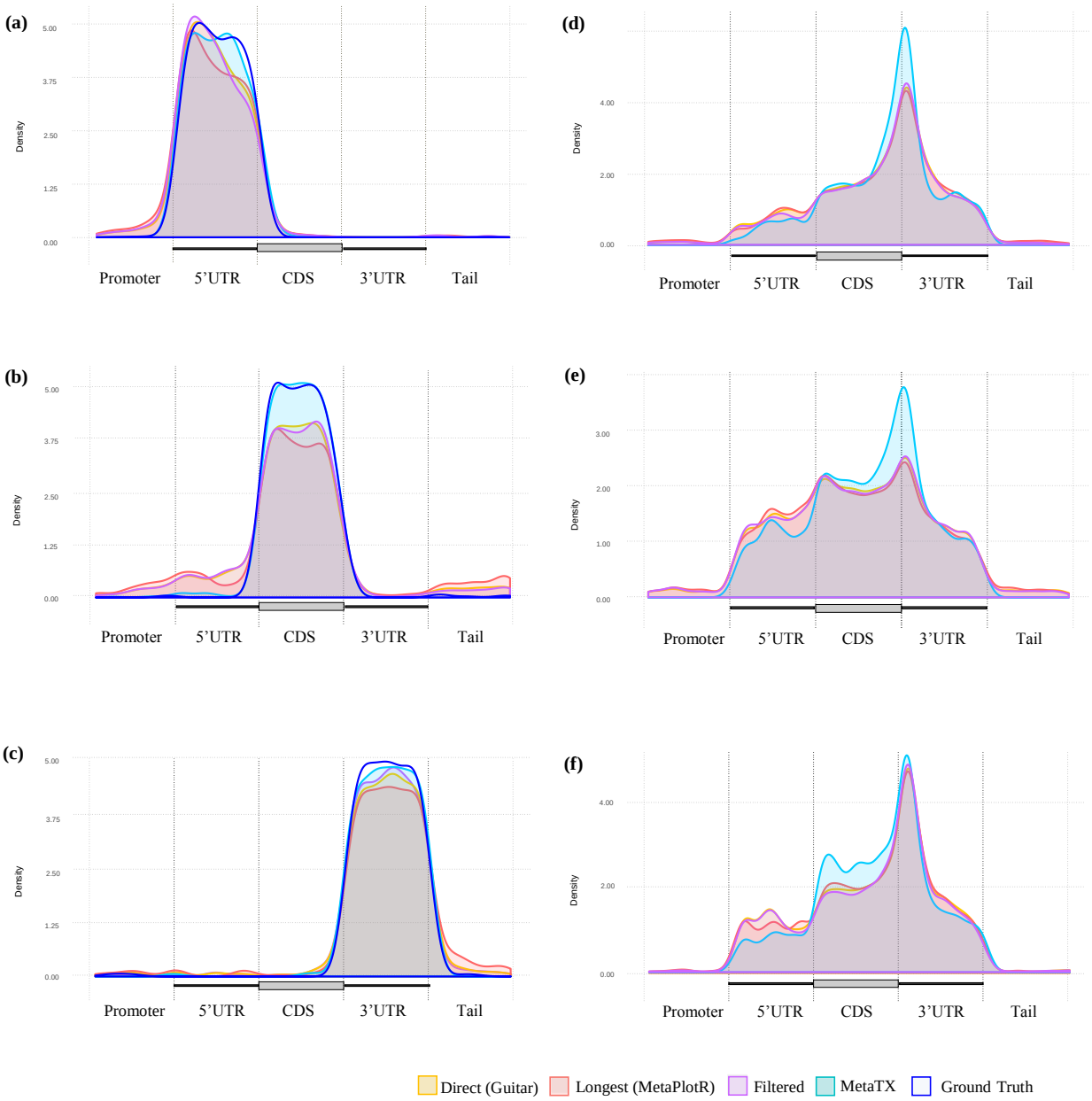

Fig. 3. Visualization of the simulated and the real data. The yellow curves are the direct estimates, the red curves are the longest estimates, the purple curves are the filter estimates and the blue curves are the estimates via the MetaTX. Plots (a)-(c) show the estimates of simulated data for 5'UTR, CDS and 3'UR. The dark blue curves are the ground truth distribution of each simulated case. Plots (d)-(e) show the estimated distribution of m6A RNA methylation sites derived from three different techniques, which respectively corresponds to the miCLIP-seq dataset, the PA--m^{6}A-seq dataset and the m^{6}A-seq dataset. We show here the smoothed distribution curves. An example of the original bin-based density plot is shown in **Supplementary Figure S1**.

References

Barski, A., *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* 2007;129(4):823-837.

Beauparlant, C.J., *et al.* Metagene profiles analyses reveal regulatory element's factor-specific recruitment patterns. *PLoS computational biology* 2016;12(8):e1004751.

Chen, K., *et al.* High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. *Angewandte Chemie (International ed. in English)* 2015;54(5):1587-1590.

Cui, X., *et al.* Guitar: An R/Bioconductor Package for Gene Annotation Guided Transcriptomic Analysis of RNA-Related Genomic Features. *BioMed research international* 2016;2016:8367534-8367534.

Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 1977;39(1):1-22.

Dominissini, D., *et al.* Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc* 2013;8(1):176-189.

Dominissini, D., *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;485(7397):201-206.

Dominissini, D., *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;485(7397):201-206.

Fustin, J.-M., *et al.* RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* 2013;155(4):793-806.

Geula, S., *et al.* Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* 2015;347(6225):1002-1006.

Ginestet, C. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2011;174(1):245-246.

Ke, S., *et al.* A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes & development* 2015;29(19):2037-2053.

Kundaje, A., *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research* 2012;22(9):1735-1747.

Lawrence, M., *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;9(8):e1003118-e1003118.

Linder, B., *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 2015;12(8):767-772.

Liu, H., *et al.* MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. *Nucleic Acids Res* 2018;46(D1):D281-D287.

Liu, N., *et al.* N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 2015;518(7540):560-564.

Liu, Q. and Gregory, R.I. RNAmod: an integrated system for the annotation of mRNA modifications. *Nucleic Acids Res* 2019;47(W1):W548-W555.

Mauer, J., *et al.* Reversible methylation of m(6)A(m) in the 5' cap controls mRNA stability. *Nature* 2017;541(7637):371-375.

Meyer, K.D., *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 2012;149(7):1635-1646.

Meyer, K.D., *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 2012;149(7):1635-1646.

Olarerin-George, A.O. and Jaffrey, S.R. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. *Bioinformatics* 2017;33(10):1563-1564.

Pendleton, K.E., *et al.* The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* 2017;169(5):824-835.e814.

Schaefer, M., *et al.* RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res* 2009;37(2):e12-e12.

Schwartz, S., *et al.* Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Rep* 2014;8(1):284-296.

Shin, H., *et al.* CEAS: cis-regulatory element annotation system. *Bioinformatics* 2009;25(19):2605-2606.

Slobodin, B., *et al.* Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* 2017;169(2):326-337.e312.

Wang, X., *et al.* N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* 2015;161(6):1388-1399.

Wang, Y., *et al.* N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol* 2014;16(2):191-198.

Yan, Z., *et al.* txCoords: A Novel Web Application for Transcriptomic Peak Re-Mapping. *IEEE/ACM Trans Comput Biol Bioinform* 2017;14(3):746-748.