Published in final edited form as:
Cell Syst. 2017 May 24; 4(5): 568–574.e7. doi:10.1016/j.cels.2017.04.007.

PROBer Provides a General Toolkit for Analyzing Sequencing-based Toeprinting Assays

Bo Li[1], **Akshay Tambe**[2], **Sharon Aviran**[3], and **Lior Pachter**[4,5,6]

[1]Center for RNA Systems Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[2]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[3]Department of Biomedical Engineering and Genome Center, University of California, Davis, Davis, CA 95616

[4]Departments of Biology and Computing & Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA

SUMMARY

A number of sequencing-based transcriptase drop-off assays have recently been developed to probe post-transcriptional dynamics of RNA-protein interaction, RNA structure, and RNA modification. Although these assays survey a diverse set of 'epitranscriptomic' marks, we term them as toeprinting assays since they share methodological similarities. As such, their interpretation is predicated on addressing similar computational challenge: how to learn isoform-specific chemical modification profiles in the face of complex read multi-mapping. We introduce PROBer, a statistical model and associated software, that addresses this challenge for the analysis of toeprinting assays. PROBer takes sequencing data as input and outputs estimated transcript abundances and isoform-specific modification profiles. Results on both simulated and biological data demonstrate that PROBer significantly outperforms individual methods tailored for specific toeprinting assays. Since the space of toeprinting assays is ever expanding and these assays are likely to be performed and analyzed together, we believe PROBer's unified data analysis solution will be valuable to the RNA community.

eTOC Blurb

[5]Correspondence: lpachter@caltech.edu.
[6]Lead Contact

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

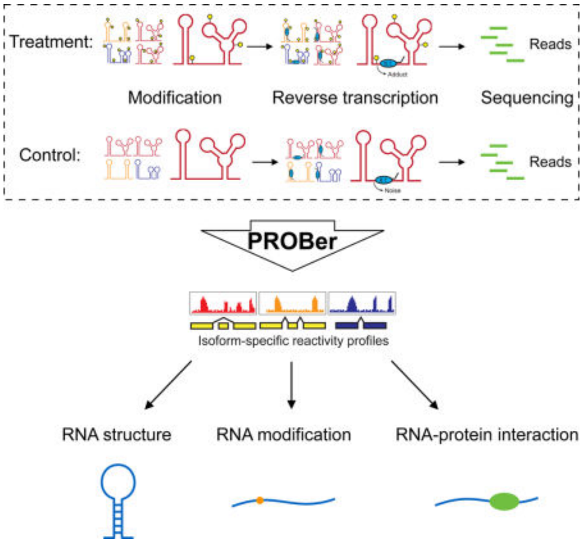
AUTHOR CONTRIBUTIONS

Methodology, B.L., A.T., S.A., and L.P.; Software and analysis, B.L.; Writing – Original Draft, B.L., A.T., S.A., and L.P.; Writing – Review & Editing, B.L., A.T., S.A., and L.P; Funding Acquisition, L.P.; Supervision, S.A. and L.P.

ADDITIONAL RESOURCES

We implemented a Snakemake (Köster and Rahmann, 2012) workflow that can be used to replicate all our analyses: https://github.com/pachterlab/PROBer_paper_analysis. PROBer parameters set for sepcific experimental protocols could be found at https://github.com/pachterlab/PROBer_paper_analysis/blob/master/exp/README.md.
Method S1. Related to STAR Methods. Detailed description of PROBer model and methods.

PROBer is a statistical method that could learn isoform-specific chemical modification profiles from sequencing-based transcriptase drop-off assays. It could be used to detect RNA structure, RNA modification & RNA-protein interaction, and thus is valuable to RNA research.



INTRODUCTION

While much of the control of gene expression occurs via transcriptional regulation, it is becoming increasingly clear that post-transcriptional regulation also plays a key role in modulating expression products (Schwanhäusser et al., 2011). Several mechanisms contribute to this phenomenon, including covalent posttranscriptional chemical modification of RNA molecules (Roundtree and He, 2016), protein binding and the assembly of higher-order ribonucleoprotein complexes (Glisovic et al., 2008), and the ability of RNA molecules to fold into and switch between intricate 2- and 3- dimensional folds (Mortimer et al., 2014; Schwanhäusser et al., 2011; Wan et al., 2011). Understanding both the expression level and the ‘meta-information’ (post-transcriptional marks) associated with a given transcript can shed light not only on the functions that an individual sequence performs, but also on the cellular pathways that it participates in and controls.

Recent advances in massively parallel DNA sequencing have enabled the transcriptome-wide investigation of several ‘epitranscriptomic’ layers. Although the specifics of the assays differ widely depending on what is being measured (and how), there are several experiments that share a common theme. We term these experiments ‘toeprinting’ (Hartz et al., 1988) by high-throughput sequencing as they share a similar workflow (Figure 1A): chemically modifying RNAs to encode a signal of interest, decoding these chemical signals by reverse transcriptase drop-off, and sequencing and mapping the resulting cDNA toeprints to recover the chemical modification signatures.

Within this framework, iCLIP and eCLIP protocols (König et al., 2010; Van Nostrand et al., 2016) explore RNA-protein interactions through crosslinking, SHAPE and DMS probing (Ding et al., 2014; Rouskin et al., 2014; Spitale et al., 2015; Talkish et al., 2014) explore

RNA secondary structure by using selective chemical probes to modify and mark unpaired flexible nucleotides, and Pseudo-seq (Carlile et al., 2014) detects RNA pseudouridylation by utilizing a reagent which specifically forms adducts at pseudouridine sites (Ψs). Furthermore, similar assays have also been developed to detect G-quadruplex structures (Kwok et al., 2016) and RNA 2′-O-methylation (Incarnato et al., 2017), indicating that the space of toeprinting assays is continuing to expand.

In each of these experiments, the upstream chemical modification is widely variable, but the library preparation and sequencing techniques are essentially the same: reverse transcription in a manner where cDNAs preferentially terminate at the sites of chemical modification, adaptor ligation to the site of reverse transcriptase drop-off, and PCR amplification followed by sequencing of the resulting cDNA library. Additionally, the number of characterizable epitranscriptomic marks is ever expanding, as are the associated chemical toolkits (Dominissini et al., 2012; Dominissini et al., 2016; Sakurai et al., 2014). As a result, toeprinting by high-throughput sequencing is becoming an essential tool for studying post-transcriptional regulation.

A key step in analyzing toeprinting experiments is to accurately learn reverse transcriptase drop-off profiles from the sequence data. These profiles are subsequently used to infer, for example, sequence motifs, secondary structure predictions, or sites of post-transcriptional chemical modification. Each sequenced read produced by the experiments potentially contains multiple layers of valuable information about both chemical modification frequencies as well as about the identity and abundance of RNA transcripts. The ability to make full use of this information becomes the key for accurate estimation of drop-off profiles and requires conjointly addressing associated bioinformatics problems including the conflation of read counts by reverse transcriptase noise, variable transcript abundances, and read mapping ambiguity. However, to date, the proposed approaches address these problems separately and therefore only yield suboptimal solutions (Choudhary et al., 2017).

RESULTS

Bioinformatics challenges

Accurately determining the transcript abundances and drop-off profiles in transcriptome-wide toeprinting experiments is complicated by several factors (Figure 1B) (Aviran and Pachter, 2014). Such experiments face a problem that is fundamental in RNA-Seq: reads align ambiguously to multiple transcripts, and appropriately handling ambiguously mapped reads (which can represent a significant proportion of alignable reads in such experiments, see Table S1) is imperative to correctly learning transcript abundances (Bray et al., 2016; Li and Dewey, 2011; Li et al., 2010; Roberts and Pachter, 2013; Trapnell et al., 2010). Incorrectly allocating multi-mapping reads adversely affects the estimated abundances of not only the transcripts that the reads were misallocated to, but also abundance estimates of related transcripts.

In toeprinting experiments, the multi-mapping problem is further exacerbated by the fact that accurate estimation of the RNA chemical modification probabilities depends on both correctly allocating multi-mapped reads, and deconvolving chemical modification profiles from adduct-independent noisy reverse transcriptase (RT) drop-off. All of these factors are inter-related and poor estimation of any one of them may significantly skew estimates of the others. Yet all of these factors must be accounted for to quantitatively estimate modification rates.

The PROBer software

To address the computational challenges associated with the interpretation and analysis of toeprinting assays, we have developed a statistically rigorous approach that serves the dual purpose of unifying these assays via a shared computational framework, while providing an inference approach that is robust to small variances in experimental protocol. Our methods are implemented in software, termed PROBer, that is based on a statistical model to jointly infer transcript abundance and modification probabilities, as well as several other parameters (see STAR★METHODS and Figure S1A) and was developed by building on previous work on RNA-Seq (Bray et al., 2016; Li et al., 2010; Li and Dewey, 2011; Roberts and Pachter, 2013; Trapnell et al., 2010), as well as models for simpler structure-probing SHAPE-Seq experiments (Aviran et al., 2011a; 2011b) where transcript abundance is not a confounding factor. The PROBer model assumes that the input data consists of raw reads (either single- or paired- end) obtained separately from a chemically treated sample, containing information about modification probabilities, and from a mock-treated control, informing about noise parameters. It assumes that cDNA fragments were generated by first selecting a transcript from the transcriptome (according to its abundance and length), randomly priming (or fragmenting) that transcript, and primer extending one nucleotide at a time. At each nucleotide encountered by the reverse transcriptase in this process, there is some probability of terminating the reverse transcription, due to modification, RT noise, primer collision, or encountering the end of the template fragment. A cDNA fragment generated by this process is observed as sequenced read if it passes a size-selection filter, which is dependent on the fragment length. From this the extent to which all the parameters in the experiment are inter-related becomes clear.

We implemented an Expectation-Maximization algorithm (Dempster et al., 1977) in PROBer to infer the parameters of the model (see STAR★METHODS). In many cases it is of interest to have transcript-specific modification profiles rather than genes; indeed in structure probing experiments it is meaningless to consider the secondary structure of a gene rather than that of a specific transcript isoform. For these reasons we focused on modeling chemical modification at the isoform–rather than the gene–level.

The PROBer workflow, shown schematically in Figure 1C, begins with a set of read alignments (separately for the chemically-treated experiment and the untreated control). Starting with initial parameter estimates, reads are allocated to transcripts based on both abundance and modification parameters. The allocated read ‘pseudocounts’ are then used to estimate *maximum a posterior* (MAP) modification probabilities as well as RT noise and maximum likelihood (ML) estimates of transcript abundances. These steps are repeated until convergence. Although PROBer implements inference with respect to a complex model, it is practical for the analysis of standard toeprinting datasets (Table S2).

PROBer outperforms alternative approaches in profiling RNA structures

To test the accuracy of PROBer on structure-probing experiments, we investigated its performance on both simulated and experimental data. In simulations, we generated a dataset in a manner consistent with the chemical mapping protocol (see STAR★METHODS) and attempted to recover parameter estimates from these simulated reads alone. At a global scale, PROBer yielded significantly improved parameter estimates when compared with alternative approaches, including StructureFold (Tang et al., 2015), Mod-seeker (Talkish et al., 2014), and icSHAPE (Spitale et al., 2015), using Pearson's correlations (Figure 2A). These results also hold for Spearman's rank correlations (Figure S2) and were representative of multiple simulations (Figure S3). In addition, because PROBer takes structure information into consideration, it is able to provide better transcript abundance estimates from structure probing datasets when compared with conventional RNA-Seq analysis software (Figure S4).

PROBer's performance at recovering secondary structure constraints for transcripts with moderate expression levels (between 100 and 1000 TPM) vastly improves on alternative approaches at the highest expression levels (greater than 10,000 TPM). This result indicates that PROBer requires approximately 90% less data (when compared to alternative approaches) to produce structural estimates of equal or better accuracy. As transcript abundances follow an exponential distribution, a moderate improvement in the range of expression levels that yields useful structural constraints translates to a large increase in the number of transcripts that can be probed. Thus, PROBer allows the experimenter to access a larger fraction of the transcriptome at the same sequencing depth and experimental cost.

Since the chemical modification parameters (i.e. ground truth) used in simulations were learned from real data using PROBer, we were concerned that our simulations would artificially inflate the apparent performance of PROBer. We therefore included in our simulated transcriptome a set of control transcripts whose chemical modification profiles were measured by the related but orthogonal method SHAPE-MaP (Siegfried et al., 2014). Like SHAPE-Seq or DMS-Seq, the SHAPE-MaP assay measures RNA secondary structure using chemical probes to label unpaired or flexible positions; however this assay encodes these chemical marks as mismatches/sequencing errors (rather than RT drop-off/toeprinting), insulating our simulations from any (unknown) systematic biases from the protocol itself. These transcripts (see STAR ★ METHODS) served as a digital "spike-in", allowing us to verify that our simulated experiments were not biased with respect to the simulation parameters. The accuracy of PROBer was confirmed by these digital spike-in experiments (Figure S5).

We further tested whether this improvement was also evident in real datasets by examining modification probability estimates for ribosomal RNAs, which have well-characterized structures (Cannone et al., 2002). We calculated precision-recall (PR) and receiver operating characteristic (ROC) curves on two yeast structure-probing data sets (Hector et al., 2014; Talkish et al., 2014) that adopted different library preparation methods (random priming or fragmentation and ligation; paired-end or single-end). We performed these analyses using crystallographically informed solvent-accessible secondary structure as a ground truth and demonstrated that PROBer was better able to estimate DMS chemical modification profiles when compared with alternative approaches (Figures 2B and S6). We also compared PROBer with alternative approaches using ROC curves on available Arabidopsis and mouse structure-probing data and the area under curve (AUC) values were documented in Table S3.

Lastly, we compared PROBer with alternative approaches on predicting yeast ribosomal RNA structures. For each method, we first rescaled the reactivity profiles used to generate the PR curve into SHAPE constraints, that could be fed into the secondary structure prediction software RNAstructure (Deigan et al., 2009). Then we predicted secondary structures of yeast 18S and 25S rRNAs for each method and evaluated the predicted structures using commonly accepted measures such as sensitivity and positive predictive value (PPV) (Sloma and Mathews, 2015). In this evaluation (Figure 2C), as well as in a comparison (Table S4) with BUM-HMM (Selega et al., 2017), PROBer's performance was superior to the alternative approaches. In addition, we have demonstrated that PROBer is robust to priming biases with respect to modification probability estimation (Figure S10 and S11).

PROBer identifies more true Ψs than alternative approaches

To demonstrate PROBer's ability of identifying epitranscriptomic marks, we analyzed the Pseudo-seq data (Carlile et al., 2014) for pseudouridine detection. We used all known Ψ sites in ribosomal and small nucleolar RNAs as a ground truth, with which we compared PROBer estimated modification profiles. Precision-recall curve analysis of the Pseudo-seq data (Figure 3A) revealed that PROBer outperformed alternative approaches, including the Pseudo-seq method used in (Carlile et al., 2014) and structure-probing methods StructureFold, Mod-seeker, and icSHAPE that we previously compared with, for predicting Ψ. Importantly, PROBer was able to detect an experimentally validated pseudo-U site ($m^1acp^3Ψ1191$ in 18S rRNA) that was not detected by alternative approaches (Figure S7). This indicates that PROBer is capable of capturing biologically relevant information that would be otherwise lost.

PROBer extracts more information from eCLIP and iCLIP data than current common practice

Next, we tested PROBer on eCLIP and iCLIP data. The eCLIP protocol (Van Nostrand et al., 2016) is an improvement over the iCLIP protocol (König et al., 2010). Both protocols encode protein binding information in a toeprinting-type manner by crosslinking RNA to proteins and degrading the cross-linked protein by proteolysis. This leaves a short peptide fragment attached to the site on the RNA where it was cross-linked, and that can therefore cause RT drop-off. One major improvement of the eCLIP protocol is the inclusion of a sequenced input control, which is lacking in iCLIP experiment.

The eCLIP and iCLIP protocols differ from other toeprinting protocols in that the RNase degradation process produces fragments that are only around the crosslink sites. This results in sparse eCLIP or iCLIP read alignment to the genome and makes it hard for us to estimate transcript abundance and reduce background noise. Therefore, for eCLIP and iCLIP data, PROBer focuses only on appropriate allocation of multi-mapping read with a simpler model (Figure S1B).

We reanalyzed eCLIP data sets (Van Nostrand et al., 2016) for four RNA-binding proteins: RBFOX2, TARDBP, TRA2A, and PUM2. We in addition reanalyzed iCLIP data sets for two RNA-binding proteins: RBFOX2 (Van Nostrand et al., 2016) and hnRNP C (Zarnack et al., 2013). These five proteins have known canonical binding motifs that have been validated both in vitro and in vivo (Van Nostrand et al., 2016; Zarnack et al., 2013), providing an independent ground truth for our evaluation.

As expected, our analysis (Figure 3B and Table S6) of these datasets demonstrated that PROBer could be used to detect significantly more peaks than the common practice that only uses uniquely-mapping reads, while kept the percentage of peaks overlapping with a canonical motif roughly the same. To demonstrate our results are robust to the peak-calling threshold, we in addition plotted the performance of each method by varying the peak-calling threshold (Figure S9). Our results clearly demonstrate that multi-mapping reads contain valuable information and that the common practice of restricting analysis to unique mappings is suboptimal. Since PROBer was only used to allocate multi-mapping reads and thus was not involved in the peak calling process, our results should hold regardless of the chosen peak-calling algorithm. We also compared PROBer with a baseline method that distributes multi-mapping reads evenly to all aligned locations, and demonstrated that PROBer worked better on allocating multi-mapping reads than the baseline (Table S5 and S7).

DISCUSSION

We present PROBer, a statistically rigorous approach to quantify chemical modification profiles from transcriptome-wide sequencing data. PROBer contains a huge amount of model parameters. To robustly estimate these parameters, we assume the transcript abundances in the treatment and control experiments are the same, and impose beta distribution priors to the chemical modification and RT noise profiles.

We have evaluated PROBer's performance with three diverse chemical modification protocols, as well as a variety of library preparation protocols. In each of these cases, PROBer outperformed alternative approaches in analysis of the data. As it is becoming clear that a systems-wide view of such post-transcriptional regulation processes is highly informative, we believe that multiple of these chemical modification toeprinting protocols will be performed within the same study. As such a unified pipeline such as PROBer is even more valuable.

PROBer is freely available with open-source at http://pachterlab.github.io/PROBer. All experiments can be replicated using the Snakemake workflow at https://github.com/pachterlab/PROBer_paper_analysis.

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resource sharing may be directed to, and will be fulfilled by the corresponding author, Dr. Lior Pachter at California Institute of Technology (lpachter@caltech.edu).

METHOD DETAILS

PROBer's generative probabilistic model—We model sequencing-based toeprinting experiments using a generative probabilistic model (Figure S1A). Our model combines the strengths of previous works on modeling RT drop-off (Aviran et al., 2011b; Aviran and Pachter, 2014) and quantifying transcript abundance (Bray et al., 2016; Li et al., 2010; Li and Dewey, 2011; Roberts and Pachter, 2013; Trapnell et al., 2010). The key parameters in our model include the relative abundances for the set of transcripts in the sample, as well as modification probabilities, and RT noise probabilities for each site on a transcript. In order to reduce the number of parameters we have to estimate, we assume the abundances in the modification-treated experiment are the same as abundances in the mock-treated experiment. We in addition assume that the reference contains M known transcripts and we number these transcripts from 1 to M.

To generate a read from the modification-treated experiment, we first pick a transcript at a rate proportional to the product of transcript abundance and length. We denote this rate by α_i, where i is the transcript number. Then we choose a priming site uniformly across all valid priming sites in the transcript. We denote the total number of available priming sites by l'_i. Once we have a priming site, reverse transcription starts in 3′ to 5′ direction. At each site j, there is a probability that RT stops due to either chemical modification (denoted by β_{ij}) or background noises such as RT natural drop-off, primer collision or reaching the end of a fragment (denoted by γ_{ij}). Once the RT stops, a cDNA fragment is generated. Thus, the probability of generating a cDNA fragment of length l, priming at j, and from transcript i is

$$\alpha_i \cdot \frac{1}{l'_i} \cdot \left(1-\left(1-\beta_{i,j-l}\right)\left(1-\gamma_{i,j-l}\right)\right) \prod_{k=j-l+1}^{j-l_p} (1-\beta_{ik})(1-\gamma_{ik}).$$

The term l_p in the above equation is the random primer length. In the Ding et al. protocol, this term is equal to 6; however if RNA fragmentation-based protocols were used, this number would be 0.

The next step is to decide if the obtained fragment passes the size selection. If not, this fragment will not be sequenced and therefore considered hidden. Otherwise, a sequence read will be produced according to our sequencing error model $\mathcal{E}$. Our sequencing error model $\mathcal{E}$ can generate either single-end or paired-end reads and allows both substitution and indel errors to occur during the sequencing step. Since the error model $\mathcal{E}$ is more complicated, we omit formulae for generating read sequences from $\mathcal{E}$. For details, please refer to Method S1, section 1.3 and section 2.

To generate a read from the mock-treated experiment is similar, excepting that the chemical modification probabilities are not involved. Thus the probability of generating a similar cDNA fragment becomes

$$\alpha_{\mathrm{i}} \cdot \frac{1}{\mathrm{l}_{\mathrm{i}}^{\prime}} \cdot \gamma_{\mathrm{i,j-l}} \prod_{\mathrm{k=j-l+1}}^{\mathrm{j-l_p}} (1-\gamma_{\mathrm{ik}}).$$

Our generative model is applicable to fragment-based probing protocols (Carlile et al., 2014; Rouskin et al., 2014; Spitale et al., 2015; Talkish et al., 2014) as well. We just need to replace the assumption of uniform priming with the assumption of uniform fragmentation. For more details about our generative model, please refer to Method S1, section 2.

PROBer parameter estimation and overfitting avoidance—Our goal is to estimate toeprinting parameters and relative abundances in the sample. Toeprinting parameters include both modification probabilities (βs) and RT noise probabilities (γs) per transcript site. Because our model contains a lot of parameters, we use two approaches to avoid overfitting. First, we reduce the total number of parameters by assuming that the transcript abundances in treatment and control are the same. Secondly, we introduce a beta distribution prior for each toeprinting parameter (β and γ) and calculate maximum a posteriori (MAP) estimate instead of maximum likelihood (ML) estimate for the toeprinting parameter. In our model, all βs share a same set of tunable hyper-parameters for their beta distribution priors and all γs share another set of hyper-parameters. By default, we set these two sets of hyper-parameters the same such that the mode of the beta prior is 0.0001. We calculate ML estimates for transcript abundances.

We have three types of hidden data. First, due to alignment ambiguity we cannot be sure about which transcript a read originates from; we can only infer a set of highly possible origins for the read using its alignments. Second, for data sets with single-end reads, we cannot observe the full cDNA fragment from a read; therefore we have to guess the priming or fragmentation site for each single-end read. Lastly, if a cDNA fragment does not pass the size selection, we cannot observe a read from it. For reasons explained in Method S1, section 2.2, we only consider the first two types of hidden data.

We use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to learn above model parameters. The workflow of our EM algorithm is shown in Figure 1C. In the E step, we interpolate the hidden data – the locations of multi-mapping reads and the priming or fragmentation sites of single-end reads – given the estimated abundances and toeprinting parameters. In the M step, we calculate the ML and MAP estimates based on both the observed data and interpolated hidden data. The E and M steps are repeated until convergence. Method S1, section 3 provides a detailed discussion about our EM algorithm.

Assessing estimation variation for PROBer-estimated modification probabilities—PROBer can provide intervals representing estimation variations of modification probabilities (βs) for any transcript of interest. These intervals account for variations in multi-mapping read allocation and toeprinting parameter estimation. PROBer produces these intervals using a three-step procedure. First, we quantify variation in multi-mapping read allocation. We fix the toeprinting parameters (β and γ) as the MAP estimates produced by PROBer and sample "true" alignments for reads in treatment and control using a collapsed Gibbs sampler. The Gibbs sampler is similar to the one used in RSEM (Li and Dewey, 2011). We run the sampler for 200 rounds in the burn-in period and then keep 20 sampled "true" alignment settings by selecting one sampled setting every 10 rounds. A "true" alignment setting is defined as the set of sampled "true" alignments for every read in treatment and control. Secondly, for the transcript of interest, we produce 50 bootstrapped samples (Efron and Tibshirani, 1993) from the "true" alignments belong to this transcript for each of the 20 "true" alignment settings. This step accounts for the variation in estimating modification parameters. Lastly, we run PROBer on each of the 1,000 bootstrapped samples and pool the PROBer estimates together to produce 95% intervals for every position in the transcript of interest.

PROBer's iCLIP and eCLIP model—Because protein-binding signals are sparse in the genome and because these signals could occur in both exons and introns, it is challenging to either model modification reactivity per nucleotide or estimate transcript abundance from iCLIP or eCLIP data sets. Therefore, PROBer focuses on allocating multi-mapping reads for iCLIP and eCLIP data with a simpler generative model (Figure S1B). To generate an iCLIP or eCLIP read, PROBer first picks a crosslink site and then generates the read sequence according to a sequencing error model. PROBer uses an Expectation-Maximization-Smoothing (EMS) algorithm (Silverman et al., 1990), which is similar to Chung et al.'s work on ChIP-Seq data (Chung et al., 2011), to allocate multi-mapping reads. Please refer to Method S1, section 5 for more details.

QUANTIFICATION AND STATISTICAL ANALYSIS

Transcriptome and genome references

Arabidopsis thaliana: We downloaded the latest genome and gene annotation (TAIR10) from The Arabidopsis Information Resource. Following Ding et al. (Ding et al., 2014), we extracted every mRNA, rRNA, tRNA, ncRNA, snRNA, miRNA, and snoRNA annotated in the GFF3 file. We also discovered and thus removed 568 duplicate sequences. In addition, we found two copies of 18S rRNA with minor differences and no 25S rRNA (but a subsequence of it, AT2G01021.1) in the extracted sequences. Thus, we added 25S rRNA sequence from the RNA structure database (Cannone et al., 2002) and removed one copy of 18S rRNA, AT3G41768.1, and the 25S subsequence, AT2G01021.1. The final reference consists of 36,264 transcripts in total.

Saccharomyces cerevisiae: We downloaded the genome (R64-1-1) and gene annotation (build R64-1-1.84) from Ensembl. After removing duplicate sequences, the final reference consists of 6,841 transcripts.

Mus musculus: We downloaded the genome (GRCm38) and gene annotation (build GRCm38.74) from Ensembl. The annotation contains no 18S or 25S rRNAs, and 353 variants of 5S rRNA. We added 18S sequence from the RNA structure database and removed all but one variant of 5S rRNA. We could not add 25S sequence because it is not included in the RNA structure database (Cannone et al., 2002). After removing duplicate sequences, the final reference consists of 93,362 transcripts.

Homo sapiens: We downloaded the human genome (GRCh38) from Ensembl.

Escherichia coli: We downloaded the E. coli 16S rRNA sequence used in (Poulsen et al., 2015) at http://people.binf.ku.dk/~jvinther/data/HRF-Seq/ecoli_rRNA.fa.

RNA structure probing data—Structure probing data from (Ding et al., 2014) were downloaded through Sequence Read Archive (SRP027216). This data set contains two biological replicates and we pooled them together. Data were pre-processed following (Ding et al., 2014), which includes removing ssDNA linker and trimming adapter sequence. We used cutadapt v1.10 (Martin, 2011) to trim adapters for all data sets mentioned in this paper. The pre-processed data contain 117,242,295 and 81,596,350 single-end reads in modification-treated and mock-treated experiments respectively.

Structure probing data from (Hector et al., 2014) were downloaded through Gene Expression Omnibus (GSE52878). We selected and pooled together only the three in vitro DMS biological replicates (GSM1277430-GSM1277435). Data were pre-processed following (Hector et al., 2014), which includes trimming adapter sequence and removing PCR duplicate using both unique molecule identifier (UMI) and read sequence information. The pre-processed data contain 29,660,269 and 18,403,322 paired-end reads in modification-treated and mock-treated experiments respectively.

Structure probing data from (Talkish et al., 2014) were downloaded through Sequence Read Archive (SRP029192). Only wild-type data were used and the two biological replicates were pooled together. Data were pre-processed following (Talkish et al., 2014), which includes trimming adapter sequence and 5′-end untemplated nucleotides. The pre-processed data contain 7,729,251 and 9,199,721 single-end reads in modification-treated and mock-treated experiments respectively.

Structure probing data from (Spitale et al., 2015) were downloaded through Gene Expression Omnibus (GSE64169). We used data from biological replicate 2 of v6.5 mouse ES cells, which consist of three conditions: mock-treated, in vitro modification-treated, and in vivo modification-treated. Data were pre-processed following (Spitale et al., 2015), which includes trimming adapter sequence and removing PCR duplicate. After pre-processing, the three conditions contain 668,854,264, 241,988,034, and 379,309,553 single-end reads respectively.

Structure probing data from (Poulsen et al., 2015) were downloaded through http://people.binf.ku.dk/~jvinther/data/SHAPES-Seq/. We only used the E. coli data, which consists of three conditions: mock-treated, modification-treated, and modification-treated followed by biotinylation selection (SHAPES). Data were pre-processed following (Poulsen et al., 2015), which includes trimming adapter sequence and 5′-end untemplated nucleotides. Afther pre-processing, the three conditions contain 1,939,723, 2,930,176, and 7,831,047 paired-end reads respectively.

Pseudouridine detection data—Pseudo-seq data from (Carlile et al., 2014) were downloaded through Gene Expression Omnibus (GSE58200). Following advice of the authors, samples GSM1403085 and GSM1403086 were picked as mock-treated experiments and samples GSM1403087 and GSM1403088 were picked as modification-treated experiments. Adapter sequences were trimmed as documented in (Carlile et al., 2014). The resulting pre-processed data contain 31,103,632 and 39,167,224 single-end reads in modification-treated and mock-treated experiments respectively.

RNA-protein interaction data—Biological replicate 1 of RBFOX2 eCLIP data from (Van Nostrand et al., 2016) were downloaded through Gene Expression Omnibus (GSE77634: GSM2055432 and GSM2055433). In addition, biological replicate 1 of TARDBP, TRA2A, and PUM2 eCLIP data from (Van Nostrand et al., 2016) were downloaded through ENCODE (ENCLB754PZD, ENCLB615MWL; ENCLB985UUR, ENCLB547DSI; ENCLB175BHT, ENCLB809FVC). Data were pre-processed following (Van Nostrand et al., 2016), which includes trimming adapter sequence, removing PCR duplicate, and filtering out reads that align to a customized set of human repetitive elements used in (Van Nostrand et al., 2016). The pre-processed data for these four proteins contain 35,708,030 & 9,765,434, 3,982,157 & 4,035,137, 5,938,623 & 4,224,764, and 3,788,923 & 19,153,400 paired-end reads in CLIP-treated and mock-treated experiments respectively.

We also analyzed two iCLIP data sets: run 2 of RBFOX2 iCLIP data from (Van Nostrand et al., 2016) were downloaded through Sequence Read Archive (SRR3147675) and biological replicate 1 of hnRNP C iCLIP data from (Zarnack et al., 2013) were downloaded through ArrayExpress (E-MTAB-1371). These two iCLIP data sets were pre-processed the same way as we did for the eCLIP data. The pre-processed iCLIP data for RBFOX2 and hnRNP C contain 17,424,135 and 8,608,578 single-end reads respectively.

Alternative methods that PROBer compared with—For probing RNA structure, we compared PROBer with three alternative methods: StructureFold (Tang et al., 2015), Mod-seeker (Talkish et al., 2014), and icSHAPE (Spitale et al., 2015). Because StructureFold can only be run via Galaxy, which makes it unsuitable for benchmarking on large data sets, we re-implemented it according to (Tang et al., 2015). In addition, since Mod-seeker only calculates modification intensities at gene level, and we focus on isoform-level modification signals, we re-implemented its formula so that we can use it to estimate isoform-level intensities. We have compared our re-implementations with the original StructureFold and Mod-seeker software on Talkish et al. and Carlile et al. yeast data and confirmed that our re-implementations faithfully reflect how the original software works. We have successfully run the icSHAPE software on all data sets used in this paper.

For detecting pseudouridine sites, we in addition compared PROBer with Pseudo-seq, the method used in (Carlile et al., 2014). Because there is no publicly available software implementing the Pseudo-seq method, we re-implemented it according to (Carlile et al., 2014).

We compared PROBer with BUM-HMM (Selega et al., 2016) on the Hector et al. paired-end yeast data, which contain 3 biological replicates. Majority of the reads in this data set are from 18S and 25S rRNAs. To calculate the BUM-HMM-estimated modification intensities for 18S, we used all three replicates. For 25S, we only fed BUM-HMM with reads from biological replicate 2 and 3 since biological replicate 1 contains no 25S reads and BUM-HMM would crash if we use all three replicates.

Alignment criteria—To make sure that the differences in performance between methods are not due to different alignment criteria (e.g. number of mismatches), we fed each method with alignments produced beforehand using the same alignment criteria. We used Bowtie v1.1.2 (Ben Langmead et al., 2009) to align single-end reads and Bowtie 2 v2.2.9 (Ben Langmead and Salzberg, 2012) to align paired-end reads. Because structure-probing protocols are strand-specific, we only aligned reads to the forward strand. Regarding to alignment criteria, we required at most 3 mismatches in each qualified alignment for the (Ding et al., 2014) Arabidopsis data. For all other data sets, we used Bowtie or Bowtie 2's default setting. Regarding to the number of alignments reported per read, PROBer asked aligners to report all qualified alignments of a read. In addition, reads with more than 200 qualified alignments were filtered out. StructureFold and icSHAPE used all qualified alignments in the best stratum (least number of mismatches in either entire read or the "seed" region). Mod-seeker and BUM-HMM used only the best single qualified alignment. These alignment settings were chosen according to the papers describing each method.

Setting PROBer parameters—PROBer's protocol-specific options, such as –primer-length, –size-selection-min, –size-selection-max, and –read-length, were set differently according to the characteristics of each protocol. (Spitale et al., 2015) used biotin to selectively enrich structural signals in modification-treated experiments. This step significantly reduces the background noise contained in the modification-treated channel and also makes it hard to interpret the relationship between mock-treated and modification-treated channels. Thus, for Spitale et al. data, we only used modification-treated data as PROBer's input. For further details, please refer to our Snakemake (Köster and Rahmann, 2012) workflow.

Simulation of structure-probing experiments and digital spike-in experiments —To assess the variability of the simulation, we simulated two sets of 30 million 37 nt single-end reads in both the modification-treated and mock-treated experiments, using the generative model described before. The model parameters used in the simulation, including ground truth modification (β) and RT drop-off (γ) probabilities, were learned from the Ding et al. structure-probing data by running PROBer. To access if structure information can affect RNA-Seq quantification process (Figure S4), we in addition simulated 30 million 37 nt single-end reads using the RSEM simulator (Li and Dewey, 2011) (which ignores structure information) with the same simulation parameters. For digital spike-in experiments, our transcriptome was augmented with sequences of nine RNAs that have SHAPE-MaP reactivities available (Siegfried et al., 2014): tRNA^{Phe} (E. coli, 76 nt), TPP riboswitch (E. coli, 79 nt), 5S rRNA (E. coli, 120 nt), 16S rRNA(E. coli, 1542 nt), 23S rRNA (E. coli, 2904 nt), IRES domain (HCV, 336 nt), Group II intron (O. iheyensis, 412 nt), Group I intron (T. thermophila, 425 nt), and HIV-1 genome (9173 nt).

The modification probabilities (βs) for these nine RNAs were converted from their SHAPE-MaP reactivities as follows: 1) negative reactivities were set to 0; 2) all other reactivities were scaled so that the highest scaled reactivity is equal to the largest ground truth β of Arabidopsis 18S rRNA. The RT drop-off probabilities (γs) were randomly sampled from the ground truth γs of Arabidopsis 18S rRNA. In order to explore the effect of expression level on estimation accuracy, we generated 4 sets of simulated data by varying the ground truth expression levels of the nine RNAs between 100, 1000, 10,000, and 100,000 *Transcripts Per Million* (TPM). Each set of simulated data consists of 30 million 37 nt single-end reads for both the modification-treated and mock-treated experiments.

Comparison with alternative methods on simulated and digital spike-in data— Our main simulation results are box plots comparing PROBer with alternative methods. In these box plots, we only focused on 1,802 transcripts that we may obtain reasonable RNA structure estimates. These transcripts were selected according to the following criteria: 1) its ground truth expression level ≥ 50 TPM; 2) its length ≥ 100 nt, and 3) its mappability score > 0. The mappability score is defined as the ratio between the number of 21 mers that can be mapped back uniquely and the total number of 21 mers in the same transcript. We further partitioned the 1,802 transcripts into 4 expression ranges in TPM: 887 transcripts in [50, 100], 849 transcripts in (10^2, 10^3], 60 transcripts in (10^3, 10^4], and 6 transcripts in (10^4, 10^6]. For each transcript and method, we calculated Pearson's correlation coefficient and Spearman's rank correlation coefficient between the ground truth modification probabilities and the estimates. In the calculation, we only used sites containing adenosines or cytosines because DMS mainly modifies adenosines and cytosines. In addition, we excluded the last 36 nt (read length is 37 nt) of each transcript from the analysis because there were little reads aligned to the 3′ end. We observed that icSHAPE had a median correlation of zero (Figures 2A, S2A, S3E, and S3F) in expression range [50, 100]. This is because icSHAPE did not provide structural estimates for these lowly expressed transcripts – icSHAPE outputted NULLs for most transcript positions in expression range [50, 100].

In addition to the results shown in Figure 2A, we also investigated the effects of interpolating hidden fragments that failed to pass size selection. We named PROBer with this interpolation enabled as the full model (see Method S1, section 2.2). With respect to Pearson's correlation (Figures S2B and S3A), the full model significantly increased the variance for structural estimates in low and medium expression ranges, which contain over 96% of investigated transcripts. For Spearman's rank correlation (Figures S2C and S3B), PROBer performed better than the full model with respect to the median correlations. These results validate our decision of taking off the size selection correction step from PROBer. To demonstrate the improvement in joint estimation of structural parameter and transcript abundance, we also compared PROBer with the RSEM + PROBer* pipeline. RSEM (Li et al., 2010; Li and Dewey, 2011) is a popular RNA-Seq transcript quantification tool that is not aware of RNA structure information. PROBer* is a modified version of PROBer that only works on a single transcript and thus is not aware of multi-mapping reads. Figures S2D, S2E, S3C, and S3D confirm our hypothesis – PROBer performs better at all expression ranges than the RSEM+PROBer* pipeline.

For digital spike-in experiments, we compared PROBer with alternative methods using boxplots of Pearson's correlations and Spearman's rank correlations on the nine spike-in RNAs (Figure S5). Since SHAPE reagent modifies all four RNA nucleotides, we included all but last 36 nt of each spike-in RNA in the calculation of correlations between the ground truth modification probabilities and the estimates. Similarly, icSHAPE had a median correlation of 0 when spike-in RNAs were lowly expressed (100 TPM). In addition, we observed that StructureFold had a median correlation of zero when spike-in RNAs were expressed at 100,000 TPM. This is because StructureFold failed to complete on the data where we set spike-in expression at 100,000 TPM.

Comparison with alternative methods using precision-recall and ROC curves —We compared PROBer's MAP estimates of chemical modification probabilities with alternative methods' scores using previously reported ribosomal RNA secondary structures (Cannone et al., 2002). Secondary structures for Arabidopsis 18S and 25S rRNAs, yeast 18S and 25S rRNAs, and mouse 18S and 12S mitochondrial rRNAs were obtained as BPSeq files. Sites on these rRNAs that participate in a base-pairing interaction were assigned an idealized modification rate of 0, and unpaired sites were assigned an idealized modification rate of 1. ROC curves comparing PROBer estimated MAP chemical modification rate and alternative method scores with this binary ground truth vector were produced and the areas under the ROC curves were calculated using PRROC v1.1 (Keilwagen et al., 2014). We excluded the last "read length – 1" nucleotides for each RNA from our analysis since there were little reads aligned to the 3′ end of RNAs. The number of nucleotides excluded in Ding et al. Arabidopsis data, Hector et al. yeast data, Talkish et al. yeast data, and Spitale et al. mouse data are 36, 48, 49, and 86 respectively. In addition, we only analyzed positions that are adenosines and cytosines for assays using DMS reagent. Because chemical reagents, such as DMS and SHAPE, might not be able to modify unbase-paired nucleotides that are blocked by other proteins and RNAs, it is hard to interpret these ROC curve results that were produced without solvent accessibility information.

Fortunately, for the two yeast data sets (Hector et al., 2014; Talkish et al., 2014), we have known crystal structure of the yeast ribosome complex (Ben-Shem et al., 2011). Thus, we in addition calculated precision-recall (PR) and ROC curves using the crystallographically informed solvent-accessible secondary structures as ground truth. To generate these PR and ROC curves, we further constrained our analysis to positions that are either base-paired or unbase-paired with solvent accessible surface areas of greater than 2Å^2. Solvent accessible surface area for adenosine and cytosine was calculated using PyMOL by following Rouskin et al. (Rouskin et al., 2014): 1) the crystal structure of yeast ribosome A (3U5B, 3U5C, 3U5D, and 3U5E) were downloaded from Protein Data Bank (http://www.rcsb.org/pdb/explore/obsolete.do?obsoleteId=3U5B) and formed into a complex in PyMOL; 2) DMS was modeled as a sphere with radius 3Å and solvent accessible surface area of N1-adenosine and N3-cytosine was calculated using the get_area function. We observed that the ground truth for yeast rRNA is highly skewed: 18S rRNA has 52 positives (A or C, unbase-paired, solvent accessible area > 2Å^2) and 435 negatives (A or C, base-paired); 25S rRNA has 93 positives and 865 negatives. It is known that when the ground truth is highly skewed, ROC curves tend to be overly optimistic and the PR curves are recommended to be used instead (Davis and Goadrich, 2006). Therefore we only showed PR curves (Figure 2B) in the main text and put ROC curves (Figure S6) in the supplement.

Comparison with alternative methods on predicting yeast rRNA structure—RNAstructure (Deigan et al. 2009) can use SHAPE constraints to help it better predicting RNA secondary structure. Using RNAstructure (v5.8.1), we compared PROBer with alternative methods on how well their estimates could be used as constraints for improving yeast 18S and 25S rRNA structure predictions. First, for each method, we selected only the modification estimates from positions that were used in the previous precision-recall curve analysis. Secondly, we converted these estimates into SHAPE constraints for RNAstructure using the normalization procedure recommended in (Sloma and Mathews, 2015). Lastly, we ran RNAstructure with default parameters and the converted SHAPE constraints, and evaluated the resulting minimum free energy structures using sensitivity and positive predictive value (PPV). Sensitivity is defined as the fraction of pairs in the ground truth that are correctly predicted and PPV is defined as the fraction of predicted pairs that also exist in the ground truth. We in addition compared PROBer with a baseline method that runs RNAstructure with no constraints.

Experiments on Poulsen et al. SHAPES data—Poulsen et al. data were aligned to E. coli 16S rRNA by Bowtie 2 with parameters used in (Poulsen et al., 2015). Since their data have strong priming biases, Poulsen et al. implemented a method to correct these biases in RNAprobR (Kielpinski et al., 2015). Using ROC and PR curves, we compared PROBer with raw read counts and RNAprobR-corrected counts on the E. coli SHAPES data (Figures S10A and S10B) and SHAPE-Seq data (Figures S10C and S10D). In the ROC and PR curves, we used the crystallographically informed solvent-accessible secondary structure of 16S rRNA as the ground truth. The crystallographically informed solvent-accessible secondary structure was obtained as follows: First, the 16S secondary structure was downloaded from the Comparative RNA Web database (Cannone et al., 2002) and the yeast ribosome crystal structure (3OFA, 3OFC) was downloaded from Protein Data Bank (http://www.rcsb.org/pdb/explore/obsolete.do?obsoleteId=3OFA). Secondly, the solvent accessible surface area at the 2′-hydroxyl group of RNA backbone of each nucleotide was calculated using PyMOL. Lastly, nucleotides that are unbase-paired according to the secondary structure and have solvent accessible areas no greater than 3Å^2 were excluded from the analysis. Following Poulsen et al. (Poulsen et al., 2015), we additionally restricted our analysis to the first 1,350 nucleotides.

Experiments on Carlile et al. Pseudo-seq data—We produced both PR and ROC curves using PRROC for Carlile et al.'s Pseudo-seq data. When we calculated the curves, we only considered 1,905 thymines in the yeast rRNAs and snoRNA. Since for yeast, we only have 49 known Ψ sites, which are part of the 1,905 thymines, the ground truth used here is also highly skewed. For this reason, we only presented PR curves in the main text (Figure 3A) and put the ROC curves (Figure S8) in the supplement. In addition, we observed a strange read count pattern at the 5′ end of 25S rRNA. Normally, the 5′ end base of a transcript should have a very high read count because of RT run-off. However, for 25S, the high read count appeared at the 3rd base. We hypothesized that this might be due to a small amount of degradation in the input RNA and therefore excluded the first 2 bases of 25S from our analysis.

Comparison with alternative methods on eCLIP and iCLIP data—We compared PROBer with a baseline method on the accuracy of allocating multi-mapping reads as follows: First, we aligned iCLIP and eCLIP reads to the human genome by Bowtie and Bowtie 2 using the default alignment criteria. We asked aligners to report all alignments and filtered out reads that aligned to more than 100 locations. Secondly, we allocated multi-mapping reads using both PROBer and the baseline method that allocates each multi-mapping read uniformly to all of its alignments. After this step, each alignment should be assigned a fractional weight. Lastly, we calculated the weighted motif hit rate for multi-mapping reads allocated by PROBer and the baseline method. Note that each alignment implies a putative crosslink site. Given a radius, we could test if there is a canonical motif around the crosslink site within the radius. We assigned 1 if the answer was yes and assigned 0 otherwise. Then the weighted motif hit rate was calculated as the alignment-weighted average of assigned values. We calculated the weighted motif hit rate for four eCLIP data sets (Table S5) and two iCLIP data sets (Table S7). For each data set, we varied the radius between 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, and 100 nt.

We compared PROBer with the current common practice, which only uses unique-mapping reads, on the number of peaks called from data. For eCLIP data, we followed the computational protocol documented in (Van Nostrand et al., 2016), which included calling peaks from aligned reads using CLIPper (Lovci et al., 2013) and normalizing the CLIPper peaks by input data. Since CLIPper could not process fractional read count, we sampled one “true” alignment for each multi-mapping read based on its alignment weights for both CLIP-treated and mock-treated (input) data. In Figure 3B, we showed the number of input-normalized peaks after controlling false discovery rate at 0.05 and the percentage of reported peaks that overlapped with a canonical motif for both PROBer and the common practice (unique method) on four distinct proteins. Following Van Nostrand et al., we considered that a peak overlapped with a canonical motif if and only if the motif was within the 100 nt radius of the peak center. We followed a similar procedure for comparing PROBer with the common practice on iCLIP data, except that we did not call input-normalized peaks since there was no input control in iCLIP data.

DATA AND SOFTWARE AVAILABILITY

Our main contribution, PROBer, is freely available with open source at http://pachterlab.github.io/PROBer.

PROBer contains five commands: prepare, estimate, simulate, iCLIP and version.

The first step in running PROBer is to build reference transcriptome indices using PROBer *prepare*. This command accepts either a genome or a set of transcript sequences as input. If a gnome is provided, users need to in addition provide a GTF/GFF3 file containing gene annotation information. PROBer will automatically extract transcript sequences from the genome using the specified annotation file. For iCLIP and eCLIP data, –genome option should be specified so that PROBer knows that genome indices are required instead.

PROBer *prepare* can optionally build Bowtie (Ben Langmead et al., 2009) and Bowtie 2 (Ben Langmead and Salzberg, 2012) indices by enabling –bowtie and –bowtie2 options. This command only needs to be run once per organism.

Once we have transcriptome indices built, we can obtain modification reactivity and transcript abundance estimates by running PROBer *estimate* on toeprinting data. PROBer *estimate* accepts either raw reads in FASTA/FASTQ format or alignments in SAM/BAM/CRAM format as input. It can handle single-end reads, paired-end reads and indel alignments. By default, Bowtie is used to align raw reads against the reference transcriptome. Bowtie 2 could be used instead by specifying the –bowtie2 option. PROBer *estimate* outputs ML estimates of transcript abundances and MAP estimates of modification and RT noise probabilities. If –output-bam is enabled, PROBer *estimate* in addition outputs BAM files consisting of posterior-probability-annotated read alignments. PROBer *estimate* can run with only modification-treated data if mock-treated control is not available. In that case, the estimated modification probabilities might not be as accurate. PROBer *estimate* provides the following options to describe key factors in toeprinting protocols: –primer-length, –size-selection-min, –size-selection-max, and –read-length. –primer-length specifies random primer length. This option should be set to 6 if random hexamer priming is used and to 0 if the protocol is fragmentation-based. –size-selection-min and –size-selection-max set the minimum and maximum fragment lengths in cDNA libraries after size selection. –read-length is only used for single-end reads and specifies the untrimmed read length. It helps PROBer to determine which single-end reads are adapter trimmed and thus can be regarded as full fragments.

For iCLIP and eCLIP data, we run PROBer *iCLIP* instead. Similar to PROBer *estimate*, PROBer *iCLIP* accepts iCLIP and eCLIP data either as raw reads in FASTA/FASTQ format or as alignments in SAM/BAM/CRAM format. If inputs are raw reads, either Bowtie or Bowtie 2 could be used to align these reads. If inputs are eCLIP data, we should turn on the –eCLIP option. For each crosslink site implied by the data, PROBer *iCLIP* outputs its genomic coordinate, unique read count, and expected multi-mapping read count.

PROBer *simulate* simulates toeprinting reads based on parameters learned from real data using PROBer *estimate*. PROBer *simulate* currently cannot simulate iCLIP or eCLIP reads.

PROBer *version* prints out the current version information.

If users want to assess the variation in PROBer-estimated modification probabilities, they should turn on –run-gibbs option in PROBer *estimate* and then run PROBer-bootstrap and PROBer-generateVariationPlot to generate variation plots for their transcripts of interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Yiliang Ding, Yin Tang, Joel McManus, and Thomas Carlile for discussions and clarifications on the StructureFold, Mod-seeker, and Pseudo-seq methods. We thank Yeon Lee, Julian König, Eric Van Nostrand, Gabriel

Pratt, and Gene Yeo for discussions on the iCLIP and eCLIP protocols. This work is supported by National Institutes of Health (NIH) grants R01 HG006129 to L.P. and R00 HG006860 to S.A., and by the Center for RNA Systems Biology at UC Berkeley (NIH P50GM102706 grant) to B.L. A.T. was partially supported by NIH Molecular Biophysics Training grant (NIH GM08295).

References

Aviran S, Lucks JB, Pachter L. RNA structure characterization from chemical mapping experiments. Proceedings of the 49th Allerton Conference on Communication, Control, and Computing. 2011a: 1743–1750.

Aviran S, Pachter L. Rational experiment design for sequencing-based RNA structure mapping. RNA. 2014; 20:1864–1877. [PubMed: 25332375]

Aviran S, Trapnell C, Lucks JB, Mortimer SA, Luo S, Schroth GP, Doudna JA, Arkin AP, Pachter L. Modeling and automation of sequencing-based characterization of RNA structure. Proc Natl Acad Sci USA. 2011b; 108:11069–11074. [PubMed: 21642536]

Ben-Shem A, de Loubresse NG, Melnikov S, Jenner L, Yusupova G, Yusupov M. The structure of the eukaryotic ribosome at 3.0 Å resolution. Science. 2011; 334:1524–1529. [PubMed: 22096102]

Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34:525–527. [PubMed: 27043002]

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. 2002; 3:2. [PubMed: 11869452]

Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014; 515:143–146. [PubMed: 25192136]

Choudhary K, Deng F, Aviran S. Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. Quant Biol. 2017; 5:3–24. [PubMed: 28717530]

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleş S. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol. 2011; 7:e1002111. [PubMed: 21779159]

Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning. 2006:233–240.

Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA. 2009; 106:97–102. [PubMed: 19109441]

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B Stat Methodol. 1977; 39:1–38.

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014; 505:696–700. [PubMed: 24270811]

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012; 485:201–206. [PubMed: 22575960]

Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, Dai Q, Di Segni A, Salmon-Divon M, Clark WC, et al. The dynamic N^1-methyladenosine methylome in eukaryotic messenger RNA. Nature. 2016; 530:441–446. [PubMed: 26863196]

Efron, B., Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall/CRC; 1993.

Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Letters. 2008; 582:1977–1986. [PubMed: 18342629]

Hartz D, McPheeters DS, Traut R, Gold L. Extension inhibition analysis of translation initiation complexes. Meth Enzymol. 1988; 164:419–425. [PubMed: 2468068]

Hector RD, Burlacu E, Aitken S, Le Bihan T, Tuijtel M, Zaplatina A, Cook AG, Granneman S. Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. Nucleic Acids Res. 2014; 42:12138–12154. [PubMed: 25200078]

Incarnato D, Anselmi F, Morandi E, Neri F, Maldotti M, Rapelli S, Parlato C, Basile G, Oliviero S. High-throughput single-base resolution mapping of RNA 2′-O-methylated residues. Nucleic Acids Res. 2017; 45:1433–1441. [PubMed: 28180324]

Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted Data. PLoS ONE. 2014; 9:e92209. [PubMed: 24651729]

Kielpinski LJ, Sidiropoulos N, Vinther J. Reproducible analysis of sequencing-based RNA structure probing data with user-friendly tools. Meth Enzymol. 2015; 558:153–180. [PubMed: 26068741]

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010; 17:909–915. [PubMed: 20601959]

Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics. 2012; 28:2520–2522. [PubMed: 22908215]

Kwok CK, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nat Methods. 2016; 13:841–844. [PubMed: 27571552]

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26:493–500. [PubMed: 20022975]

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013; 20:1434–1442. [PubMed: 24213538]

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011; 17:10–12.

Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014; 15:469–479. [PubMed: 24821474]

Poulsen LD, Kielpinski LJ, Salama SR, Krogh A, Vinther J. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. RNA. 2015; 21:1042–1052. [PubMed: 25805860]

Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013; 10:71–73. [PubMed: 23160280]

Roundtree IA, He C. RNA epigenetics–chemical messages for posttranscriptional gene regulation. Curr Opin Chem Biol. 2016; 30:46–51. [PubMed: 26625014]

Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2014; 505:701–705. [PubMed: 24336214]

Sakurai M, Ueda H, Yano T, Okada S, Terajima H, Mitsuyama T, Yoyoda A, Fujiyama A, Kawabata H, Suzuki T, et al. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. Genome Res. 2014; 24:522–534. [PubMed: 24407955]

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

Selega A, Sirocchi C, Iosub I, Granneman S, Sanguinetti G. Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments. Nat Methods. 2017; 14:83–89. [PubMed: 27819660]

Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods. 2014; 11:959–965. [PubMed: 25028896]

Silverman BW, Jones MC, Wilson JD. A smoothed EM approach to indirect estimation problems, with particular, reference to stereology and emission tomography. J R Stat Soc Series B Stat Methodol. 1990; 52:271–324.

Sloma MF, Mathews DH. Improving RNA secondary structure prediction with structure mapping data. Meth Enzymol. 2015; 553:91–114. [PubMed: 25726462]

Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, Kuchelmeister HY, Batista PJ, Torre EA, Kool ET, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015; 519:486–490. [PubMed: 25799993]

Talkish J, May G, Lin Y, Woolford JL, McManus CJ. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA. 2014; 20:713–720. [PubMed: 24664469]

Tang Y, Bouvier E, Kwok CK, Ding Y, Nekrutenko A, Bevilacqua PC, Assmann SM. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. Bioinformatics. 2015; 31:2668–2675. [PubMed: 25886980]

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016; 13:508–514. [PubMed: 27018577]

Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. Nat Rev Genet. 2011; 12:641–655. [PubMed: 21850044]

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu Elements. Cell. 2013; 152:453–466. [PubMed: 23374342]

Highlights

Sequencing-based transcriptase drop-off assays probe post-transcriptional dynamics

PROBer learns isoform-specific modification profiles accurately from these assays

Simulated and real data results suggest PROBer significantly outperforms alternatives

PROBer is valuable to RNA research as a general analysis tool for drop-off assays

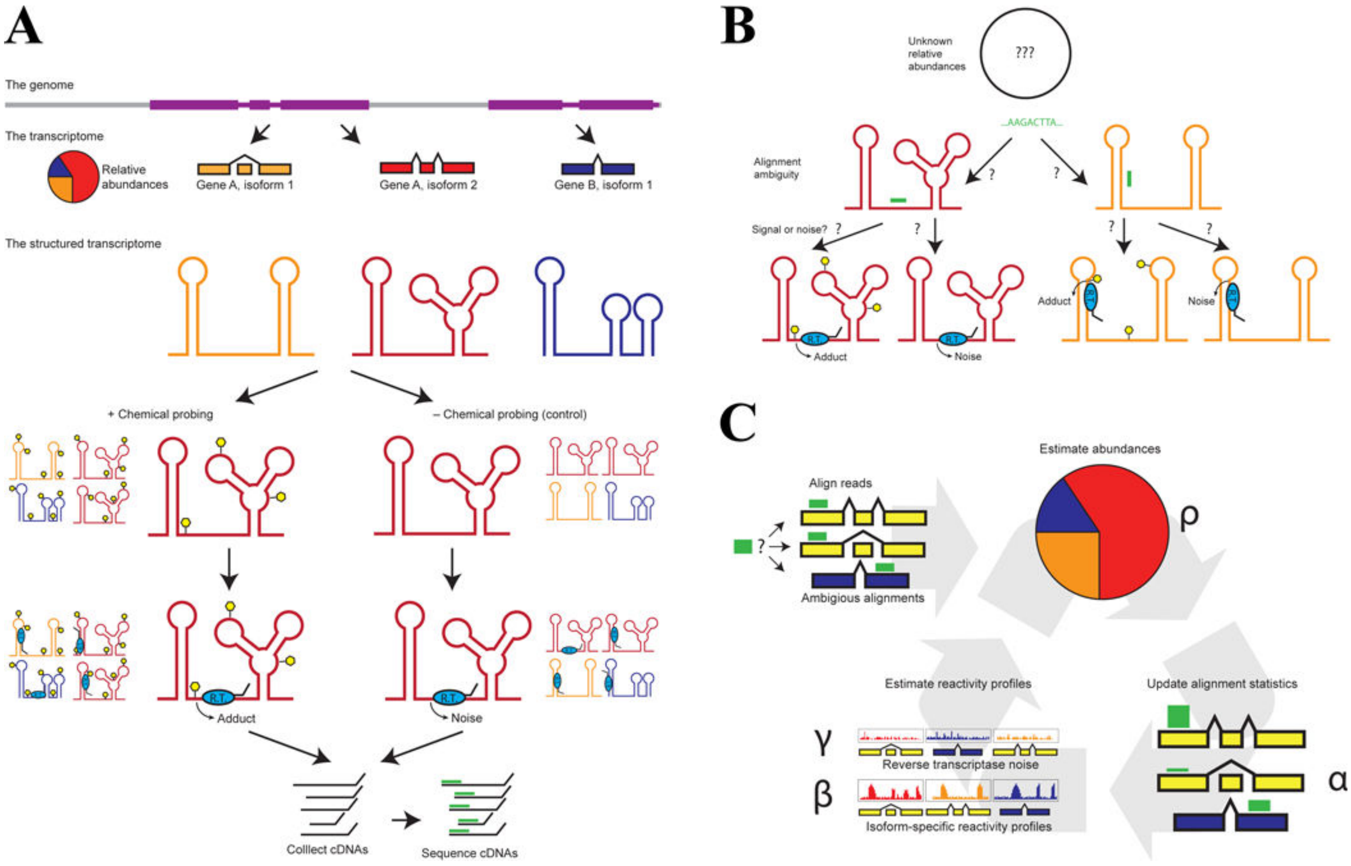
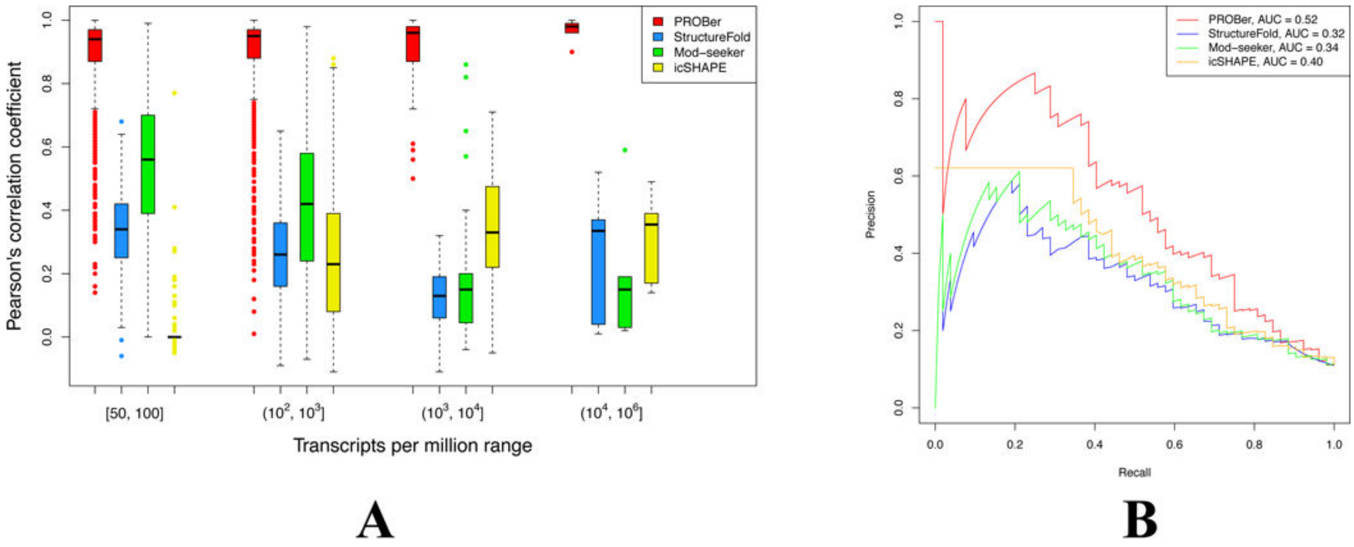


Figure 1. Cartoon depictions of sequencing-based toeprinting experiments, the associated Bioinformatics challenges, and our solution

(A) Cartoon depiction of an idealized toeprinting experiment. The genome is transcribed and RNAs are spliced and folded to form the structured transcriptome. This pool of RNAs is split into two, and either treated with a chemical probe, or mock-treated without the chemical probe. These chemical adducts are detected by reverse transcriptase (RT) drop-off, but the signal is convoluted by reverse transcriptase noise. Reverse transcription products are collected and sequenced. (B) Potential bioinformatics challenges. The structured transcriptome that gave rise to a given toeprinting dataset consists of known transcripts of unknown relative abundance. Reads from this dataset might align ambiguously to one or more transcripts, and might have been generated by either RT drop-off at a chemical modification, or by RT noise. (C) Conceptual workflow of PROBer. Sequencing data (both treatment and control datasets) from a toeprinting experiment are used as the input. In the E-step, reads are assigned to transcripts depending on an initial alignment, and the relative abundances & toeprinting parameters of the transcripts estimated in the M-step. In the M-step, transcript abundances and toeprinting parameters are learned, using the read assignments calculated in the E-step.

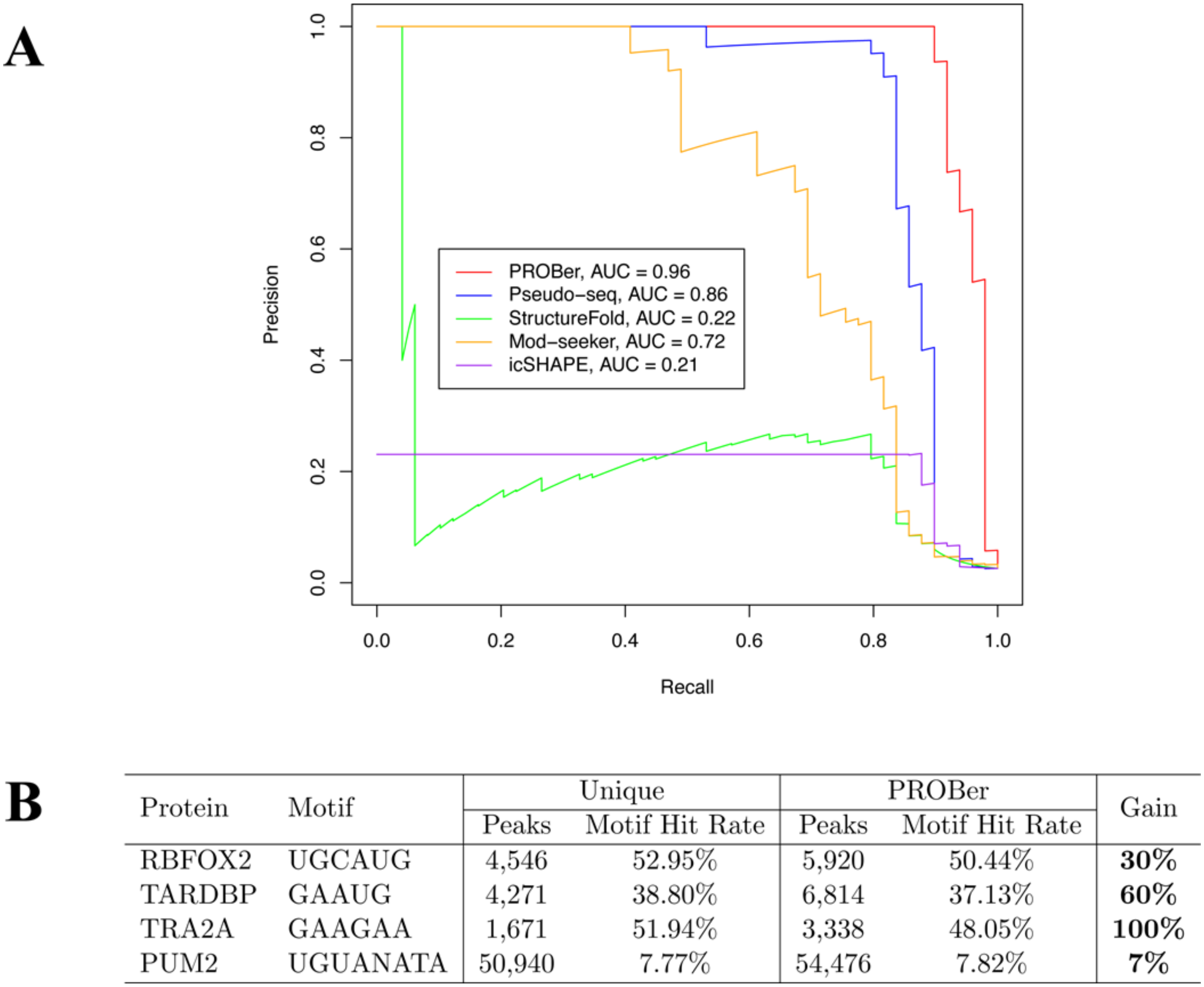


Method	Talkish *et al.*				Hector *et al.*			
	18S		25S		18S		25S	
	Sensitivity	PPV	Sensitivity	PPV	Sensitivity	PPV	Sensitivity	PPV
PROBer	**46.49%**	**41.06%**	**54.63%**	**51.59%**	**44.89%**	**39.68%**	48.66%	45.74%
StructureFold	NA	NA	NA	NA	43.29%	36.67%	48.46%	44.52%
Mod-seeker	35.07%	31.03%	36.52%	35.23%	34.87%	30.30%	46.37%	44.42%
icSHAPE	45.69%	40.35%	49.15%	47.14%	28.26%	25.13%	36.82%	34.61%
Baseline	35.67%	30.19%	51.54%	47.83%	35.67%	30.19%	**51.54%**	**47.83%**

C

Figure 2. Performance of PROBer on RNA structure probing assays

(A) A simulated RNA structure-probing dataset was generated in a manner consistent with Ding et al. protocol (Ding et al., 2014), and used as the input for a number of structure-probing quantification methods, which include PROBer, StructureFold (Tang et al., 2015), Mod-seeker (Talkish et al., 2014), and icSHAPE (Spitale et al., 2015). Accuracy was evaluated by comparing the results from these methods with the ground truth modification reactivity profiles (β values) using Pearson's correlation coefficients. PROBer consistently outperforms alternative approaches across a wide range of expression levels. See also Figures S2 and S3. (B) PROBer was compared with alternative methods for in vitro probing yeast 18S rRNA on a paired-end data set (Hector et al., 2014). Methods were evaluated by Precision-Recall (PR) curves and area under curve (AUC) values using crystallographically informed solvent-accessible secondary structures as ground truth. PROBer outperforms alternative methods significantly. See also Figure S6 and Table S3. (C) PROBer was compared with alternative methods on two yeast structure-probing data sets (Hector et al., 2014; Talkish et al., 2014) for predicting 18S and 25S rRNA secondary structures. For each method, the estimated modification reactivity profiles were converted into SHAPE constraints (see STAR★METHODS) and then the SHAPE constraints were fed to RNAstructure (Deigan et al., 2009) to produce secondary structure predictions. We additionally compared with a baseline method, which ran RNAstructure with no SHAPE constraints. The resulting minimum free energy structures were evaluated on two commonly used metrics: sensitivity and positive predictive value (PPV). We highlight the best performer of each column in bold. PROBer outperforms alternative approaches in all 4 cases and outperforms the baseline in 3 out of 4 cases.



Protein	Motif	Unique		PROBer		Gain
		Peaks	Motif Hit Rate	Peaks	Motif Hit Rate	
RBFOX2	UGCAUG	4,546	52.95%	5,920	50.44%	**30%**
TARDBP	GAAUG	4,271	38.80%	6,814	37.13%	**60%**
TRA2A	GAAGAA	1,671	51.94%	3,338	48.05%	**100%**
PUM2	UGUANATA	50,940	7.77%	54,476	7.82%	**7%**

Figure 3. Performance of PROBer on detecting pseudouridine modifications and identifying protein-RNA binding sites

(A) PROBer was compared with alternative methods on data for predicting known pseudouridine (Ψ) sites in yeast rRNAs and snoRNA (Carlile et al., 2014). Methods were evaluated by precision-recall (PR) curves and area under curve (AUC) values. PROBer outperforms alternative approaches significantly. See also Figures S7 and S8. (B) PROBer was compared with the common practice (unique method) that uses only uniquely-mapping reads on eCLIP (Van Nostrand et al., 2016) data sets for 4 distinct RNA-binding proteins. The first two columns in the table give the protein name and canonical binding motif. The binding motifs have been validated both in vitro and in vivo (Van Nostrand et al., 2016). The next four columns give the number of input-normalized peaks called at false discovery rate of 0.05 (see STAR★METHODS) and the percentage of input-normalized peaks overlapping with canonical motifs for the unique method and PROBer respectively. Peaks were called using CLIPper (Lovci et al., 2013). The last column gives the percentage of more peaks

PROBer detected comparing with the unique method. PROBer enables us to extract significantly more information from the eCLIP data sets. See also Table S5, S6 and S7.