Identifying A- and P-site locations on ribosome-protected mRNA fragments using Integer Programming

Nabeel Ahmed[1, ¶], Pietro Sormanni[2, ¶], Prajwal Ciryam[2, #], Michele Vendruscolo[2], Christopher M. Dobson[2] and Edward P. O'Brien[1, 3, 4, *]

[1]Bioinformatics and Genomics Graduate Program, The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA

[2]Department of Chemistry, University of Cambridge, Cambridge, UK

[3]Department of Chemistry, Pennsylvania State University, University Park, PA, USA

[4]Institute for Cyber Science, Pennsylvania State University, University Park, PA, USA

[¶] These authors contributed equally to this work.

[#] Present address: Department of Neurology, Columbia University College of Physicians and Surgeons, New York, NY, USA

*To whom correspondence should be addressed. Tel: (814) 867-5100; Fax: (814) 865-2927; Email: epo2@psu.edu

Abstract

Identifying the A- and P-site locations on ribosome-protected mRNA fragments from Ribo-Seq experiments is a fundamental step in the quantitative analysis of transcriptome-wide translation properties at the codon level. Many analyses of Ribo-Seq data have utilized heuristic approaches applied to a narrow range of fragment sizes to identify the A-site. In this study, we use Integer Programming to identify A-site by maximizing an objective function that reflects the fact that the ribosome's A-site on ribosome-protected fragments must reside between the second and stop codons of an mRNA. This identifies the A-site location as a function of the fragment's size and its 5□ end reading frame in Ribo-Seq data generated from *S. cerevisiae* and mouse embryonic stem cells. The correctness of the identified A-site locations is demonstrated by showing that this method, as compared to others, yields the largest ribosome density at established stalling sites. By providing greater accuracy and utilization of a wider range of fragment sizes, our approach increases the signal-to-noise ratio of underlying biological signals associated with translation elongation at the codon length scale.

Introduction

Translation is a fundamental cellular process and an important step of gene expression resulting in the production of proteins in cells[1]. In the past decade the advent of Ribo-Seq (also known as Ribosome profiling), a high-throughput Next-Generation Sequencing method[2,3], has enabled the transcriptome-wide study of translation. Ribo-Seq involves rapidly halting translation in cells through the use of antibiotics or flash freezing followed by cell lysis and then digestion of the lysate using an RNase enzyme[4]. The resulting pool of ribosome-protected mRNA fragments is then amplified and sequenced. The number and length of mRNA fragments that map to the coding sequences (CDSs) of transcripts is a function of the location and number of ribosomes that were sitting at a particular location on different copies of the same transcript. Where the ribosome's A- and P-sites were located on a fragment during the digestion step is not known *a priori*, additional information and assumptions must be introduced to estimate their locations. Since translation occurs at the A- and P-sites, the identification of these sites is critical to address translation-related questions. If the A- and P-sites are not accurately identified, then systematic or random error can diminish the statistical power of any underlying biological signal that might exist. The identification of the A- and P-sites within ribosome footprints is therefore fundamental to quantitatively understanding translation at the codon length scale.

Because of the importance of this assignment problem, a number of methods for identifying the A- and P-sites have been created[2,5–13]. Many of these approaches utilize the biological fact that only the P-site is permitted to occupy the start codon during translation initiation and only the A-site is permitted to occupy the stop codon during termination. Using such approaches, the A-site location in *S. cerevisiae* Ribo-Seq datasets, for example, has been estimated to be 15 nt from the 5 end of ribosome-protected mRNA fragments of size 28 nt[2,14]; 16 nt for fragment size 29 nt[14]; 15 nts from the 5 end of fragments that are 30 nt in length[15] and frame-specific offsets of 14 to 17 nts from the 5 end for fragments between 28 and 30 nt in length[12,16]. The P-site location offset is 3 nt prior to the A-site. Similarly, in mouse embryonic stem cells (mESCs), such approaches have yielded specific offsets for different fragment lengths[11].

Here, we utilize the fundamental biological fact that the A-site on ribosome-protected fragments must reside within the CDS of a gene under normal growth conditions and without any upstream open reading frames. We use this fact to create an objective function that, when maximized, identifies where the ribosome's A- and P-sites are most likely to be located on a ribosome-protected mRNA fragment. We apply our method to *S. cerevisiae* and mESCs Ribo-Seq datasets and show that, compared to other methods, our approach has greater accuracy and statistical power in identifying A- and P-site locations and assigning read density.

Methods

Integer Programming Algorithm

2

66   In the analysis of Ribo-Seq data, mRNA fragments are initially aligned onto the reference transcriptome
67   and their location is reported with respect to their 5ꞌ end. This means that one fragment will contribute
68   one read that is reported on the genome coordinate to which the 5ꞌ end nucleotide of the fragment is
69   aligned (Fig. 1A). In Ribo-Seq data, fragments of different lengths are observed that can arise from
70   incomplete digestion of RNA and from the stochastic nature of mRNA cleavage by the RNase used in the
71   experiment (Figs. 1C-D, Supplementary Fig. S1). A central challenge in quantitatively analyzing Ribo-Seq
72   data is to identify from these Ribo-Seq reads where the A- and P-sites were located at the time of
73   digestion. It is non-trivial to do this since incomplete digestion and stochastic cleavage can occur at both
74   ends of the fragment. For example, mRNA digestion resulting in a fragment of size 29 nt can occur in
75   different ways, two of which are illustrated in Fig. 1B. The quantity that we need to accurately estimate is
76   the number of nucleotides that separate the codon in the A-site from the 5ꞌ end of the fragment, which we
77   refer to as the offset and denote Δ. Knowing Δ determines the position of the A-site as well as the P-site
78   since the P-site will always be at Δ minus 3 nt.

79   Our solution to this problem relies on the biological fact that for canonical transcripts with no upstream
80   translation the A-site of actively translating ribosomes must be located between the second codon and
81   stop codon of the CDS of a transcript[17]. Therefore, the optimal offset value Δ for fragments of a particular
82   size (S) and reading frame (F) that map onto gene i is the one that maximizes the total number of
83   reads $T(\Delta|i, S, F)$ between these codons. The size of an mRNA fragment S is measured in nucleotides,
84   and the frame F has values of 0, 1 or 2 and corresponds to the frame in which the 5ꞌ end nucleotide of
85   the fragment is located. The 5ꞌ end frame F is a result of RNase digestion and it is distinct from the
86   reading frame of the ribosome that is typically translating in-frame (frame 0 of A-site). This concept can be
87   expressed in terms of Integer Programming[18], a mathematical optimization procedure in which an
88   objective function is maximized subject to integer and linear restraints. With Δ as the integer variable to
89   optimize, the objective function in this case is $T(\Delta|i, S, F) = \sum_{j=4}^{N_{C,i}} R(j, \Delta|i, S, F)$, where $N_{C,i}$ is the number of
90   nucleotides in the CDS of gene i and $R(j, \Delta|i, S, F)$ is the number of reads from fragments of size S and
91   frame F mapped onto gene i whose 5ꞌ end is at nucleotide position j on the CDS after being shifted
92   along the transcript by Δ nucleotides. The optimal Δ, denoted Δ', for a given (S, F) for gene i is determined
93   as $\max\{T(\Delta|i, S, F)\}$ subject to the constraints (*i*) that $0 \leq \Delta \leq S$, and (*ii*) that the modulus of $\frac{\Delta}{3} = 0$.
94   Constraint (*i*) enforces the requirement that the A-site is located between the first and last nucleotide of
95   the fragment of size S nts. Constraint (*ii*) maintains the frame of the 5ꞌ-most nucleotide of the fragment as
96   the Ribo-Seq reads are shifted by an amount Δ. We enforce Constraint (*ii*) because we are interested in
97   the assignment of reads to the A-site at the resolution of a codon, not an individual nucleotide. If we did
98   not enforce constraint (*ii*), our algorithm would simply yield equal $T(\Delta|i, S, F)$ scores for the two other
99   values of Δ that would still map the reads on the A-site codon, but in the two frames where the 5ꞌ end
100   was not in. Therefore, to simplify the determination of offsets we implemented constraint (*ii*). Thus, by
101   maximizing $T(\Delta|i, S, F)$ for the CDS of each gene in a data set of N_g genes, we will obtain a set of N_g

3

102   values of Δ'. From this distribution of Δ' values, the A-site location corresponds to the most probable Δ'
103   value.

104   While identifying the Δ' value for each gene in our data set, we also minimize the occurrence of false
105   positives by ensuring that the highest score, $T(\Delta'|i, S, F)$, is significantly higher than the next highest
106   score, $T(\Delta''|i, S, F)$, which occurs at a different offset Δ''. If the difference between the top two scores is
107   less than the average number of reads per codon, we apply the following additional selection criteria. To
108   choose between Δ' and Δ'', we select the one that yields a number of reads at the start codon that is at
109   least one-fifth less than the average number of reads at the second, third and fourth codons. We further
110   require that the second codon have a greater number of reads than the third codon. The biological basis
111   for these additional criteria are that the true offset (*i.e.,* the actual location of the A-site) cannot be located
112   at the start codon, and that the number of reads at the second codon should be higher on average than
113   the third codon due to contributions from the initiation step of translation, during which the ribosome is
114   assembling on the mRNA with the start codon in the P-site. In the Results section, we demonstrate that
115   the results from our method are largely robust to changes in these thresholds.

Ribo-Seq datasets

117   ***S. cerevisiae.*** Published Ribo-Seq data from *S. cerevisiae* were obtained from GSM1557447 used in the
118   study of Pop and co-workers[19]. The raw reads were pre-processed according to the method stated in the
119   original study. Raw fastq files were downloaded and preprocessed using Fastx-toolkit (v0.013)
120   (http://hannonlab.cshl.edu/fastx_toolkit/index.html) as stated in the methods of the original study. The
121   adapter sequence CTGTAGGCACCATCAAT was stripped using FastQ clipper and low-quality reads
122   were filtered by FastQ quality filter. The processed reads were aligned first to the ribosomal RNA
123   sequences using Bowtie 2 (v2.2.3)[20]. The reads which did not align to the ribosomal sequences were then
124   aligned to the *Saccharomyces cerevisiae* assembly R64-2-1 (UCSC: sacCer3) using Tophat (v2.0.13)[21]
125   with up to two mismatches allowed. Gene annotations were obtained from Saccharomyces Genome
126   Database (http://www.yeastgenome.org/) on May 4, 2016 for 6,572 protein-coding genes. Reads were
127   assigned to the nucleotide positions according to the 5ᵇ end.

128   The pooled Ribo-Seq dataset was formed by combining reads from all replicates *of S. cerevisiae*
129   Ribo-Seq data published in studies in which cycloheximide (CHX) was not used to induce translation
130   arrest[14–16,19,22–28]. It has been demonstrated that CHX pre-treatment leads to distortion of ribosome
131   profiles due to ribosome slippage even after CHX treatment[12,22]. The distorted ribosome profiles can spill
132   across the CDS boundaries thus limiting the application of Integer Programming algorithm. Hence, our
133   analysis only used those datasets without CHX pre-treatment. The list of all the utilized datasets is
134   reported in Supplementary Table S1. The raw reads from each study were processed according to the
135   reported method in the original study. If the method is not reported in the original study, we used cutadapt
136   (v1.14)[29] to pre-process the raw reads. The alignment and assignment of reads to gene transcripts was
137   done as above for the Pop dataset[19].

138   **Mouse embryonic stem cells.** The "no drug" sample for mouse embryonic stem cells (mESCs)

139   measured by Ingolia and co-workers[11] was utilized in this study. Since CHX treatment has been shown to

140   artificially alter ribosome profiles in *S. cerevisiae*, we believed it prudent to not use mESC samples pre-

141   treated with CHX. To increase the coverage we pooled reads from another untreated Ribo-Seq sample

142   of mESCs published in the study of Hurt and co-workers[30]. The linker sequence

143   CTGTAGGCACCATCAATTCGTATGCCGTCTTCTGCTTGAA for Ingolia's dataset and the poly-A adapter

144   sequence for Hurt's dataset were trimmed using cutadapt (v1.14)[29]. The trimmed reads were first aligned

145   to ribosomal RNA sequences using Bowtie2 (v2.2.3)[20] and the filtered reads were subsequently aligned to

146   mm10 reference transcriptome consisting of 21,185 genes obtained from UCSC knownGene database

147   using Tophat (v2.0.13)[21] with up to two mismatches allowed. For a gene with multiple isoforms, only the

148   isoform with the longest CDS was included in the reference transcriptome. For transcripts with no

149   information on the 5ʹ UTR region, we included 40 nt of genomic sequence upstream from the start codon

150   for successful alignment of reads around start codon and effective application of Integer Programming

151   algorithm. Translation initiation site data was obtained from Table S3 of study of Ingolia and co-

152   workers[11]. We selected genes that have only one translation initiation site coding for only a canonical

153   CDS product. From these genes, only genes containing a single isoform were selected, resulting in 430

154   genes in our final dataset.

155   ***Escherichia coli.*** Wild-type Ribo-Seq data for *E.coli* were obtained from studies of Li and co-workers

156   (2012)[31], Li and co-workers (2014)[32] and Woolstenhulme and co-workers[33]. The accession numbers of

157   the samples used are provided in Supplementary Table S1. The respective linker sequences in each

158   sample were trimmed using cutadapt (v1.14)[29]. Reads were initially aligned to ribosomal RNA sequences

159   using Bowtie2 (v2.2.3)[20] and the rest of reads aligned to the *E.coli* reference genome build NC_000913.3

160   using Tophat (v2.0.13)[21] with up to two mismatches allowed. Gene annotations were obtained for 4314

161   genes from RefSeq database corresponding to NC_000913.3.

162   ## Gene selection, analyses and statistical tests

163   **Selection of genes.** To obtain good sampling statistics, we selected for analysis only those genes that

164   have on average greater than 1 read per codon per fragment length per reading frame. This means that

165   different sets of genes can be used in the Integer Programming algorithm depending on the fragment

166   length and frame under scrutiny. The average number of reads per codon was calculated on the CDS

167   region of the gene and an additional upstream region corresponding to the size of the fragment length

168   being considered. Genes in which more than 1% of the total number of mapped reads, for a given

169   S and F, mapped to multiple locations across the genome were discarded from further analysis.

170   **Identifying unique offsets.** We defined the most probable offset Δ' to have a unique, unambiguously

171   identified A-site if at least 70% of genes in the dataset had an offset equal to Δ', and further require that

172   there be at least 10 genes in the dataset. Otherwise, the A-site location is defined as ambiguous for the

5

173   fragment size and frame under scrutiny. In the Results section, we show the A-site location is largely
174   robust to moderate variation in this 70% threshold.

175   **High coverage test.** To test for the effect of depth of coverage on the A-site location we increased the
176   average number of reads per codon required for a gene to be included in the analyzed dataset from 1 to
177   values up to 50. Three requirements have to be met for an ambiguous offset to be identified as unique as
178   coverage is increased. As before, 70% of the genes had to have the most probable offset with at least 10
179   genes in the dataset. In addition, there must to be a statistically significant increasing trend in the most
180   probable offset with increasing coverage. This requirement prevents fluctuations above 70% due to
181   statistical error as being counted as a unique offset. This trend is calculated using Linear Regression
182   Analysis.

183   **Test using Artificial Ribo-Seq data:** To construct artificial ribosome occupancies, we used Gillespie's
184   algorithm[34] to simulate translation across *S. cerevisiae* mRNA transcripts. During the simulations, we
185   saved snapshots every X steps recording the A-site codon location and creating a histogram of ribosome
186   occupancies across the transcript. To be consistent with the sampling statistics of the experimental
187   Pooled *S. cerevisiae* data, we carried out our analysis on the same 4,487 transcripts that met our filtering
188   criteria for $(S, F) = (28, 0)$, and normalized our simulated ribosome occupancies such that they sum up
189   to the total number of reads mapped to that transcript in the experimental data. We then created different
190   fragment size and 5⊡ end reading frame distributions (Supplementary Fig. S3A, B). Specifically, since the
191   reads are counts, we use Poisson statistics by treating each (S, F) as an event in the order:
192    $(20, 0), (20, 1), (20, 2), \dots, (35, 0), (35, 1)$ and $(35, 2)$. Six Poisson distributions of different variances
193    $(\lambda = 4, 8, 16, 24, 48, 80)$ were generated. The distributions were shifted such that the mode of the
194   distribution was at $(S, F) = (28, 0)$, which is typically found in experiments, with probabilities summing up
195   to 1 between $(20, 0)$ and $(35, 2)$. Two additional read length distributions were also considered with
196   modes at $(S, F) = (24, 0)$ and $(S, F) = (32, 0)$ with $\lambda = 8$. Four different sets of offset tables were used
197   as an input to generate the artificial Ribo-Seq reads from the simulated ribosome occupancies for each of
198   these distributions. These four offset sets are i) a constant offset of 15 nucleotides for all (S, F)s, ii) a
199   constant offset of 18 for all (S, F)s iii) a constant offset of 12 for $S = 20, 21, \dots, 26, 27$ and constant offset
200   of 18 for $S = 28, 29, \dots, 34, 35$ iv) the "top offset" values for (S, F) combinations identified using our
201   algorithm in the experimental Pooled *S. cerevisiae* data (*i.e.,* the offset values of Table 1). These input
202   offset tables were compared to the 'output' offset table generated by applying the IP algorithm on the
203   artificial Ribo-Seq data to test the correctness of our method.

204   **Statistical significance of PPX and XPP motifs.** To test if the normalized read density distribution of a
205   PPX or XPP motif is not due to random chance, we calculated the *P*-value using a permutation test[35]. For
206   the total number of instances of a PPX/XPP motif, we randomly selected an equal number of instances of
207   any other three-residue motif and determined the median normalized read density at the third codon

6

208   position of the motif, thereby creating a random distribution. We repeated this procedure 10,000 times

209   and calculated the fraction of iterations that had a median density equal to or greater than the one

210   observed for that PPX/XPP motif. This fraction is equal to the *P*-value. The instances of PPX and XPP

211   motifs are identified from those transcripts that have at least 50% of codon positions with 1 read or more.

212   **Comparison with other A-site mapping methods.** We compared the performance of Integer

213   Programming algorithm with other methods by calculating the difference in normalized read density

214   between the Integer Programming A-site value and the compared method's A-site value at the third

215   codon of PPG and PPE motifs, which are associated with ribosome pausing in *S. cerevisiae* and mESCs

216   respectively.

217   In *S. cerevisiae*, A-site ribosome profiles were obtained for Integer Programming method by applying

218   the offsets listed in Table 1 for fragment sizes 24 to 34 nt. For methods used by Martens and co-workers[5]

219   and Hussmann and co-workers[12] specifically in *S. cerevisiae*, A-site profiles were obtained by applying

220   the offsets for specific fragment sizes as stated in the Methods sections of those studies. We included a

221   constant heuristic offset of 15 nt which has been used in several studies of *S. cerevisiae* Ribo-Seq

222   data[2,36–38]. The constant offset of 15 nt has been applied to a wide range of fragment lengths across

223   studies including 22-32 nt[2], 27-30 nt[36], 28 nt[37], 27-34 nt[38]. To be conservative, we apply a constant offset

224   of 15 nt to fragments between 27 and 30 nt only. Similarly, we also include a method where a constant

225   offset of 18 nt is applied to fragments between 27 and 30 nt to compare to the performance of the Integer

226   Programming method.

227   For mESCs, Ingolia and co-workers[11] implemented length specific offsets of 15, 16 and 17 nts from

228   the 5ꞌ end, respectively, for fragments of size 29-30 nt, 31-33 nt and 34-35 nt. Several studies have also

229   implemented a constant offset of 15 for range of fragment sizes 25-35 nt[39,40]. Similar to *S. cerevisiae*, we

230   also implement a constant offset of 18 nt to fragment size range of 25-35 nt.

231   Few general methods have been proposed to determine A-site locations in any organism. We

232   implemented the methods Plastid[7], RiboProfiling[8] and riboWaltz[9] which are publicly available as R

233   packages. The A-site offset tables generated using these methods for our analyzed datasets in *S.*

234   *cerevisiae* and mESCs are presented in Supplementary Table S9. To determine the A-site profiles using

235   the 'ribodeblur' method created by Wang and co-workers[6], we ran the source code available in GitHub

236   (https://github.com/Kingsford-Group/ribodeblur-analysis/releases/tag/v0.1) on our datasets and added a

237   custom Python script to generate the 'deblurred' A-site profiles. For Rpbp[41], the publicly available

238   software was downloaded and run locally to obtain the A-site offsets. We also applied the center-

239   weighted method as described by Becker and co-workers[42]; for reads greater than 23 nt, we trim 11 nt

240   from both ends of the fragment and distribute the read equally among the remaining nucleotides. For

241   scikit-ribo method[10], the source code was downloaded and was successfully run for *S. cerevisiae*

242   datasets to obtain the A-site profiles. Scikit-ribo could not be run on mouse ESC data as the current

243   available version of the source code contains bugs resulting in inaccurate annotation assignments for

244   higher eukaryotic genomes.

7

245   Instances of PPG motifs (in *S. cerevisiae*) and PPE motifs (in mESCs) used for analysis are selected

246   from genes in which at least 90% of codon positions have at least 1 read in their 5ʹ aligned ribosome

247   profiles in the CDS region and an upstream region of 18 nt. An instance of a motif is included for analysis

248   only if its ribosome density is greater than 1.5 of average ribosome density at the third codon position in

249   the A-site profile of any compared methods. We use the Wilcoxon signed rank test to determine if there is

250   a statistically significant difference between the normalized read density at the third codon of motif

251   instances obtained by Integer Programming and other methods.

Results

Illustrating the Integer Programming optimization procedure

254   To illustrate this Integer Programming algorithm in action we provide an example using the hypothetical

255   mRNA shown in Figure 2. The algorithm is as follows: First, for gene i, consider $RP(j, \Delta = 0|i, S, F)$

256   composed of those fragments of size S $(= [20, 21, ..., 35]$ nt$)$ and whose 5ʹ end has been aligned to

257   reading frame F $(= 0, 1$ or $2)$. Second, for this ribosome profile, determine the Δ that

258   maximizes $T(\Delta|i, S, F)$. Do this by starting from the 5ʹ-end-aligned ribosome profile (Δ=0) and shift it three

259   nucleotides at a time (*i.e.*, obey Constraint 2 described in Methods) towards the 3ʹ end of the transcript

260   such that $\Delta = 0, 3, 6, 9, ..., \leq S$. At each value of Δ, calculate $T(\Delta|i, S, F)$ and record its value. Third, after all

261   Δ values have been tested, the Δ that maximizes $T(\Delta|i, S, F)$ is denoted Δ', which is the putative location of

262   the A-site relative to 5ʹ end of fragments of size S and frame F for gene i. Check if the secondary-

263   selection criteria are required and apply them when the scores for the top two offsets differ by less than

264   the average number of reads per codon in the mRNA. Finally, repeat these steps for every fragment size

265   between 20-35 nts in length and every reading frame. Thus, for one gene, this procedure yields 48

266   (=16x3) independent values for Δ', one for each fragment size and frame combination.

267   The fragment-size and frame distributions of ribosome-protected fragments (Figs. 1C, D) in *S.*

268   *cerevisiae* are not gene dependent (Supplementary Fig. S2), and therefore, neither should be the offset

269   values. Thus, the location of the A-site, relative to the 5ʹ end of a fragment of size S and frame F,

270   corresponds to the most probable value of the offset across all the genes in the dataset.

A-site locations in *S. cerevisiae* Ribo-Seq data are fragment size and frame dependent

273   We first applied the Integer Programming method to Ribo-Seq data from *S. cerevisiae* published by Pop

274   and co-workers[19]. For each combination of S and F we first identified those genes that have at least 1

275   read per codon on average in their corresponding ribosome profile. The number of genes meeting this

276   criterion is reported in Supplementary Table S2. We then applied the Integer Programming method to this

277   subset of genes. The resulting distributions of Δ values are shown in Fig. 3A for different combinations of

278   fragment length and frame. We only show results for fragment sizes between 27 and 33 nt because

8

279 greater than 90% of reads map to this range (Fig. 1C). The most probable offset value for all fragment
280 sizes between 20 to 35 nt is reported as an offset table (Supplementary Table S4).

281   We see that the optimal Δ value - that is, the A-site location - changes for different combinations of S
282 and F, with the most probable values either at 15 or 18 nt. Thus, the location of the A-site depends on S
283 and F. In most cases, there is one dominant peak for a given pair of S and F values. For example, for
284 fragments of size 27 through 30 nt in frame 0, greater than 70% of their per-gene optimized Δ values are
285 15 nt from the 5⃞ end of these fragments. Similar results are found for other combinations such as sizes
286 30, 31 and 32 nt in frame 1 and 28 through 32 nt in frame 2, where optimized Δ values are 18 nt. Thus,
287 across the transcriptome, the A-site codon position on these fragments is uniquely identified.

288   There are, however, S and F combinations that have ambiguous A-site locations based on these
289 distributions. For example, for fragments of size 27 nt in frame 1, 47% of the gene-optimized Δ values are
290 at 15 nt while 30% are at 18 nt. Similar results are observed for fragments 28 and 29 nt in frame 1, and
291 31 and 32 nt in frame 0. Thus, for these S and F combinations there is a similar probability of the A-site
292 being located at one codon or another, and therefore we cannot uniquely identify the A-site's location.

Higher coverage leads to more unique offsets

294 We hypothesized that ambiguity in identifying the A-site for particular S and F combinations may be due
295 to low coverage (*i.e.,* sampling poor statistics). To test this hypothesis we pooled the reads from different
296 published Ribo-Seq datasets into a single dataset with consequently higher coverage and more genes
297 that meet our selection criteria (Supplementary Table S2). Application of our method to this Pooled
298 dataset gives unique offsets for more S and F combinations compared to the original Pop dataset (Fig. 3B
299 and Supplementary Table S4), supporting our hypothesis. For example, for fragments of size 27 and
300 frame 1, now we have the unique offset of 15 nt with 72% of gene-optimized Δ values at 15 nt (Fig. 3B).
301 However, we still see the ambiguity present for certain (S, F) combinations.

302   We employed an additional strategy to increase coverage by restricting our analysis to genes with
303 greater and greater average reads per codon. If the hypothesis is correct, then we should see a
304 statistically significant trend of an increase in the most probable Δ value with increasing read depth. We
305 applied this analysis to the Pooled dataset and find that some initially ambiguous S and F combinations
306 become unambiguous as coverage increases. For example, at an average of 1 read per codon, (S, F)
307 combinations of $(25, 0)$, $(27, 2)$ and $(30,1)$ are ambiguous as they fall below our 70% threshold. However,
308 we see a statistically significant trend $(slope = 0.5, \ p \ = \ 3.94 \times 10^{-6})$ for fragments of $(25, 0)$ that the 15
309 nt offset becomes more probable upon increasing the coverage, eventually crossing the 70% threshold
310 (Fig. 4A). Similarly, for $(27, 2)$ $(slope = 0.58, \ p = 5.77 \times 10^{-5})$ and $(30,1)$ $(slope = 0.25, \ p = 0.009)$ there
311 is a trend towards an offset of 18 nt, with more than 70% of genes having this offset at the highest
312 coverage (Figs. 4B, C). Hence, for these fragments, increasing coverage uniquely identifies Δ' and hence
313 the A-site location. For a few combinations of (S, F), like $(32, 0)$, the ambiguity is not resolved even upon

9

314   very high coverage (Fig. 4D), which we speculate may be due to inherent features of nuclease digestion

315   being equally likely for more than one offset.

316   Thus, high enough coverage yields the optimal offset table represented in Table 1, where the offset is

317   the most probable location of the A-site relative to the 5⊠ end of the mRNA fragments generated in *S.*

318   *cerevisiae.*

Consistency across different datasets

320   Ribo-Seq data is sensitive to experimental protocols that can introduce biases in the digestion and

321   ligation of ribosome-protected fragments. Pooling datasets together offers the advantage of higher

322   coverage but it may mask the biases specific to an individual dataset. To determine whether our unique

323   offsets (Table 1) are consistent with results from individual data sets we applied the Integer Programming

324   algorithm to each individual dataset. Most of these datasets have low coverage resulting in fewer genes

325   meeting our filtering criteria (Supplementary File S1). For each unique offset in Table 1, we classify it as

326   consistent with an individual data set provided that the most probable offset from the individual dataset

327   (even if it does not reach the 70% threshold due to limitations in the depth of coverage) is the same as in

328   Table 1. We find that the vast majority of unique offsets (18 out of 20) in Table 1 are consistent across

329   75% or more of the individual datasets (statistics reported in Supplementary Table S5). Just two (S, F)

330   combinations show frequent inconsistencies. (S, F) combinations $(27, 1)$ and $(27, 2)$ are inconsistent in

331   33% or more of the individual datasets (Supplementary Table S5). This suggests that researchers who

332   wish to minimize false positives should discard these (S, F) combinations when creating A-site ribosome

333   profiles.

Robustness of the offset table to threshold variation

335   The Integer Programming algorithm utilizes two thresholds to identify unique offsets. One is that 70% of

336   genes exhibit the most probable offset, the other, designed to minimize false positives arising due to

337   sampling noise in the Ribo-Seq data, is that the reads in the first codon be less than one-fifth of the

338   average reads in the second, third and fourth codon. While there are good reasons to introduce these

339   threshold criteria, the exact values of these thresholds are arbitrary. Therefore, we tested whether varying

340   these thresholds changes the results reported in Table 1. We varied the first threshold to 60% and 80%,

341   and recomputed the offset table. We report whether the unique offset changed by listing an 'R' or 'S' (for

342   robust and sensitive, respectively) alongside the reported offset in Supplementary Table S5. We find that

343   two-thirds of the unique (S, F) combinations do not change (Supplementary Table S5). (S, F)

344   combinations $(25, 0), (25, 2), (27, 0), (27, 1), (28, 1), (31, 0), (33, 0)$ and $(33, 2)$ become ambiguous when

345   we increased the threshold to 80%.

346   We varied the second, aforementioned threshold from one-fifth up to one and down to one-tenth, and

347   we find that all unique (S, F) combinations except $(25, 2), (33, 0), (33, 2)$ and $(34, 1)$ remain unchanged

10

348   (reported as 'R' in Supplementary Table S5). Thus, in summary, in the vast majority of cases, the unique

349   offsets reported in Table 1 depend very little on specific values of these thresholds.

Testing the Integer Programming algorithm against artificial Ribo-Seq data

351   To test the correctness and robustness of our approach we generated a dataset of simulated ribosome

352   occupancies across 4,487 *S. cerevisiae* transcripts and asked whether our method could accurately

353   determine the A-site locations. Artificial Ribo-Seq reads were generated from these occupancies

354   assuming a Poissonian distribution in their (S, F) values using random footprint lengths similar to that

355   found in experiments (see Methods and Supplementary Fig. S3A, B). We investigated the ability of our

356   method to correctly determine the true A-site locations for four different sets of pre-defined offset values

357   (see Methods). The Integer Programming algorithm was then applied to the resulting artificial Ribo-Seq

358   data. We find the offset table generated from the algorithm reproduces the input offsets used

359   (Supplementary Fig. S3C and Supplementary Table S6). This procedure was repeated for different read

360   length distributions as well as with different input offsets and we find that the offset tables generated by

361   our algorithm reproduce the input offset tables in greater than 93% of all (S, F) combinations

362   (Supplementary Fig. S3B, C and Supplementary File S2). The method identifies a small number of

363   ambiguous offsets due to the low read coverage at the tails of the distributions. A finding that emphasizes

364   further the importance of read coverage as a critical factor in accurately identifying the A-site.

A-site offsets in mouse embryonic stem cells

366   The biological fact that A-site of a ribosome resides only between the second and stop codon is not

367   limited to *S. cerevisiae* and hence the Integer Programming algorithm should be applicable to Ribo-Seq

368   data from any organism. Therefore, we applied our method to a Pooled Ribo-Seq dataset of mouse

369   embryonic stem cells (mESCs). The resulting A-site offset table exhibited ambiguous offsets at all but

370   three (S, F) combinations (Supplementary Table S7). In mESCs there is widespread translation

371   elongation that occurs beyond the boundaries of annotated CDS regions in upstream open reading

372   frames (uORFs)[39]. Enrichment of ribosome-protected fragments from these translating uORFs can make

373   it difficult for our algorithm to find unique offsets because they can contribute reads around the start

374   codon of canonical annotated CDSs. Therefore, we hypothesized that if we apply our algorithm to only

375   those transcripts devoid of uORFs and possessing a single initiation site then our algorithm should

376   identify more unique offsets. Ingolia and co-workers[11] have experimentally identified for well-translated

377   mESCs transcripts its number of initiation sites and whether uORFs are present using translation-initiation

378   inhibiting drug Harringtonine. Therefore, we selected those genes that have only one translation initiation

379   site near the annotated start codon and further restricted our analysis to transcripts with a single isoform,

380   as multiple isoforms can have different termination sites.

381   Application of Integer Programming algorithm to this set of genes increases the number of unique

382   offsets from 3 to 13 (S, F) combinations (Supplementary Table S8). Applying the same robustness and

383   consistency tests as we did in *S. cerevisiae* reveals that 77% of the unique offsets are robust to threshold

11

384   variation, and a similar percentage is consistent across both individual datasets used to create the Pooled

385   data (Supplementary Table S8). Thus, the unique offsets we report for mESCs are robust and consistent

386   in the vast majority of datasets. This result also indicates that successful identification of A-site locations

387   requires analysing only those transcripts that do not contain uORFs.

Integer Programming does not yield unique offsets for *E.coli*

389   As a further test of how widely we can apply our algorithm, we applied it to a Pooled Ribo-Seq data from

390   the prokaryotic organism *E. coli*. The number of genes meeting our filtering criteria is reported in

391   Supplementary Table S3. MNase, the nuclease used in the *E. coli* Ribo-Seq protocol, digests mRNA in a

392   biased manner - favoring digestion from the 5′ end over the 3′ end[33,43]. Therefore, as done in other

393   studies[33,43,44], we applied our algorithm such that we identified the A-site location as the offset from the 3′

394   end instead of the 5′ end. Polycistronic mRNAs (*i.e.*, transcripts containing multiple CDSs) can cause

395   problems for our algorithm due to closely spaced reads at boundaries of contiguous CDS being scored for

396   different offsets in both the CDSs. To avoid inaccurate results, we restrict our analysis to the 1,915

397   monocistronic transcripts that do not have any other transcript within 40 nt upstream or downstream of

398   the CDS. Based on our experience in the analysis of mESCs dataset, we filter out transcripts with multiple

399   translation initiation sites as well as transcripts whose annotated initiation sites have been disputed.

400   Nakahigashi and co-workers[45] have used tetracycline as translation inhibitor to identify 92 transcripts in

401   *E.coli* with different initiation sites from the reference annotation and we exclude these transcripts from

402   our analysis. However, for this high coverage pooled dataset, we find ambiguous offsets for all (S, F)

403   combinations (Supplementary Table S7). A meta-gene analysis of normalized ribosome density in the

404   CDS and 30 nt region upstream and downstream reveal signatures of translation beyond the boundaries

405   of the CDS (Supplementary Fig. S4), especially a higher than average enrichment of reads a few

406   nucleotides before the start codon. We speculate that the base-pairing of the Shine-Dalgarno (SD)

407   sequence with the complementary anti-SD sequence in 16S rRNA[46] protects these few nucleotides

408   before the start codon from ribonuclease digestion and hence results in an enrichment of Ribo-Seq reads.

409   Since these "pseudo" ribosome-protected fragments cannot be differentiated from actual ribosome-

410   protected fragments containing a codon with the ribosome's A-site on it, our algorithm is limited in its

411   application for this data.

Reproducing known PPX and XPP motifs that lead to translational slowdown

413   In *S. cerevisiae*[47] and *E. coli*[33,48] certain PPX and XPP polypeptide motifs (in which X corresponds any

414   one of the 20 amino acids) can stall ribosomes when the third residue is in the A-site. Elongation factors

415   eIF5A (in *S. cerevisiae*) and EF-P (in *E. coli*) help relieve the stalling induced by some motifs but not

416   others[47]. Even in mESCs, Ingolia and co-workers[11] detected PPD and PPE as strong pausing motifs.

417   Therefore, we examined whether our approach can reproduce the known stalling motifs. We did this by

418   calculating the normalized read density at the different occurrences of a PPX and XPP motif.

12

419   In *S. cerevisiae*, we observed large ribosome densities at PPG, PPD, PPE and PPN (Fig. 5A), all of

420   which were classified as strong stallers in *S. cerevisiae*[47] and also in *E. coli*[48]. In contrast, there is no

421   stalling, on average, at PPP, consistent with other studies[47]. This is most likely due to the action of eIF5A.

422   For the XPP motifs, the strongest stalling was observed for GPP and DPP motifs, which are consistent

423   with the results in *S. cerevisiae* and in *E. coli* (Fig. 5B). In mESCs, we see the strongest stalling at PPE

424   and PPD, reproducing the results of Ingolia and co-workers[11] (Supplementary Fig. S5A). For XPP motifs,

425   we observed very weak stalling only for DPP (Supplementary Fig. S5B). Thus, our approach to map the

426   A-site on ribosome footprints enables the accurate detection of established translation pausing at

427   particular PPX and XPP nascent polypeptide motifs.

428   A study of Ribo-Seq data of mammalian cells[49] observed a sequence-independent translation pause

429   when the 5th codon of the transcript is in the P-site. This post-initiation pausing was also observed in an *in*

430   *vitro* study of poly-phenylalanine synthesis where stalling was observed when the 4th codon was in the P-

431   site[50]. With the A-site profiles obtained using our offset tables for *S. cerevisiae* and mESCs; we also

432   observe these pausing events when both the 4th and 5th codons are at the P-site (Supplementary Fig. S6).

Greater A-site location accuracy than other methods

434   There is no independent experimental method to verify the accuracy of identified A-site locations using

435   our method or any other method[4,5,52–55,6,8–10,12,41,42,51]. We argue that the well-established ribosome

436   pausing at particular PPX sequence motifs is the best available means to differentiate the accuracy of

437   existing methods. The reason for this is that these stalling motifs have been identified in *E.coli*[56,57] and *S.*

438   *cerevisiae*[58] through orthogonal experimental methods (including enzymology studies and toe printing),

439   and the exact location of the A-site during such a slowdown is known to be at the codon encoding the

440   third residue of the motif[56]. Thus, the most accurate A-site identification method will be the one that most

441   frequently assigns greater ribosome density to X at each occurrence of the PPX motif.

442   We applied this test to the strongest stalling PPX motifs, *i.e.*, PPG in *S. cerevisiae* and PPE in mESCs.

443   In *S. cerevisiae*, the Integer Programming method yields the greatest ribosome density at the glycine

444   codon of PPG motif when applied to both the Pooled (Fig. 6A) and Pop datasets (Supplementary Fig.

445   S7A). Examining each occurrence of PPG in the transcriptome, we find that in a majority of instances our

446   method assigns more ribosome density to glycine than every other method when applied to both the

447   Pooled (Fig. 6B, Wilcoxon signed-rank test $(n = 224)$, $P < 0.0005$ for all methods except Hussmann

448    $(P = 0.164)$) and Pop datasets (Supplementary Fig. S7B, Wilcoxon signed-rank test $(n = 35)$, $P <$

449   10^{-5} for all methods except Hussmann $(P = 0.026)$ and Ribodeblur $(P = 0.01)$). The same analyses

450   applied to mESCs at PPE motifs shows that our method outperforms the other nine methods (Figs. 6C-D)

451   with our method assigning greater ribosome density at glutamic acid for at least 85% of the PPE motifs in

452   our dataset as compared to all other methods (Fig. 6D, Wilcoxon signed-rank test $(n = 104)$, $P < 10^{-15}$

453   for all methods). Thus, for *S. cerevisiae* and mESCs our Integer Programming approach is more accurate

454   than other methods in identifying the A-site on ribosome-protected fragments.

13

455   A large number of molecular factors influence codon translation rates and ribosome density along

456   transcripts[59]. One factor is the cognate tRNA concentration, as codons decoded by cognate tRNA with

457   higher concentrations should have on average lower ribosome densities[15,16,60]. Therefore, as an additional

458   qualitative test, we expect that the most accurate A-site method will yield the largest anti-correlation

459   between the ribosome density at a codon and its cognate tRNA concentration. This test is only qualitative

460   as the correlation between codon ribosome-density and cognate tRNA concentration may be affected by

461   other factors, including codon usage and reuse of recharged tRNAs in the vicinity of the ribosome

462   influence the relationship[61,62]. Using tRNA abundances previously estimated from RNA-Seq experiments

463   on *S. cerevisiae*[16], we find that our Integer Programming method yields the largest anti-correlation

464   compared to the eleven other methods considered (Supplementary Table S10), further supporting the

465   accuracy of our method. We were unable to run this test in mESCs as measurements of tRNA

466   concentration have not been reported in the literature.

Discussion

468   We have introduced a method to determine the A- and P-site locations on ribosome-protected mRNA

469   fragments, and shown that it is more accurate than other methods in correctly assigning ribosome density

470   to the glycine residue in PPG motifs and glutamic acid residue in PPE motifs, which are strong

471   translation-stalling sites in *S. cerevisiae* and mESCs, respectively. Our method is unique amongst existing

472   methods because it (*i*) uses a probabilistic approach to identify the A-site location through Integer

473   Programming optimization and (*ii*) has an objective function rooted in the biology of translation – meaning

474   that its optimization enforces the fact that the A-site location of most reads must have been between the

475   second and stop codons of the CDSs. To be sure, several methods use biological features to assign the

476   A-site (such as having more reads around the start and stop codons than in the UTR[2,11]). However, ours

477   is the only method that also utilizes feature (*i*), which is beneficial because the stochastic nature of mRNA

478   cleavage during the digestion-step of Ribo-Seq necessitates a probabilistic perspective. Our method is

479   not entirely probabilistic since we have to set thresholds and apply a secondary criterion to arrive at a

480   final offset value. These measures are unavoidable due to the variability in coverage between different

481   genes. However, we find that the results are robust to variation in thresholds and mostly consistent

482   across different Ribo-Seq datasets. Hence, the respective A-site offset tables provided for *S. cerevisiae*

483   (Table 1) and mouse embryonic stem cells (Supplementary Table S8) can be applied to any dataset from

484   these organisms.

485   Noteworthy about our test for accuracy is that it is based on results from orthogonal experimental

486   techniques. The stalling of translation at glycine in PPG motifs is well-documented[33,47,56–58] and in *S.*

487   *cerevisiae* the Integer Programming method assigns higher Ribo-Seq reads at the glycine codon at most

488   instances of PPG compared to other A-site methods. In mESCs PPE is the strongest stalling motif[11]. The

489   Integer Programming method outperforms other methods by assigning, on average, 1.76 times more

490   reads at the glutamic acid codon compared to other methods. These results indicate that the Integer

14

491   Programming method presented in this study is more accurate than existing methods. One reason for this

492   increase in accuracy, among many possible reasons, may be that most methods only use reads from

493   around the start codon, while our method uses reads from around both the start and stop codons.

494   A potential point of confusion may arise from the distributions shown in Fig. 3 in which there are two

495   highly probable offset values, raising the question of whether or not there are multiple A-site locations for

496   a given fragment size and frame. In almost all fragment length and frame combinations, there is one

497   unique most probable A-site location, but this ambiguity can arise from poor read coverage on a gene or

498   stochastic fluctuations in the extent of digestion on one side of an mRNA fragment compared to the other.

499   Consider fragment size 28 in frame 1. In the Pop data set (top, middle panel of Fig. 3A), approximately

500   half of the genes have $\Delta = 15$ nt, while the others have $\Delta = 18$ nt, meaning the A-site could be at either

501   location. When we increase the read coverage of the genes, however, we see that the vast majority of the

502   offsets shift to 15 nt (bottom, middle panel in Fig. 3B). Thus, the original A-site ambiguity was not due to

503   multiple, equally possible A-site locations, but rather the true A-site location was hard to detect without

504   better coverage. Consider another example. For $S = 27$ and $F = 1$ we observe in Fig. 3A that 8% of genes

505   have an optimal $\Delta = 0$, seemingly suggesting that the A-site is located at the 5ʹ-end on a subset of

506   fragments. Spot-checking the ribosome profiles of these genes, we find that these genes contain no

507   reads in the 27 nt region upstream of the second codon and 27 nt upstream of the stop codon (data not

508   shown). Thus, the values of $T(\Delta|i, S, F)$ for all Δ were equal and the optimal Δ was arbitrarily assigned a

509   value of 0. In the higher coverage Pooled dataset, however, there are only 2% of genes with optimal $\Delta = 0$

510   for $S = 27$ and $F = 1$. Hence, as we increase coverage, the proportion of genes with spurious offsets

511   decreases. Thus, offsets away from the most probable offset arise from sampling issues, not from

512   multiple A-site locations. This result is also seen in the analysis of the artificial Ribo-Seq data where our

513   algorithm correctly predicts the true offsets for a majority of (S, F) combinations while ambiguous offsets

514   occur only for those (S, F) combinations with the lowest read coverage.

515   We note that we set a threshold of 70% to determine a most-probable offset for each fragment size

516   and reading frame and demonstrated that the results are robust to variation with this threshold

517   (Supplementary Table S5). Therefore, the A-site assignments reported in Table 1 represent the most

518   likely location of the A-site relative to the 5ʹ end of mRNA fragments produced from Ribo-Seq

519   experiments on *S. cerevisiae*.

520   Some (S, F) combinations (such as $S = 32$ and $F = 0$, in Table 1) appear to be inherently ambiguous,

521   that is, increasing their coverage does not lead to a unique A-site assignment (Fig. 4D). We do not know

522   the reason for this result, but we speculate that these are situations where there are truly multiple equally

523   probable A-site locations. Another possibility is that the ribosome adopts different conformations in these

524   situations that result in different read lengths and offsets, leading to ambiguity[14]. The important point is

525   that the A-site cannot be accurately assigned in these situations. We therefore recommend that

526   researchers discard reads from these (S, F) combinations to minimize chances of erroneous A-site

15

527   assignments. We believe it will have negligible effect on the A-site profiles since these combinations

528   contribute only 2.9% of total reads in the Pooled dataset.

529   We have found that the Integer Programming algorithm is sensitive to reads arising from outside the

530   boundaries of annotated CDS regions from non-canonical sources like upstream ORFs (uORFs) or

531   Internal Ribosome Entry Sites (IRES). Specifically, applying our method to Ribo-Seq data from mESCs

532   yielded few unique offsets. It was only after removing genes that had multiple translation initiation sites,

533   some arising from uORFs, that the number of unique offsets increased more than four-fold. The reason

534   for this improvement was that by removing the uORFs, our method's assumption was met that the reads

535   within 40 nt of the start codon only arise from the annotated CDS. Our method was not able to identify

536   any unique offsets in *E. coli* Ribo-Seq data even after we controlled for multiple translation initiation sites.

537   We observed in *E. coli* a high enrichment of reads before the start codon after applying the conventional

538   12 nt offset from 3 end[33] (Supplementary Fig. S4) which we speculate may be due to protection of

539   mRNA segments involved in binding of the Shine-Dalgarno sequence to the ribosome[63] and could limit

540   the accuracy of our method.

541   The next best method to the Integer Programming method is the Hussmann approach[12]. Besides more

542   frequently assigning greater ribosome density to glycine in PPG motifs and exhibiting strong correlation

543   with cognate tRNA abundances, the Integer Programming method is also superior because it provides

544   greater statistical power and is based on biological features of translation rather than heuristic

545   assumptions. Specifically, Hussmann's method only uses reads that are 28, 29 and 30 nt in length,

546   whereas our method uses reads between 24 to 34 nt in length. This greater coverage results in greater

547   statistical power for our method. Hussmann's method uses a nearest-neighbour heuristic to determine

548   frame-specific offsets of +14, 15 or 16 for lengths 28 and 29 and offset of +15, 16 or 17 for length 30,

549   whereas our method is based on the feature that the A-site be located within the CDS. The reason

550   Hussman's method yields comparable results is that its offset table is highly similar to Table 1. If the

551   reading frame is maintained after applying the offset from the 5 end, then 8 out of 9 of Hussmann's

552   offsets are the same as in Table 1 with the 9th offset of $(29,1)$ being ambiguous in our method.

553   Our method preserves the original 3 nt periodicity found in the original 5-end aligned mRNA

554   fragments. Therefore, it is not designed for detecting frame-shifting, translation of upstream ORFs, or

555   novel short peptides. Nevertheless, correct assignment of reads to the A-site codon is essential in a

556   variety of other analyses, such as determining translation kinetics, and our method provides the most

557   accurate assignment of ribosome density compared to other methods (Fig. 6 and Supplementary Table

558   S10).

559   In summary, we have created a method for A-site identification that is more accurate than existing

560   methods in *S. cerevisiae* and mouse embryonic stem cells, utilizes a fundamental feature of translation to

561   identify the A-site, and have revealed how the A-site location changes based on the size of the mRNA

562   fragment and its frame. By increasing the accuracy and range of fragment sizes for which the A-site can

16

563   be identified, our approach can help future studies to measure translation elongation properties at the

564   length scale of individual codons.

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

17

References

1. Calkhoven, C. F., Müller, C. & Leutz, A. Translational control of gene expression and disease. *Trends Mol. Med.* **8,** 577–583 (2002).

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

3. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165,** 22–33 (2016).

4. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7,** 1534–1550 (2012).

5. Martens, A. T., Taylor, J. & Hilser, V. J. Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic Acids Res.* **43,** 3680 (2015).

6. Wang, H., McManus, J. & Kingsford, C. Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. *J. Comput. Biol.* **24,** 486–500 (2017).

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

8. Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R. & Barbry, P. RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research* **5,** 1309 (2016).

9. Lauria, F., Tebaldi, T., Bernabò, P., Groen, E. J. N., Gillingwater, T. H. & Viero, G. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Comput. Biol.* **14,** 1–20 (2018).

10. Fang, H., Huang, Y.-F., Radhakrishnan, A., Siepel, A., Lyon, G. J. & Schatz, M. C. Scikit-ribo Enables Accurate Estimation and Robust Modeling of Translation Dynamics at Codon Resolution. *Cell Syst.* **6,** 180–191.e4 (2018).

11. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

12. Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S. & Press, W. H. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet.* **11,** e1005732 (2015).

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

14. Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **2014,** 1–16 (2014).

15. Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. & Futcher, B. Measurement of average

18

626   decoding rates of the 61 sense codons in vivo. *Elife* **3,** e03735 (2014).

627 16.   Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B. & Bartel, D. P.

628   Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and

629   Regulation of Yeast Translation. *Cell Rep.* **14,** 1787–1799 (2016).

630 17.   Cooper, G. *Translation of mRNA. The Cell: A Molecular Approach*. (Sinauer Associates, 2000). at

631   <https://www.ncbi.nlm.nih.gov/books/NBK9839/>

632 18.   Sierksma, G. *Linear and Integer Programming Theory and Practice*. (CRC Press, 2001). at

633   <http://openlibrary.org/books/OL8124799M/Linear_Integer_Programming>

634 19.   Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S. & Koller, D. Causal

635   signals between codon bias , mRNA structure , and the efficiency of translation and elongation.

636   *Mol. Syst. Biol.* **10,** 770 (2014).

637 20.   Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–

638   359 (2012).

639 21.   Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. TopHat2: accurate

640   alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome*

641   *Biol.* **14,** R36 (2013).

642 22.   Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in ribosome

643   profiling experiments. *Nucleic Acids Res.* **42,** (2014).

644 23.   Guydosh, N. R. & Green, R. Dom34 rescues ribosomes in 3′ untranslated regions. *Cell* **156,** 950–

645   962 (2014).

646 24.   Jan, C. H., Williams, C. C. & Weissman, J. S. 'Principles of ER cotranslational translocation

647   revealed by proximity-specific ribosome profiling'. *Science* **346,** 748–751 (2014).

648 25.   Williams, C. C., Jan, C. H. & Weissman, J. S. Targeting and plasticity of mitochondrial proteins

649   revealed by proximity-specific ribosome profiling. *Science* **346,** 748–751 (2014).

650 26.   Nedialkova, D. D. & Leidel, S. A. Optimization of Codon Translation Rates via tRNA Modifications

651   Maintains Proteome Integrity. *Cell* **161,** 1606–1618 (2015).

652 27.   Young, D. J., Guydosh, N. R., Zhang, F., Hinnebusch, A. G. & Green, R. Rli1/ABCE1 Recycles

653   Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs In Vivo. *Cell* **162,** 872–884

654   (2015).

655 28.   Nissley, D. A., Sharma, A. K., Ahmed, N., Friedrich, U. A., Kramer, G., Bukau, B. & O'Brien, E. P.

656   Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to

657   co-translational folding. *Nat. Commun.* **7,** 10341 (2016).

658 29.   Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

659   *EMBnet.journal* **17,** 10 (2011).

660 30.   Hurt, J. a, Robertson, A. D. & Burge, C. B. Global analyses of UPF1 binding and function reveals

661   expanded scope of nonsense-mediated mRNA decay. *Genome Res.* **23,** 1636–1650 (2013).

662 31.   Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine–Dalgarno sequence drives translational

663   pausing and codon choice in bacteria. *Nature* **484,** 538–541 (2012).

664   32.   Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates

665   reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635 (2014).

666   33.   Woolstenhulme, C. J., Guydosh, N. R., Green, R. & Buskirk, A. R. High-Precision analysis of

667   translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep.* **11,** 13–21 (2015).

668   34.   Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81,**

669   2340–2361 (1977).

670   35.   Good, P. *Permutation, Parametric, and Bootstrap Tests of Hypothesis.* (Springer Series in

671   Statistics, 2005). doi:10.1007/978-0-387-98135-2

672   36.   Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role for

673   proline in stalling translation. *Genome Res.* **24,** 2011–2021 (2014).

674   37.   Qian, W., Yang, J. R., Pearson, N. M., Maclean, C. & Zhang, J. Balanced codon usage optimizes

675   eukaryotic translational efficiency. *PLoS Genet.* **8,** e1002603 (2012).

676   38.   Diament, A. & Tuller, T. Estimation of ribosome profiling performance and reproducibility at various

677   levels of resolution. *Biol. Direct* **11,** 24 (2016).

678   39.   Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E.,

679   Wills, M. R. & Weissman, J. S. Ribosome Profiling Reveals Pervasive Translation Outside of

680   Annotated Protein-Coding Genes. *Cell Rep.* **8,** 1365–1379 (2014).

681   40.   Reid, D. W. & Nicchitta, C. V. Primary role for endoplasmic reticulum-bound ribosomes in cellular

682   translation identified by ribosome profiling. *J. Biol. Chem.* **287,** 5518–5527 (2012).

683   41.   Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H. & Dieterich, C. Bayesian

684   prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* **45,** 2960–2972 (2016).

685   42.   Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. Selective ribosome profiling as a

686   tool for studying the interaction of chaperones and targeting factors with nascent polypeptide

687   chains and ribosomes. *Nat. Protoc.* **8,** 2212–39 (2013).

688   43.   O'Connor, P. B. F., Li, G. W., Weissman, J. S., Atkins, J. F. & Baranov, P. V. RRNA:mRNA pairing

689   alters the length and the symmetry of mRNA-protected fragments in ribosome profiling

690   experiments. *Bioinformatics* **29,** 1488–1491 (2013).

691   44.   Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational

692   Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.* **14,** 686–694 (2016).

693   45.   Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., Yoshikawa, H.,

694   Wanner, B. L., Ishihama, Y. & Mori, H. Comprehensive identification of translation start sites by

695   tetracycline-inhibited ribosome profiling. *DNA Res.* **23,** 193–201 (2016).

696   46.   Malys, N. Shine-Dalgarno sequence of bacteriophage T4: GAGG prevails in early genes. *Mol.*

697   *Biol. Rep.* **39,** 33–39 (2012).

698   47.   Schuller, A. P., Wu, C. C. C., Dever, T. E., Buskirk, A. R. & Green, R. eIF5A Functions Globally in

699   Translation Elongation and Termination. *Mol. Cell* **66,** 194–205.e5 (2017).

700 48. Peil, L., Starosta, A. L., Lassak, J., Atkinson, G. C., Virumae, K., Spitzer, M., Tenson, T., Jung, K.,
701   Remme, J. & Wilson, D. N. Distinct XPPX sequence motifs induce ribosome stalling, which is
702   rescued by the translation elongation factor EF-P. *Proc. Natl. Acad. Sci.* **110,** 15265–15270
703   (2013).

704 49. Han, Y., Gao, X., Liu, B., Wan, J., Zhang, X. & Qian, S. B. Ribosome profiling reveals sequence-
705   independent post-initiation pausing as a signature of translation. *Cell Res.* **24,** 842–851 (2014).

706 50. Haase, N., Holtkamp, W., Lipowsky, R., Rodnina, M. & Rudorf, S. Decomposition of time-
707   dependent fluorescence signals reveals codon-specific kinetics of protein synthesis. *Nucleic Acids*
708   *Res.* **46,** (2018).

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

711 52. Diament, A. & Tuller, T. Estimation of ribosome profiling performance and reproducibility at various
712   levels of resolution. *Biol. Direct* **11,** 24 (2016).

713 53. Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of
714   Ribosomal Velocity. *PLoS Biol.* **11,** e1001508 (2013).

715 54. Dana, A. & Tuller, T. Determinants of Translation Elongation Speed and Ribosomal Profiling
716   Biases in Mouse Embryonic Stem Cells. *PLoS Comput. Biol.* **8,** (2012).

717 55. Sabi, R. & Tuller, T. A comparative genomics study on the effect of individual amino acids on
718   ribosome stalling. *BMC Genomics* **16,** S5 (2015).

719 56. Doerfel, L. K., Wohlgemuth, I., Kothe, C., Peske, F., Urlaub, H. & Rodnina, M. V. EF-P Is Essential
720   for Rapid Synthesis of Proteins Containing Consecutive Proline Residues. *Science* **339,** 85–88
721   (2013).

722 57. Ude, S., Lassak, J., Starosta, A. L., Kraxenberger, T., Wilson, D. N. & Jung, K. Translation
723   elongation factor EF-P alleviates ribosome stalling at Polyproline Stretches. **339,** 82–86 (2013).

724 58. Gutierrez, E., Shin, B. S., Woolstenhulme, C. J., Kim, J. R., Saini, P., Buskirk, A. R. & Dever, T. E.
725   eif5A promotes translation of polyproline motifs. *Mol. Cell* **51,** 35–45 (2013).

726 59. Sharma, A. K. & O'Brien, E. P. Non-equilibrium coupling of protein structure and function to
727   translation–elongation kinetics. *Curr. Opin. Struct. Biol.* **49,** 94–103 (2018).

728 60. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids*
729   *Res.* **42,** 9171–9181 (2014).

730 61. Brackley, C. A., Romano, M. C. & Thiel, M. The dynamics of supply and demand in mRNA
731   translation. *PLoS Comput. Biol.* **7,** (2011).

732 62. Rudorf, S. & Lipowsky, R. Protein synthesis in E. coli: Dependence of codon-specific elongation
733   on tRNA concentration and codon usage. *PLoS One* **10,** 1–22 (2015).

734 63. Sonenberg, N. & Hinnebusch, A. G. Regulation of Translation Initiation in Eukaryotes:
735   Mechanisms and Biological Targets. *Cell* **136,** 731–745 (2009).

736

Acknowledgements

Author Contributions

PS, PC and EPO conceived the study. NA, PS and EPO designed the computational analyses. PC, MV, CMD contributed to design of the computational analyses. NA and PS analyzed the data. NA and EPO wrote the manuscript. All authors reviewed and commented on the manuscript.

Data Availability

All source code is made available on the GitHub repository https://github.com/nabeel1990/Asite_IP_method

Additional Information

Competing interests: The authors declare no competing interests.

Figure legends

Figure 1: The A-site location can be defined as an offset from the 5′ end of ribosome-protected fragments. (A) A schematic representation of a ribosome beginning translation (top drawing) and of the offset between the Ribo-Seq reads mapped with respect to the 5′ end of footprints and centered on the A-site (orange bar plots). The ribosome is shown protecting a 28 nt fragment with its 5′ end in reading frame 0. The start codon of a gene can only occupy the P-site and hence the A-site was determined to be at an offset of 15 nt from the 5′ end for fragment size 28 which is in frame 0 (highlighted in yellow) [2]. The P-site and A-site within the fragment are indicated. The reads are then shifted from the 5′ end to the A-site by the offset value. **(B)** The boundaries of the 28 nt ribosome-protected footprint are indicated by red bars. Stochastic nuclease digestion can result in different fragment sizes. Two most probable variants of a 29 nt footprint with the 5′ end in frame 1 (highlighted in yellow) are shown with dashed arrows which can result in offsets of 15 nt (top) and 18 nt (bottom), respectively. **(C-D)** mRNA fragment size distribution for *S. cerevisiae* Ribo-Seq dataset from Pop and co-workers **(C)** and the Pooled dataset **(D)**

Figure 2: An illustration of the application of the Integer Programming algorithm to a Ribosome profile. For a hypothetical transcript that is 60 nt in length the first panel shows the ribosome profile originating from reads assigned to the 5′ end of fragments of size 33 in frame 0. The start and the stop codon are indicated while the rest of the CDS region is colored light peach. The algorithm shifts this ribosome profile by 3 nt and calculates the objective function $T(\Delta \mid i, S, F)$. The extent of the shift is the offset Δ. Values of $T(\Delta \mid i, S, F)$ for $\Delta = 12, 15, 18, 21$ nts are indicated. In this example, the average number of reads per codon is 7.85. The difference between the top two offsets, 18 (T=222) and 15 (T=215), is less than the average. Hence, we check the secondary criteria (Methods). Offset 18 meets the criteria that the number of reads in the start codon is less than one-fifth of the average of reads in second, third and fourth codons and also that number of reads in the second codon is greater than reads in third codon. Hence, Δ=18 nt is the optimal offset for this transcript.

Figure 3: Distribution of offset values from the Integer Programming algorithm applied to transcripts from *S. cerevisiae*. The data plotted in **(A)** are from the Pop dataset, and **(B)** the Pooled dataset. The distributions are plotted as a function of the offset value and for fragment sizes of 27 to 33 nt, are shown, from left to right, for frames 0, 1 and 2. For a given fragment size and frame, the A-site location is at the most probable Δ value in the distribution, provided the offset occurs for more than 70% of the genes (dashed lines in panels). Error bars represent 95% Confidence intervals calculated using Bootstrapping. Sample sizes are reported in Supplementary Table S2.

Figure 4: Increasing coverage identifies A-site locations for S and F combinations that were initially ambiguous. Plotted is the percentage of transcripts with a particular Δ value for different S and F combinations from the Pooled dataset of *S. cerevisiae*. In each panel, multiple distributions are plotted

23

798   corresponding to transcripts with increasing coverage, indicated by the legend at the bottom. For

799   example, the distributions in blue and red arise from transcripts with, respectively, at least 1 or 2 reads

800   per codon on average. We observe the A-site location tends towards 15 nt for $S = 25,\ F = 0$ **(A)** and

801   towards 18 nt for $S = 27, F = 2$ **(B)**, and $S = 30, F = 1$ **(C)**. For $S = 32, F = 0$ **(D)**, there is no trend even at

802   higher coverage. Note that for $S = 27, F = 2$ (panel B), there are less than 10 genes with an average

803   greater

804   than 50 reads per codon and hence we do not include the data point beyond average greater than 45

805   reads per codon (see Methods). Error bars represent 95% Confidence intervals calculated using

806   Bootstrapping.

807   **Figure 5: Several PPX and XPP motifs lead to ribosomal stalling in *S. cerevisiae*.** The median

808   normalized ribosome density is obtained for all instances of **(A)** PPX and **(B)** XPP motifs in which X

809   corresponds to any one of the 20 naturally occurring amino acids. Using a permutation test, we determine

810   if the median ribosome density is statistically significant or occurs by random chance. Statistically

811   significant motifs are highlighted in dark red. This analysis was carried out on the Pop dataset for

812   transcripts in which at least 50% of codon positions have reads mapped to them. Error bars are 95%

813   Confidence Intervals for the median obtained using Bootstrapping.

814   **Figure 6: The Integer Programming algorithm correctly assigns greater ribosome density than**

815   **other methods to the Glycine in PPG motifs in *S. cerevisiae* and to Glutamic acid in PPE motifs in**

816   **mESCs. (A)** Normalized ribosome density obtained using the various methods used to identify the A-site

817   is shown for an instance of PPG motif in gene YLR374W with G at codon position 303 in the Pooled

818   dataset of *S. cerevisiae* (see Legend and Main Text for details about methods). **(B)** The fraction of PPG

819   instances ($n = 224$) at which the Integer Programming method yields greater ribosome density at glycine

820   compared to every other method. The color-coding is the same as shown in the legend in panel (A). Our

821   method does better if it assigns greater ribosome density in more than half the instances (horizontal line

822   in panel B). The Integer Programming method does better than all other methods ($P < 0.0005$) except for

823   Hussmann, which is not statistically different ($P = 0.164$). **(C)** Normalized ribosome density is shown for

824   an instance of PPE motif in gene uc007zma.1 with E at codon position 127 in the Pooled dataset of

825   mouse ESCs (see Legend and main text for details about methods). **(D)** The fraction of PPE instances at

826   which the Integer Programming method yields greater ribosome density at glutamatic acid compared to

827   every other method. The color-coding is same as shown in the legend of panel (C). The Integer

828   Programming method does better than all other methods ($P < 10^{-15}$) in accurately assigning ribosome

829   density to Glutamic Acid in PPE motifs ($n = 104$). For both analyses, two-sided *p*-values were calculated

830   using the Wilcoxon signed rank test. Error bars represent the 95% Confidence Interval about the median

831   calculated using Bootstrapping.

832

24

Tables

Table 1: A-site locations (nucleotide offsets from 5′ end) determined by applying the Integer Programming algorithm to the Pooled dataset in *S. cerevisiae* are shown as a function of fragment size and frame. The top two offset values are listed for those S and F combinations in which the A-site location could not be uniquely determined. For unique offsets, the most-probable offset value is listed.

Fragment Size	Frame 0	Frame 1	Frame 2
24	15	15/12	18/12
25	15	12/15	18
26	15/12	18/15	18/15
27	15	15	18
28	15	15	18
29	15	15/18	18
30	15	18	18
31	15	18	18
32	18/15	18	18
33	18	18	18
34	18	18	18/21

A — Ribosome; Protected fragment; Reading Frame; Genome (beginning of a gene); 5′ end aligned reads; Reads mapped on the A site

B — Ribosome; P site; A site; Reading Frame; Frame 1 offset values

C — Pop Dataset

D — Pooled Dataset

A $S = 25, F = 0$

B $S = 27, F = 2$

C $S = 30, F = 1$

D $S = 32, F = 0$

Genes with offset (%)

Offset Δ (nucleotides)

avg > 1   avg > 2   avg > 5   avg > 10   avg > 15   avg > 20   — — Threshold

avg > 25   avg > 30   avg > 35   avg > 40   avg > 45   avg > 50

A

Normalized Ribo Density (y-axis, A)

Legend:
- Median
- - - Average Ribosome Density

PPX motif with X in A-site

B

Normalized Ribo Density (y-axis, B)

XPP Motif with second P in A-site