Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Sequence analysis

RiboProP: A Probabilistic Ribosome Positioning Algorithm for Ribosome Profiling

Dengke Zhao [1], William Baez [2], Kurt Fredrick [3,4], and Ralf Bundschuh [1,2,4,5,6,*]

[1]Interdisciplinary Biophysics Graduate Program, The Ohio State University, Columbus, OH, 43210, USA,
[2]Department of Physics, The Ohio State University, Columbus, OH, 43210, USA,
[3]Department of Microbiology, The Ohio State University, Columbus, OH, 43210, USA,
[4]Center for RNA Biology, The Ohio State University, Columbus, OH, 43210, USA,
[5]Department of Chemistry & Biochemistry, The Ohio State University, Columbus, OH, 43210, USA and
[6]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, OH, 43210, USA.

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on 02 Jun 2018; revised on 05 Sep 2018; accepted on XXXXX

Abstract

Motivation: Ribosome profiling has been widely used to study translation in a genome-wide fashion. It requires deep sequencing of ribosome protected mRNA fragments followed by mapping of fragments to the reference genome. For applications such as identification of ribosome pausing sites, it is not enough to map a fragment to a given gene, but the exact position of the ribosome represented by the fragment must be identified for each mRNA fragment. The assignment of the correct ribosome position is complicated by the broad length distribution of the ribosome protected fragments caused by the known sequence bias of micrococcal nuclease (MNase), the most widely used nuclease for digesting mRNAs in bacteria. Available mapping algorithms suffer from either MNase bias or low accuracy in characterizing the ribosome pausing kinetics.

Results: In this paper, we introduce a new computational method for mapping the ribosome protected fragments to ribosome locations. We first develop a mathematical model of the interplay between MNase digestion and ribosome protection of the mRNAs. We then use the model to reconstruct the ribosome occupancy profile on a per gene level. We demonstrate that our method has the capability of mitigating the sequence bias introduced by MNase and accurately locating ribosome pausing sites at codon resolution. We believe that our method can be broadly applied to ribosome profiling studies on bacteria where codon resolution is necessary.

Availability: Source code implementing our approach can be downloaded under GPL3 license at http://bioserv.mps.ohio-state.edu/RiboProP.

Contact: bundschuh@mps.ohio-state.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ribosome profiling, or Ribo-seq, has become increasingly popular since it was introduced in 2009 (Ingolia *et al.*, 2009). It allows for the simultaneous quantification and localization of all translating ribosomes in a cell. There are three categories of applications of ribosome profiling (Ingolia, 2014, 2016): (1) quantification of protein synthesis at the translational level, (2) identification of open reading frames (ORFs), and (3) quantification of translation kinetics and identification of ribosome pausing sites. In the first two types of applications, the accurate positions of ribosomes are not critical, since the focus of such studies is the relative abundance of the ribosomes on individual genes or the regions of the genome that are actually translated. However, for applications such as identification of peptide-mediated ribosome pausing sites (Gong and Yanofsky, 2002; Nakatogawa and Ito, 2002; Bhushan *et al.*, 2011; Woolstenhulme *et al.*, 2015), ribosome positions at codon resolution are crucial for differentiating a slowly translating codon from its nearby fast translating codons.

A typical ribosome profiling experiment requires deep sequencing of ribosome protected mRNA fragments followed by mapping the fragments to the reference genome. This mapping provides genome wide ribosome density profiles for further analysis. Since the ribosome protected fragments have a length of around 20 to 35 nucleotides, applications that rely on precise locations of individual ribosomes require assigning the functional sites of each ribosome to a particular nucleotide within

© The Author 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

the fragment it protects. While there are different conventions on which position on the ribosome (such as, e.g., the central nucleotide of the P-site codon or the final nucleotide of the A-site codon) to use as a reference, these conventions are all equivalent to each other up to a known relative offset. However, no matter the convention chosen, an important question is how to identify the location of the chosen position of the ribosome within every ribosome protected fragment sequenced in the experiment.

Sample preparation has important consequences for the assignment of ribosomal positions to ribosome protected fragments. Ribosome profiling has been performed on both prokaryotes and eukaryotes. The original ribosome profiling experiment was conducted on budding yeast by Ingolia *et al.*, 2009, in which RNAase I was used for digesting the mRNA that is not under the protection of ribosomes. RNAase I is an unbiased nuclease and is known to be able to digest the mRNA almost to the "edge" of the ribosomes, resulting in a narrow length distribution of the ribosome protected fragments. Choosing a consistent location on these fragments has not been much of a challenge. However, for ribosome profiling studies on bacteria such as *Escherichia coli*, the best available nuclease for digesting mRNA is micrococcal nuclease (MNase), since the activity of RNAase I is inhibited by the 30S subunit (Datta and Burma, 1972) rendering it unusable in bacterial systems (Woolstenhulme *et al.*, 2015). Unlike RNAase I, MNase has a strong sequence bias (Hörz and Altenburger, 1981). The ribosome protected mRNA fragments obtained from MNase digestion have a broad length distribution, making assigning the correct ribosomal location to each read challenging. Ideally, to assign the correct ribosomal location to a specific read one has to consider the sequence bias of the MNase; however, this approach has not been available to date.

In the area of ribosome profiling in bacteria, there are three broad categories of methods available for mapping the ribosome protected fragments to the second nucleotide of the P site (which is for concreteness the convention we will adopt for this study): (1) the most widely used fixed offset method, where the read is mapped to a position at a fixed distance from one extremity of the read (usually 14 or 15 nucleotides from the 3' end depending on the experimental conditions and the organism) (Woolstenhulme *et al.*, 2015; Balakrishnan *et al.*, 2014); (2) the variable offset method, which is similar to the fixed offset method except that one separates the reads into length groups and chooses a specific offset for each length group (Dunn and Weissman, 2016); and (3) the center weighted method, where the score of a read is spread over one or more adjacent positions in the center area of the read (Li *et al.*, 2012; Oh *et al.*, 2011). None of these methods explicitly takes into account the sequence of the read and thus the MNase sequence bias.

In this paper, we propose a new method for mapping the ribosome protected fragments. We first develop a model which takes into account the MNase sequence bias and the effect of ribosome protection against MNase digestion. We then optimize the free parameters in the model and apply this model to reconstruct the ribosome density profile on a per gene level. To demonstrate the performance of our method, we test it on the ribosome profiling data set by Balakrishnan *et al.*, 2014 and compare the result to all three other methods listed above. Through comparison, we show that our method is able to significantly reduce the effect of MNase bias in the process of mapping and accurately predict ribosome pausing sites at codon resolution. We thus believe that this method has a broad impact on ribosome profiling experiments in bacteria when codon resolution is needed.

2 Methods

2.1 Sequence bias of micrococcal nuclease and ribosome protection model

In this paper, we focus our discussion on bacterial ribosome profiling studies, which use MNase for mRNA digestion. We consider two factors when developing a mathematical model for the experimental system: (1) sequence bias of micrococcal nuclease (MNase) and (2) ribosomal protection against MNase digestion of mRNA. First, since MNase is known to have strong sequence bias (Hörz and Altenburger, 1981), we assume four separate cleavage rates s_A, s_U, s_G and s_C for mRNA cleavage 5' of a given nucleotide A, U, G, and C, respectively. Second, we assume that the ribosome completely protects nucleotides far inside the ribosome from MNase cleavage and does not protect nucleotides far outside of the ribosome from cleavage at all, while nucleotides in the vicinity of the "edge" of the ribosome are partially protected from cleavage. To be specific, we choose the protection induced relative cleavage efficiency $r(x) \in [0,1]$ at nucleotide position x to be given by a Gaussian error function

$$r(x) = \frac{\mathrm{erf}[a(x-b)]+1}{2} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{a(x-b)} e^{-t^2}\,\mathrm{d}t \qquad (1)$$

which smoothly transitions from 0 at $x \ll b$ to 1 at $x \gg b$ over a range of width $1/a$. Then, the MNase cleavage rates and ribosome protection function together yield the effective MNase cleavage rate at the i^{th} nucleotide $k_i = r_i \cdot s_{n_i} = \frac{1}{2}\{\mathrm{erf}[a(i-b)]+1\} \cdot s_{n_i}$, where $n_i \in \{A,U,G,C\}$ is the nucleotide of the mRNA at position i relative to a fixed position on the ribosome.

We then denote the time to complete the MNase digestion during the ribosome profiling experiment as t. Since k_i is the rate of cleavage 5' of the i^{th} nucleotide, the probability of the backbone between the $i-1^{\mathrm{st}}$ and the i^{th} nucleotide not being cut after time t is $e^{-k_i t}$. We assume that protection is complete for any nucleotide in the ribosomal A and P site and thus characterize the 3' end of a ribosome protected fragment in terms of the number m of nucleotides starting from the first nucleotide on the 3' side of the ribosomal A site. In order for this 3' end of a ribosome protected fragment to be m nucleotides 3' of the A site, cleavage cannot have occurred at any of the m intervening positions and must have occurred at the $m+1^{\mathrm{st}}$ position (the cleavage status at any nucleotide beyond the $m+1^{\mathrm{st}}$ is irrelevant for the 3' end of the ribosome protected fragment due to the cleavage at the $m+1^{\mathrm{st}}$ position). Thus, the probability of the ribosome protected fragment having m nucleotides beyond the ribosomal A site is

$$\begin{aligned} P(m|n_1 \ldots n_{m+1}) &= e^{-k_1 t} \cdot e^{-k_2 t} \cdot e^{-k_3 t} \cdot \ldots \cdot e^{-k_m t}(1-e^{-k_{m+1} t}) \\ &= \left[\prod_{i=1}^{m} e^{-k_i t}\right] \cdot (1-e^{-k_{m+1} t}), \end{aligned} \qquad (2)$$

which depends on the sequence $n_1 \ldots n_m$ of the ribosome protected fragment 3' of the ribosomal A site and the identity of the first nucleotide n_{m+1} of the mRNA 3' of the ribosome protected fragment since the k_i are given in terms of the nucleotide dependent MNase digestion rates s_{n_i}.

2.2 Optimization of the model parameters

We optimize the parameters (a, b, s_A, s_U, s_G and s_C) in equation (2) by fitting the model to a subset of ribosome profiling data with known ribosome positions. Because the time t for MNase digestion in a ribosome profiling experiment only enters in terms of its product with the s_N, we combine the time variable t and the sequence specific MNase cutting rates into single parameters $s_A \cdot t$, $s_U \cdot t$, $s_G \cdot t$ and $s_C \cdot t$ in the optimization. In order to identify reads for which we know the position of the corresponding

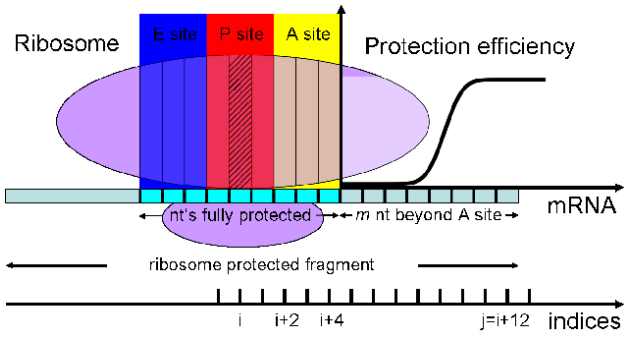

Fig. 1. Schematic of the ribosome protection of mRNA against MNase cleavage. In our model, the bare, sequence dependent MNase cleavage rate is modulated by the ribosome protection, whose efficiency is mathematically modeled by a gaussian error function (black line). The ribosome completely protects nucleotides far inside of the ribosome and does not protect nucleotides far outside of the ribosome, while near the âŁœedgeâŁž of the ribosome the nucleotides are partially protected. The diagonally striped position corresponds to the second codon position of the P site, which is used as the convention to define the ribosome position i. The light blue object indicates the ribosome protected fragment, which in this particular example ends by an MNase cut after base $j = i + 12$.

ribosome, we exploit the fact that in bacteria, ribosomes are known to arrest at the stop codon for a while until they are released. Thus, the vast majority of all reads in the vicinity of the 3' ends of genes will be derived from ribosomes located with the stop codon in the A site. In order to optimize our parameters, we select these reads from the 40 genes with the highest ribosome counts per length from the ribosome profiling dataset by Balakrishnan *et al.*, 2014. For each read j we obtain the number m_j of nucleotides it extends beyond the stop codon of its gene as well as the sequence r_j starting with the first base 3' of the stop codon and extending by one base beyond the end of the actual read with the help of the known genomic sequence. Since we assume the read is generated from a ribosome with the stop codon in its A site, the probability of a single such read is given by equation (2), and the probability of the entire pool of M reads is given by

$$\begin{aligned} P(\text{pool}) &= P(m_1|r_1) \cdot P(m_2|r_2) \cdot P(m_3|r_3) \dots P(m_M|r_M) \\ &= \prod_{j=1}^{M} P(m_j|r_j) \end{aligned} \tag{3}$$

We use maximum likelihood estimation to obtain the optimized parameters, i.e., we insert equation (2) for every $P(m_j|r_j)$ in equation (3) to obtain $P(\text{pool})$ as a function of parameters a, b, $s_A \cdot t$, $s_U \cdot t$, $s_G \cdot t$ and $s_C \cdot t$. We then use simulated annealing to vary these parameters and find the set of parameters that maximize $P(\text{pool})$.

2.3 Reconstruction of ribosome density profiles on a per gene level

We use our model to reconstruct the ribosome density profiles on a per gene level. For a given mRNA of ℓ nucleotides (not counting the stop codon), we denote the (unknown) true ribosome density at nucleotide position i (in our convention defined as the density of ribosomes with their P site centered on nucleotide i, see Figure 1) as x_i. For each position i, we then calculate the expected distribution $z_{j|i}$ of the 3' most nucleotide of a ribosome protected fragment being at position j in the gene, given the protecting ribosome is located with the P site centered on nucleotide i based on our model as

$$z_{j|i} = \begin{cases} 0 & j \le i+4 \\ P(j-i-4|n_{i+5} \dots n_{j+1}) & i+5 \le j < i+34 \\ 0 & j \ge i+34 \end{cases} . \tag{4}$$

Here, we for practical purposes do not consider the probability of the 3' most nucleotide of the fragment to be at positions after $i+34$ because the mRNA at these positions is far beyond the protection of the ribosome and thus should be completely cleaved if the mRNA is digested for a sufficient amount of time. Similarly, we set the probability of 3' ends of fragments to be in the A site or P site of the ribosome to be zero, since these areas of the fragment are fully protected by the ribosome. The resulting overall 3' end distribution y_j of the ribosome protected fragments for the entire gene is then given by

$$y_j = \sum_{i=-4}^{\min\{j,\ell-4\}} z_{j|i} \cdot x_i \qquad 1 \le j \le \ell + 30. \tag{5}$$

In the equation above, negative indices indicate positions beyond the 5' end of the gene, which implies that our calculation allows for some ribosomes to be located upstream of the start codon. Most likely, starting from a random initial guess of the x_i, the thus calculated $(y_1, y_2, y_3 \dots y_{\ell+30})$ are different from the observed distribution $(y'_1, y'_2, y'_3 \dots y'_{\ell+30})$ of 3' ends in the actual ribosome profiling data (normalized by their mean coverage). We then optimize $(x_{-4}, x_{-3}, x_{-2}, x_{-1}, x_1, x_2, x_3 \dots x_{\ell-4})$ to make the calculated $(y_1, y_2, y_3 \dots y_{\ell+30})$ as close as possible to the observed $(y'_1, y'_2, y'_3 \dots y'_{\ell+30})$ using a simulated annealing algorithm. The method we use to quantify the discrepancy between the expected and the actual distribution of the 3' ends of the fragments is the Bhattacharyya distance $d = -\log(\sum_{j=1}^{\ell+30} \sqrt{y_j \cdot y'_j})$. The resulting x_i's (normalized by their mean) are the reconstructed ribosome coverages after normalization.

Table 1. Length of the ribosome protected mRNA read and its corresponding P-site offset used to map the read to the reference genome. Note that in the ribosome footprints data we use, the reads $\le$33 nucleotides account for more than 99.9% of the total, although we map the reads up to 42 nucleotides using variable offset method, we do not expect the length dependent offset for reads $>$33 nucleotides to be accurately determined using the termination peak near stop codon (Balakrishnan et al., 2014) due to the limited amount of such reads.

Length (nt)	Offset (nt)	Length (nt)	Offset (nt)
20	15	32	16
21	14	33	20
22	12	34	9
23	13	35	15
24	14	36	25
25	15	37	5
26	14	38	8
27	14	39	13
28	15	40	14
29	16	41	9
30	15	42	28
31	16		

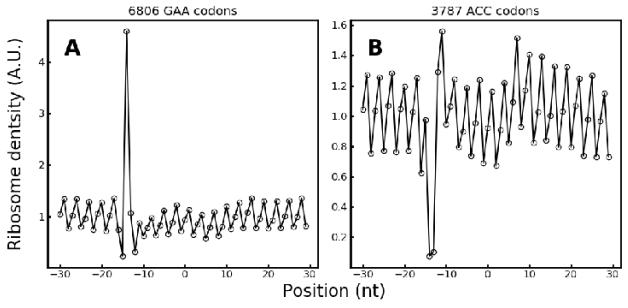

Fig. 2. Average ribosome footprint density around specific codons: (A) GAA codon (B) ACC codon. We average ribosome footprint density around 6806 GAA and 3787 ACC codons from the 500 most highly translated genes. The 0 on the horizontal axis is the position of the first nucleotide within the codon. Negative indices are toward the 5' end of the window and positive indices are toward the 3' end of the window. In this example, the method for mapping the ribosome footprints to the reference genome is the conventional fixed offset method. We count 14 nucleotides upstream of the 3' end of the ribosome protected fragment to obtain the corresponding position of the central nucleotide of the P site (Balakrishnan et al., 2014). We apply this approach to all ribosome footprints to obtain the genome-wide ribosome density profile. The upward peak at around -14 for the GAA codon and the downward peak at around -14 for the ACC codon are quantitative indications of the sequence bias of MNase used in ribosome profiling experiments. See Figure S1 for the same analysis for all other possible codons.

2.4 Comparison between different mapping algorithms

To compare our method of assigning P-site locations to other available methods, we selected three different mainstream methods and applied them to the ribosome footprint data set by Balakrishnan *et al.*, 2014.

Fixed offset: We use a similar method as described in (Balakrishnan *et al.*, 2014). Briefly, we select the ribosome densities in a ± 30 nucleotide window around the stop codon of the 1000 genes with the largest ribosome counts per length (with 0 representing the position of the third nucleotide of the stop codon), normalize the coverages of 3' ends of ribosome protected fragments in the window by the average of the window for every gene, and then average all 1000 windows to obtain the average 3' end coverages around the stop codon. The highest 3' end density on the 3' side of the stop codon is located at the 10^{th} nucleotide; thus, we calculate the most likely central nucleotide of the P site to be 14 nucleotides upstream of the 3' end of the fragments. We then use this fixed offset to map all reads.

Variable offset: Following Dunn and Weissman, 2016 we proceed similarly to the fixed offset method, except that we separate reads into different groups based on their length. For each length, we perform the same analysis as for the fixed offset method resulting in the P-site offsets for each read length shown in Table 1. We then map the reads using these length dependent P-site offsets.

Center weighted: We use the method described in (Oh *et al.*, 2011; Li *et al.*, 2012). We select the reads with a length of at least 21 nucleotides. For reads that are 21 nucleotides long, we assign the center of the read as the central nucleotide of the ribosomal P site. For reads that are longer than 21 nucleotides, we first remove 10 nucleotides on both ends of the read; for the remaining N nucleotides, we assign each position a weight of $1/N$.

3 Results

3.1 Ribosome density profiles around specific codons reveal significant MNase biases

In order to illustrate the severity of MNase bias, we perform a metagene analysis of ribosome density around particular codons. For each of the 61 non-stop codons, we identify all occurrences in the 500 most highly translated genes in the ribosome profiling data set by Balakrishnan *et al.*, 2014, obtain the ribosome density in a 60 nucleotide window centered on the first nucleotide of the selected codon, normalize this density in the window, and average over all occurrences of the codon. Figure 2 shows these averaged profiles for two representative codons (GAA and ACC) while supplementary figure S1 shows the profiles for all 61 non-stop codons, both using a fixed distance of 14 nucleotides from the 3' end of the read to define the position of the ribosomal P site when calculating ribosome density as it has been used in the original study (Balakrishnan *et al.*, 2014). As expected the averaged profiles show a pronounced periodicity of three nucleotides consistent with the in frame localization of ribosomes indicating the near nucleotide resolution of P-site assignments.

However, the three nucleotide periodicity of the ribosome density profiles is interrupted by a striking feature at position -14. This feature is a large apparent ribosome density increase for GAA codons and a large apparent ribosome density decrease for ACC codons (Figure 2) and can be found for many of the remaining non-stop codons as well (Figure S1). Since it is unlikely that ribosome density truly consistently increases or decreases significantly 14 nucleotides upstream of every GAA or ACC codon, respectively, this feature must represent an artifact of the experimental method. Since the assignment of ribosomal P sites to ribosome protected fragments in this data is performed using the fixed distance of 14 nucleotides from the 3' end of the ribosome protected fragment, the feature represents precisely those ribosome protected fragments, the 3' end of which coincides with the fixed codon. We thus hypothesize that this feature is a result of the MNase sequence bias and use it below as a diagnostic for the impact of MNase sequence bias on different analysis approaches.

3.2 Quantitative modeling of MNase bias and ribosome protection

To reduce the artifacts caused by MNase bias, we develop a ribosome positioning algorithm based on the biophysical model of the sequence bias of MNase and ribosome protection against MNase cleavage shown in Figure 1. MNase is known to cleave mRNA at different rates based on the nucleotide 5' of the cleavage site (Hörz and Altenburger, 1981; Dingwall *et al.*, 1981). We thus allow four independent cleavage rates, one for each nucleotide, in our model. These cleavage rates are further modulated by the ribosome protection efficiency. We model the ribosome as a partially flexible body that fully protects the nucleotides deep within the ribosome from cleavage, allows the full cleavage rate for nucleotides far outside the ribosome, and reduces the cleavage rates in a sequence independent manner at the "edge" of the ribosome. Since it is well established that ribosome protection of the 5' ends of bacterial ribosome profiling fragments is much more variable than protection of the 3' ends (Woolstenhulme *et al.*, 2015) we apply this approach only to the 3' side of the ribosome. Using this model, we can then calculate the probability distribution of the 3' ends of ribosome protected fragments, given the positions of the ribosomes at specific locations and the mRNA sequence. Fitting this model to the observed profile of 3' ends of ribosome protected fragments yields a distribution of ribosome positions for every gene. See the methods section for details on the model and the fitting procedures.

3.3 Our model captures the complex 3' end distributions of ribosome protected fragments near stop codons

Due to the known ribosome accumulation at stop codons in bacteria, we expect that the ribosome footprints derived from the 3' end of each gene have accurately identified P-site and A-site positions (i.e. the vast majority of these ribosome protected fragments should stem from ribosomes with the stop codon located in the A site). Our model quantifies the distribution of 3' ends of ribosome protected fragments given the known position of a ribosome using 6 fitting parameters a, b, $s_A \cdot t$, $s_U \cdot t$, $s_G \cdot t$ and

$s_C \cdot t$, (see section 2.2 for details). We optimize these parameters using the ribosome protected fragments accumulated at the stop codons of the 40 genes with the highest ribosome counts per length. The resulting values for the parameters are $a = 0.16$, $b = 17$, $s_A \cdot t = 14.2$, $s_U \cdot t = 8.4$, $s_G \cdot t = 0.4$, and $s_C \cdot t = 1$, respectively. Thus, we find that that the fastest cleavage rate s_A is 35 times faster than the slowest cleavage rate s_G, while the average of s_A and s_U is about 15 times faster than the average of s_G and s_C, consistent with previous reports on MNase bias (Hörz and Altenburger, 1981; Dingwall *et al.*, 1981).

In order to evaluate the validity of our model, we then apply our model with the parameters fit to the 40 genes with the highest ribosome counts per length to predict the distributions of 3' ends of reads near the stop codons of the next 20 genes (41 through 60) by ribosome counts per length (Figure 3) and compare the predictions to experimental observations. The predictions are visually very similar to experiment for at least 10 out of the 20 genes, such as *rpmG*, *rpsJ*, *rplW* and *eno*. Some genes, such as *hupB* and *rpsC*, have significantly unpredicted peaks further downstream. These peaks might be due to stop codon read through, i.e., due to ribosomes actually being located 3' of the annotated stop codon, which would not be captured by our model that for the purpose of this calculation assumes that all ribosomes are stalled right at the annotated stop codon and thus is a topic beyond the scope of discussion in this paper.

3.4 The identity of the nucleotide 5' of the cut site is much more important than the identity of the nucleotide 3' of the cut site

Since the cutting rate of MNase is known to largely depend on the identity of the nucleotide on the 5' side of the cut site (Hörz and Altenburger, 1981; Dingwall *et al.*, 1981), the cutting rates in our model are assumed to depend only on this nucleotide. However, in principle other factors of the library preparation could be responsible for the observed sequence bias. Thus, we asked how the MNase-inspired model with cutting rates depending solely on the identity of the nucleotide on the 5' side of the cut site compares to analogous models, where the rate either only depends on the identity of the nucleotide on the 3' side of the cut site or on the identity of both nucleotides surrounding the cut site. We thus repeated our analysis of the ribosome fragment distributions around stop codons using a model with four rates that depend on the nucleotide on the 3' side of the cut site and a model with 16 rates that depend on both nucleotides surrounding the cut site. We again optimized parameters on the 40 genes with the highest ribosome counts per length and then evaluated the validity of the models on the next 20 genes (41 through 60) by ribosome counts per length. Table 2 shows the optimal fitting parameters for all three models. In addition, the table shows the average Bhattacharyya distances (Kailath, 1967) between the distributions predicted by the model with the optimal parameters and the experimental ribosome protected fragment distributions in the vicinity of the stop codon separately for the 40 most highly translated genes (which have been used in the fitting) and the next 20 genes (which have not been considered in the fitting). The Bhattacharyya distance between two distributions measures their similarity and is zero if they are equal and infinity if the two distributions do not have any overlap.

We find that the model that only considers the nucleotide on the 3' side of the cutting side learns optimal parameters that show only very weak dependency on this nucleotide. The positional parameter b of the ribosome protection function Eq. (1) is fit to 12, which is in stark contrast to what is known about the distance between 3' ends of ribosome protected fragments and the ribosomal P-sites. The Bhattacharyya distances for this model indicate a much worse correspondence between model and experiment than the 5'-dependent model both for the 40 genes the model was fit to and for the next 20 validation genes. We conclude that the model that only takes into account the identity of the nucleotide on the 3' side of the cut site is not consistent with the data.

Table 2. Fitting parameters and fitting quality of different models for the 3' end distributions of ribosome protected fragments near stop codons. The columns correspond to models in which the cutting rate depends only on the nucleotide 5' of the cut site, only on the nucleotide 3' of the cut site, and on both nucleotides surrounding the cut site, respectively. The first two rows show the optimal values of the fitting parameters a and b for the slope and the position of the ribosome protection function defined in Eq. (1). The next four rows show the optimal values of the sequence dependent cutting rates. For the model that depends on the identities of both nucleotides surrounding the cut site, these rates are given as four columns corresponding to the four different nucleotides on the 3' side of the cut site. The last two rows provide the average Bhattacharyya distance (Kailath, 1967) between the predicted and the experimentally observed distributions of 3' ends of ribosome protected fragments in the vicinity of stop codons for the 40 most highly translated genes and the next 20 genes when ordered by overall translation, respectively.

	5'-dependent	3'-dependent	both sides			
			AN	UN	GN	CN
a	0.16	0.18	0.20			
b	17	12	17			
$s_{(N)A} \cdot t$	14.2	0.6	15.8	12.4	32.0	17.0
$s_{(N)U} \cdot t$	8.4	0.4	13.2	9.3	12.5	12.4
$s_{(N)G} \cdot t$	0.4	1.3	0.71	0.51	0.80	0.53
$s_{(N)C} \cdot t$	1.0	0.8	0.6	1.0	2.2	2.1
d_{1-40}	0.3264	0.4545	0.3495			
d_{41-60}	0.3561	0.5082	0.4079			

When the rates are allowed to depend on both nucleotides surrounding the cut site, the positional parameter b is again optimal at the same value of 17 as for the model that only considers the nucleotide on the 5' side of the cut site. The optimal cutting rates themselves show a clear pattern of strong dependence on the nucleotide on the 5' side of the cut site and weak dependence on the nucleotide on the 3' side of the cut site. The Bhattacharyya distance for the fragment end distributions of the 40 genes the model is fit to is comparable to the distance for the 5'-dependent model (note that the fitting itself is done by maximizing likelihood and not by minimizing Bhattacharyya distance). However, the Bhattacharyya distance of the distributions of the next 20 genes is quite a bit larger than the one for the 5'-dependent model hinting at the presence of some overfitting that impedes generalizability of the model to genes it was not trained on. We conclude that including a dependence of the rates on both nucleotides surrounding the cut site does not improve the performance of the model and the optimal parameters mostly revert to a dependence only on the nucleotide on the 5' side of the cut site. Thus, the original, 5'-dependent model is most consistent with the experimental data, in agreement with the known MNase biases.

3.5 Reconstructed ribosome coverages show reduced MNase bias compared to other methods

Currently, in the context of ribosome profiling studies on bacteria such as *Escherichia coli*, there are three broad categories of methods for mapping ribosome footprints: (1) the fixed offset method (Balakrishnan *et al.*, 2014), where a fixed distance from one extremity of the read, identified using the translation initiation or termination peak, is applied to all ribosome

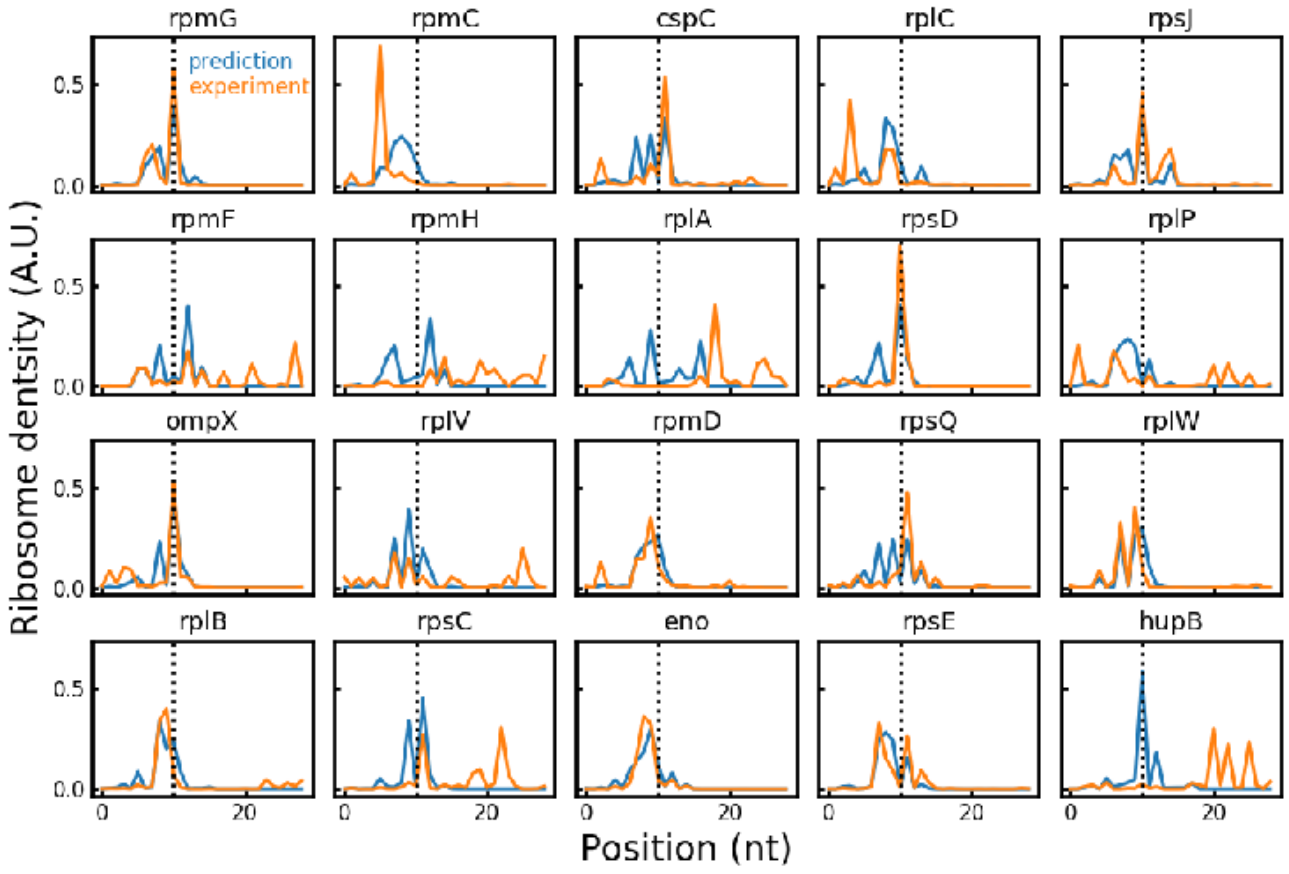

Fig. 3. 3' end distribution of ribosome protected fragments near the stop codon of the genes with the 41^{st} to 60^{th} highest ribosome counts per length (gene names shown above; the 40 genes with the highest ribosome counts per length were used to fit the model parameters and thus cannot be used in this verification of the model). 0 on the horizontal axis represents the first nucleotide 3' of the stop codon, and positive indices are toward the 3' direction of the genes. Blue: ribosome density profile predicted by our ribosome protection model based on the sequence alone. Orange: ribosome density profile from the ribosome profiling experiment by Balakrishnan et al., 2014. The vertical dashed lines at +10 represent the position of the peak from metagene analysis in reference (Balakrishnan et al., 2014), which is the site where the 3' ends of ribosome footprints near the stop codons from all genes accumulate on average.

protected reads to locate the position of the second nucleotide of the P site; (2) the variable offset method (Dunn and Weissman, 2016), which applies a specific fixed distance to the reads with specific length but allows for this distance to be different for different read lengths; (3) the center weighted method (Li *et al.*, 2012; Oh *et al.*, 2011), where the position of the second nucleotide of the P site associated with a read is fractionated over several consecutive positions, making each of the positions a possible location for the real second nucleotide of the P site. We compare our method to these three methods using the same approach as in section 3.1 (Figure 4A and B). We find that, compared to the fixed offset and variable offset methods, our method is capable of significantly reducing the effect of MNase bias. Figure 4A as an example, shows the average ribosome density on the 5' side of all GAA codons in the 41^{st} through 60^{th} most highly translated genes, with 0 being the first nucleotide of the codon. The reduced height of the upstream peak serves as an quantitative indicator of the reduced MNase bias. Figure 4B shows the difference between the highest and lowest average ribosome coverage in a -30 to +30 window for all codons and demonstrates the general ability of our method to reduce the MNase bias. The center weighted method has a similar capability of correcting MNase bias as our method. We also notice that the result of our method shows less periodicity than the fixed and variable offset method, which we believe is the consequence of the probability based nature of our algorithm. The loss of periodicity is also true for the center weighted method.

3.6 The model localizes ribosome pausing sites at codon resolution

While our method is comparable to the center weighted method in the ability to reduce MNase bias, one advantage that differentiates our method from the center weighted method is the ability to more accurately predict ribosome pausing sites at codon resolution. We look into two well known ribosome pausing sites: the stop codon and the peptide-mediated pausing site GGCCCU (encoding Gly165 and Pro166) in *secM*. Figure 4C shows the comparison at the stop codon for the 500 most highly translated genes. While all four methods show the ribosome pausing peak near the stop codon, one disadvantage of the center weighted method is that, due to its nature, it smears out the peak across several adjacent positions. Another example in Figure 4D shows the comparison at the extensively studied pausing site in *secM* (Nakatogawa and Ito, 2002; Bhushan *et al.*, 2011). Our method accurately predicts the P-site stalling site at Gly165, while the pausing site determined by the center-weighted method is spread to 5 to 6 nucleotides downstream. In this example, the fixed offset and variable offset methods, likely due to the effect of MNase bias, show the peak at a wrong position, which is off by about a codon. We conclude that our method allows more precise codon localization of ribosome density than the other methods, laying the groundwork to being able to differentiate slowly translating codons from fast translating codons.

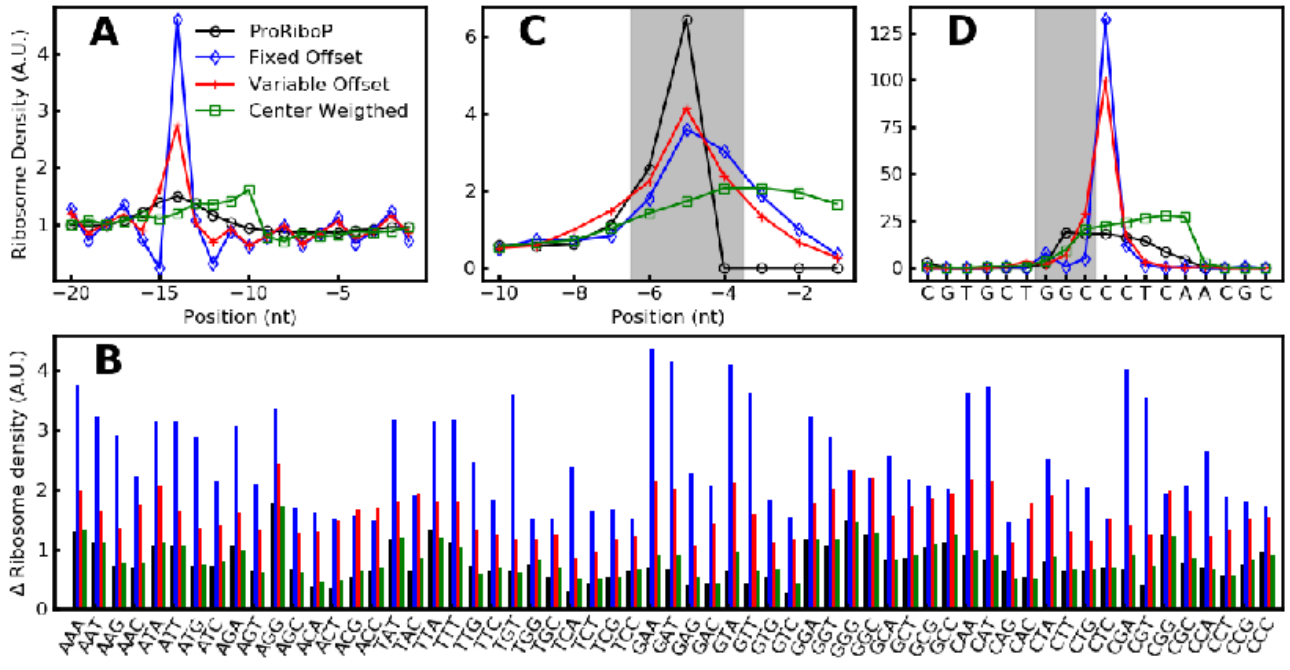

Fig. 4. A comparison between different mapping methods: RiboProP (black), fixed offset (blue), Variable offset (red) and center weighted (green). (A) Average ribosome density near all 6806 GAA codons in the 500 most highly translated genes with the first nucleotide of the GAA codon being at 0; the positive axis points in the 3' direction of the gene containing the GAA codon. The peak at 14 nt upstream of GAA from the fixed offset and variable offset methods is a quantitative indication of the MNase bias, which is significantly reduced in RiboProP and the center weighted method. (B) The difference between the highest and lowest ribosome density in a 60 nucleotide window around all 61 codons in the 500 most highly translated genes (colors represent the four different methods as in A). Large differences indicate large MNase biases. (C) Average ribosome density near stop codons in the 500 most highly translated genes. The shaded area indicates the position of the codon on the 5' side of the stop codon where the P site of stalled ribosomes should be located. (D) Average ribosome density near the peptide-mediated ribosome pausing site GGCCCU (encoding Gly165 and Pro166) in secM. The shaded area indicates the position of the GGC codon where the P site of the paused ribosome is located.

4 Discussion

In summary, we propose a new algorithm for mapping ribosome protected fragments from ribosome profiling studies on bacteria. In these experiments, micrococcal nuclease, although a biased nuclease, is considered to be the best option for digesting mRNAs, since unbiased nucleases such as RNAase I cannot be used in bacterial systems (Woolstenhulme *et al.*, 2015). The micrococcal nuclease digests the mRNAs in a sequence dependent manner, resulting in a broad length distribution of ribosome protected mRNA fragments and introducing challenges in accurately locating the corresponding P site positions for individual reads. Commonly used methods, including the fixed offset method, the variable offset method, and the center weighted method suffer from either strong sequence bias, which results in less accuracy in locating correct ribosome pausing sites, or a strong spreading out of ribosome density especially noticeable at ribosome pausing sites, which results in less accuracy in determining the relative ribosome pausing time at different locations. Our method has the ability to predict the ribosome pausing events at codon resolution while minimizing the effect of MNase bias. Since it focuses on sequence dependence of MNase digestion, it can in principle be combined with other sophisticated, but not sequence dependent methods of analyzing ribosome profiling data such as, e.g., (O'Connor *et al.*, 2016).

We note that while the motivation for our model is the known sequence bias of MNase, in the end our model cannot distinguish between MNase bias and any other possible sequence dependent biases that may be introduced at any step of the library preparation (such as, e.g., adapter ligation biases). The sequence-dependent cutting rates we find and the fact that they largely depend on the nucleotide 5' of the cut site as described in section 3.4 are consistent with the known behavior of MNase (Hörz and Altenburger, 1981; Dingwall *et al.*, 1981). Also the fact that these effects do not seem to play a major role in eukaryotic ribosome profiling, where RNAase I is used instead of MNase seem to point toward MNase being the main reason for the biases corrected by our model. However, in the end our model is able to accomodate sequence biases at the 3' end of fragments no matter their origin, making the demonstrated improvements in ribosome localization independent of the reason for the sequence biases.

One notable shortcoming of our method is that we restrict the prediction region on a gene from the 4th nucleotide upstream of the start codon to the stop codon. Our method does not include potential ribosome occupancy further upstream of the start codon (possible uORFs) or downstream of stop codons (read through). Both areas are considered to be important applications of ribosome profiling studies. Ideally, one could apply our algorithm to predict the ribosome occupancy both upstream and downstream of the gene, however, the detailed mechanisms of possible translation at uORFs and locations beyond the stop codon are not clear at this point, thus the basis of our algorithm, the ribosome protection and MNase digestion model, may or may not entirely hold.

We believe that our approach can also be more broadly applied to other sequencing experiments involving the usage of biased enzymes for digestion. An example are MNase-seq experiments for determining nucleosome positions within chromatin (Barski *et al.*, 2007). Studies show that the sequence specificity of MNase may affect the accuracy of the nucleosome positioning data from MNase-seq experiments and thus a mitigation method is needed (Chung *et al.*, 2010). Another example is the DNAse-seq experiment for determining transcription factor binding sites (He *et al.*, 2014). DNAse is also known to digest the DNA in a sequence dependent way and mitigation of the sequence bias may be beneficial.

Funding

This material is based upon work supported by the National Science Foundation under Grants No. DMR-1410172, MCB-1614990, and DMR-1719316.

References

Balakrishnan, R., Oman, K., Shoji, S., Bundschuh, R., and Fredrick, K. (2014). The conserved GTPase LepA contributes mainly to translation initiation in escherichia coli. *Nucleic Acids Res.*, **42**, 13370–13383.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Bhushan, S., Hoffmann, T., Seidelt, B., Frauenfeld, J., Mielke, T., Berninghausen, O., Wilson, D. N., and Beckmann, R. (2011). SecM-stalled ribosomes adopt an altered geometry at the peptidyl transferase center. *PLoS Biology*, **9**, e1000581.

Chung, H.-R., Dunkel, I., Heise, F., Linke, C., Krobitsch, S., Ehrenhofer-Murray, A. E., Sperling, S. R., and Vingron, M. (2010). The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One*, **5**(12), e15754.

Datta, A. K. and Burma, D. P. (1972). Association of ribonuclease I with ribosomes and their subunits. *J. Biol. Chem.*, **247**(21), 6795–6801.

Dingwall, C., Lomonossoff, G. P., and Laskey, R. A. (1981). High sequence specificity of micrococcal nuclease. *Nucleic Acids Res.*, **9**, 2659–2673.

Dunn, J. G. and Weissman, J. S. (2016). Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics*, **17**, 958.

Gong, F. and Yanofsky, C. (2002). Instruction of translating ribosome by nascent peptide. *Science*, **297**, 1864–1867.

He, H. H., Meyer, C. A., Hu, S. S., Chen, M.-W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., and Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.

Hörz, W. and Altenburger, W. (1981). Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Res.*, **9**(12), 2643–2658.

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, **15**, 205–213.

Ingolia, N. T. (2016). Ribosome footprint profiling of translation throughout the genome. *Cell*, **165**, 22–33.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**(5924), 218–223.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.*, **15**(1), 52–60.

Li, G.-W., Oh, E., and Weissman, J. S. (2012). The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–541.

Nakatogawa, H. and Ito, K. (2002). The ribosomal exit tunnel functions as a discriminating gate. *Cell*, **108**, 629–636.

O'Connor, P. B., Andreev, D. E., and Baranov, P. V. (2016). Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.*, **7**, 12915.

Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. A., Kramer, G., *et al.* (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, **147**(6), 1295–1308.

Woolstenhulme, C. J., Guydosh, N. R., Green, R., and Buskirk, A. R. (2015). High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep.*, **11**(1), 13–21.