Analytical Biochemistry 510 (2016) 1–10

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Optimization of high-throughput sequencing kinetics for determining enzymatic rate constants of thousands of RNA substrates

Courtney N. Niland [a], Eckhard Jankowsky [b], Michael E. Harris [a,] *

[a] Department of Biochemistry, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA
[b] Center for RNA Molecular Biology, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

ARTICLE INFO

Article history:
Received 29 April 2016
Accepted 3 June 2016
Available online 11 June 2016

Keywords:
High-throughput sequencing
Enzyme specificity
RNA processing

ABSTRACT

Quantification of the specificity of RNA binding proteins and RNA processing enzymes is essential to understanding their fundamental roles in biological processes. High-throughput sequencing kinetics (HTS-Kin) uses high-throughput sequencing and internal competition kinetics to simultaneously monitor the processing rate constants of thousands of substrates by RNA processing enzymes. This technique has provided unprecedented insight into the substrate specificity of the tRNA processing endonuclease ribonuclease P. Here, we investigated the accuracy and robustness of measurements associated with each step of the HTS-Kin procedure. We examine the effect of substrate concentration on the observed rate constant, determine the optimal kinetic parameters, and provide guidelines for reducing error in amplification of the substrate population. Importantly, we found that high-throughput sequencing and experimental reproducibility contribute to error, and these are the main sources of imprecision in the quantified results when otherwise optimized guidelines are followed.

Published by Elsevier Inc.

The ability of ribonucleases, ribonucleoproteins, and RNA processing enzymes to recognize multiple alternative substrates is essential to cellular gene expression. For example, the RNA substrates for key enzymes such as the ribosome, spliceosome, tRNA, and mRNA processing enzymes can vary greatly in sequence and/or structure [1–5]. Given the broad range of alternative substrates that are recognized by these enzymes, their specificity cannot be entirely captured by sequence motif analysis, homology modeling, or similar approaches that consider only genomically encoded or optimal substrates [6]. Moreover, it is well established that a biologically relevant investigation of enzyme specificity involves understanding how substrates compete for association [7–9]. In vitro structure–function experiments comparing the kinetics of individual RNA substrate variants provide a powerful way to test potential specificity determinants. However, this approach has limited throughput and, therefore, is not practical for achieving a comprehensive description of specificity.

A more complete understanding of the specificity of RNA binding proteins and RNA processing enzymes can be gained by analysis of the processing rate constant or equilibrium binding constant for all possible substrate variants [6]. Such data provide a means for identifying sequence and structure determinants of specificity and comprehensively analyzing how sequence variation affects the reaction mechanism [10]. This level of understanding is necessary for prediction of the distribution of enzyme binding sites in the transcriptome and designing RNAs and RNA binding proteins with novel specificities [11–13]. By analyzing the effect of all possible variations in substrate RNA sequence on rate constants or equilibrium constants, the effect of sequence variation at one position on the sequence preference elsewhere in the binding site is revealed [14]. Such coupling between the energetic contributions of nucleotides in the RNA substrate is expected due in part to the complex structure and folding of RNA. Quantitative analysis of the interdependence between the contributions of individual nucleotides to recognition by RNA binding proteins and RNA processing enzymes has the potential to reveal important elements of substrate structure as well as their intrinsic sequence specificity.

Recently, powerful new approaches have been developed aimed at comprehensively analyzing RNA sequence specificity, including SELEX (systematic evolution of ligands by exponential enrichment) [15], Bind-n-Seq [16], and HiTS-RAP (high-throughput sequencing–RNA affinity profiling) [17]. However, these techniques

Abbreviations: HTS-Kin, high-throughput sequencing kinetics; PCR, polymerase chain reaction; UV, ultraviolet; EDTA, ethylenediaminetetraacetic acid; CV, coefficient of variation.

* Corresponding author.
E-mail address: meh2@cwru.edu (M.E. Harris).

http://dx.doi.org/10.1016/j.ab.2016.06.004
0003-2697/Published by Elsevier Inc.

monitor only equilibrium processes or provide information on optimal substrates only and, therefore, do not analyze the full complement of substrate variants or require specialized instrumentation. We developed a new technique termed high-throughput sequencing kinetics (HTS-Kin) that overcomes these limitations, allowing quantitative measurement of the second-order rate constants of thousands of substrate variants in a single reaction using standard molecular biology methods and standard Illumina sequencing protocols. Initial application of HTS-Kin was used to comprehensively analyze the specificity of C5, the protein subunit of the transfer RNA processing ribonucleoprotein enzyme RNase P from *Escherichia coli*, for its corresponding binding site in the 5′ leader of precursor tRNA [14]. The affinity distribution of C5 was found to resemble those of highly specific nucleic acid binding proteins [14]. Unlike these specific proteins, however, C5 does not bind its physiological RNA targets with the highest affinity but rather binds them with affinities near the median of the distribution. Thus, the data not only delineated the rules governing substrate recognition by C5 but also revealed that apparently nonspecific and specific RNA-binding modes might not differ fundamentally but represent distinct parts of common affinity distributions.

HTS-Kin continues to provide important new insights into RNase P molecular recognition and is amenable to a broad range of applications. Therefore, it is necessary to consider sources of uncertainty, evaluate their contribution to error in determination of relative rate constants by this method, and propose strategies for minimizing or avoiding inaccuracies in interpretation of rate constants calculated from these data. In HTS-Kin, the relative rate constants for in vitro RNA processing reactions are determined by analyzing the change in the concentration of individual RNAs in the unreacted substrate population compared with a reference substrate using internal competition kinetics. The change in concentration of each substrate is calculated from the number of reads obtained by Illumina sequencing of the substrate population at select time points in the reaction relative to a reference substrate.

Thus, for the HTS-Kin technique, there are several factors requiring optimization in order to minimize error and that may limit accuracy. These factors include (i) accounting for the variation in initial substrate concentrations in randomized RNA populations, (ii) choosing the appropriate time scale for accurately capturing the range of rate constants in the population, (iii) selecting an appropriate reference substrate as an internal standard, (iv) preparing the cDNA library by reverse transcription and polymerase chain reaction (PCR) amplification, (v) the Illumina sequencing itself, and (vi) error due to experimental reproducibility. Here, we examine each of these factors individually with respect to its contribution to the variation in the observed affinity distributions measured by HTS-Kin. In general, for optimal HTS-Kin experiments, early reaction times should be used to minimize rate constant compression. Although substrate amplification must be maintained in the linear range, the error due to small differences in cycle number is negligible. In addition, quantification of rate constants for slow reacting substrates is subject to error from Illumina sequencing, yet a high degree of experimental reproducibility is achieved for most substrate sequence variants.

Materials and methods

Isolation and synthesis of RNase P subunits and pre-tRNA substrates

Expression and purification of *E. coli* C5 protein was done as described previously [18]. *E. coli* P RNA was synthesized using T7 RNA polymerase (NEB M0251S) in in vitro transcription reactions containing 5–10 μg of template cDNA. The synthesized RNA products were isolated using PAGE (polyacrylamide gel electrophoresis), identified and excised by ultraviolet (UV) shadowing, purified by phenol–chloroform extraction and ethanol precipitation, dissolved in 10 mM Tris–HCl (pH 8) and 1 mM ethylenediaminetetraacetic acid (EDTA, pH 8), and quantified by UV absorbance. The *E. coli* pre-tRNA^{Met82} gene was cloned into the pUC18 vector, and PCR was used to introduce all possible mutations at positions N(−1) to N(−6) in the 5′ leader by using mutant forward primers to produce cDNA used for in vitro transcription as described, above, with 20–25 μg of template. PCR conditions consisted of the following: 1 U of *Taq* DNA polymerase (Roche 04638964001), 1× supplied PCR buffer, 0.2 mM dNTP mix, 0.5 μM forward and reverse primers, and 18 nM template DNA were heated to 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 s, 55 °C for 45 s, and 72 °C for 1 min, and final extension at 72 °C for 5 min.

Multiple turnover HTS-Kin reactions

RNase P in vitro pre-tRNA processing reactions were performed in 50 mM Tris–HCl (pH 8), 100 mM NaCl, 0.005% Triton X-100, and 17.5 mM $MgCl_2$. The holoenzyme complex and pre-tRNA pool (spiked with a negligible amount of ^{32}P-labeled pre-tRNA) were treated separately by denaturation at 95 °C for 3 min, followed by renaturation in $MgCl_2$ at 37 °C for 10 min. Reactions were initiated using equal volumes of enzyme and substrate with final concentrations of 5 nM RNase P and 1 μM pre-tRNA. Aliquots of 160 μl were taken at desired reaction times and quenched in 33 mM EDTA on dry ice. Substrate and product were separated on a 10% denaturing polyacrylamide gel and exposed to a phosphorimager screen and X-ray film. Radioactivity was quantified using ImageQuant software, and fraction of reaction was calculated by taking the amount of product at each time point divided by the addition of substrate and product bands. After substrate isolation and purification, first-strand synthesis was performed using 5 μl of the equalized RNA and 1 μM reverse primer at 72 °C for 10 min, ice for 1 min before adding 100 U of SuperScript III reverse transcriptase, 0.75 mM dNTP mix, 2.5 mM DTT (dithiothreitol), and 1× supplied RT buffer. Incubation continued at 42 °C for 10 min, 50 °C for 40 min, 55 °C for 20 min, and finally 95 °C for 5 min. Samples were diluted 1:300, and 1 μl was used to amplify for high-throughput sequencing. PCR was performed as above with forward primers that bound to the 21-nt sequence at the 5′ end and contained a barcode for each time point and randomized dinucleotide sequence.

Results and discussion

Determination of relative rate constants for in vitro RNA processing reactions by HTS-Kin

Internal competition kinetics, which HTS-Kin uses to calculate relative rate constants from Illumina sequencing data, is based on the fact that variation in specificity is due to differences in the activation energies for k_{cat}/K_m of alternative substrates for the same enzyme (Fig. 1A). There are several advantages and potential disadvantages in using internal competition kinetics; therefore, it is important to consider these factors in the context of their application in HTS-Kin. The kinetics of such reactions containing multiple alternative substrates has been described previously [8,9,19], and the equations and derivations for internal competition were recently reviewed and developed for quantification of both precursor and product ratios by Anderson [20]. Briefly, as illustrated in Scheme 1, a single population of enzyme (E) can combine with multiple substrates (S_1, S_2, S_3, …, S_i).

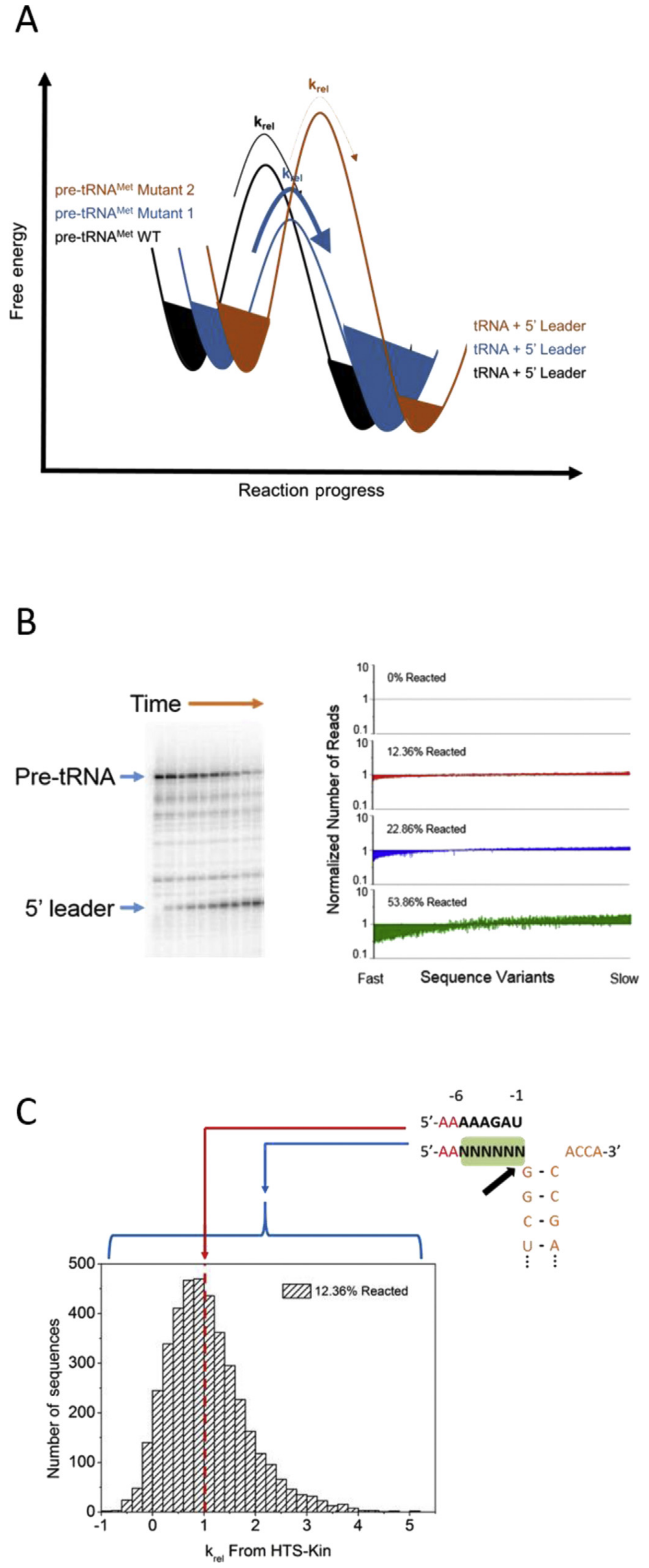

Fig.1. High-throughput sequencing kinetics (HTS-Kin) measures processing rates of thousands of RNA substrates using internal competition kinetics. (A) Reaction coordinate diagram depicting the processing of multiple pre-tRNA substrates by RNase P. As the reaction progresses, the activation energy for k_{cat}/K_m determines the relative rate of product formation; thus, favorable substrates (blue) are depleted more quickly, whereas unfavorable substrates (orange) are minimally processed and accumulate transiently relative to the wild-type substrate (black). (B) The substrate and product at different time points in the reaction are separated on a denaturing polyacrylamide gel (left), and the residual substrate population is isolated for high-throughput sequencing. Plotting the normalized reads for each substrate variant from Illumina sequencing shows that as the reaction progresses, substrates with fast k_{rel} values are depleted from the residual substrate population, whereas those with slow k_{rel} values accumulate (right). (C) An affinity distribution measured using HTS-Kin using a pre-tRNA^{Met}N(−1 to −6) randomized population is shown as the number of substrate variants with a given k_{rel} value and depicts the entire range of effects of this variation on enzyme processing. By definition, the wild-type pre-tRNA has a k_{rel} of 1, and substrates are calibrated to this as either faster (k_{rel} > 1) or slower (k_{rel} < 1) than the reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The rate of product formation of any individual substrate (v_{obs1}) is proportional to the fraction of total enzyme in the ES_1 form [21]. Additional alternative substrates deplete ES_1, and consequently the rate of formation of P_1, by acting as competitive inhibitors. For alternative substrates, here the substrate variant S_2 and wild-type reference S_1, the multiple turnover rate equation is essentially that for competitive inhibition and the ratio of the two observed rates simplifies to [8,9,22,23].

$$\frac{v_{obs2}}{v_{obs1}} = \frac{\left(k_{cat}/K_m\right)_2}{\left(k_{cat}/K_m\right)_1}\left(\frac{S_2}{S_1}\right). \tag{1}$$

Thus, the relative rate constant, or the ratio of the processing rate constants for the two competing substrates, is the ratio of their respective k_{cat}/K_m values multiplied by the ratio of their concentrations. Integration of the above general equation describes how the ratio of substrates will change over the time course for first-order and pseudo-first-order reactions [21,24]:

$$k_{rel} = \frac{\ln\left(\frac{S_2}{S_{2,0}}\right)}{\ln\left(\frac{S_1}{S_{1,0}}\right)}. \tag{2}$$

In Eq. (2), the values of $S_{1,0}$ and $S_{2,0}$ are the initial concentrations of the two substrates, and S_1 and S_2 are their concentrations after a specific time interval. This expression can be integrated and rearranged to give [14].

$$k_{rel} = \ln\frac{(1-f)}{\frac{R_{i,0}}{R_i}\left(\sum_1^i \frac{R}{R_0}X\right)} \Bigg/ \ln\frac{(1-f)}{\sum_1^i \frac{R}{R_0}X}, \tag{3}$$

where R_i is the ratio S_2/S_1 determined at remaining total substrate f and $R_{i,0}$ is the ratio S_2/S_1 at the start of the reaction. This expression is valid for any analytical method to measure S_2/S_1. In the case of HTS-Kin, these ratios are calculated from the number of Illumina sequence reads obtained from libraries made from the substrate population at the start of the reaction and at specific fractions of total substrate reacted.

HTS-Kin reactions involving RNase P require the following steps, all of which have specific features that can impact the reproducibility and contribute to the error in the calculation of relative rate constants. First, a population of pre-tRNA randomized in the 5′ leader at N(−6) to N(−1) that contacts both the RNA and protein subunits of RNase P is synthesized. Randomization is accomplished using the cloned wild-type pre-tRNA^{Met} gene as a template for PCR amplification in which the forward primers encode the randomized positions. The randomized DNA pool is then used for in vitro transcription to generate the randomized pre-tRNA substrate pool. Although the

$$E + S_1 \underset{}{\overset{K_1}{\rightleftarrows}} ES_1 \xrightarrow{V_1} E + P_1$$

$$E + S_2 \underset{}{\overset{K_2}{\rightleftarrows}} ES_2 \xrightarrow{V_2} E + P_2$$

$$E + S_3 \underset{}{\overset{K_3}{\rightleftarrows}} ES_1 \xrightarrow{V_3} E + P_3$$

$$E + S_i \underset{}{\overset{K_i}{\rightleftarrows}} ES_i \xrightarrow{V_i} E + P_i$$

Scheme 1. Association of RNase P (E) with multiple ptRNA substrates (S_1, S_2, S_3, S_i).

initial synthetic DNA population is synthesized to result in an approximate equimolar distribution of nucleotides at each position, it is unlikely that this distribution is maintained throughout the PCR and workup of the substrate pool. However, the initial distribution of substrate variants is assayed directly by Illumina sequencing. Moreover, as described in more detail below, the use of internal competition kinetics minimizes the effects of systematic inaccuracies in measurements of substrate ratios and does not rely on an equimolar distribution in the initial precursor population.

The substrate pool is reacted with RNase P, and substrate and product from individual reaction time points are separated on a denaturing polyacrylamide gel. The reaction progress is quantified and the substrate RNA populations are isolated at different time points and made into libraries for Illumina sequencing using reverse transcription and PCR amplification using a unique barcodes for each time point (Fig. 1B). By monitoring the number of Illumina reads of each sequence as a function of time, it is clear that as the reaction progresses, favorable substrates deplete from the residual substrate population while those with slow rate constants accumulate (Fig. 1B). Using Eq. (3) above, this information is used to calculate k_{rel} values for all 4096 substrate variants. These data represent the entire range of effects of this 5′ leader variation on enzyme processing. This is best exemplified in an affinity distribution as shown in Fig. 1C as a histogram of the number of substrate variants with a specific relative rate constant, k_{rel}.

The application of internal competition kinetics in this method offers several important advantages with respect to accuracy and precision of the resulting rate constant distribution. Due to the use of substrate ratios to calculate rate constants, systematic inaccuracies in the determination of these ratios, which may occur during several steps in the process, are canceled. In addition, experimental variation in k_{rel} calculation is minimized because all substrates react in the same reaction vessel and under identical reaction conditions. Nonetheless, disadvantages include the necessity to optimize several key reaction parameters and the potential for contributions from multiple sources of stochastic error that may propagate through the experiment. In the following sections, we consider the advantages and disadvantages at each step in the application of HTS-Kin with respect to reproducibility and minimization of error. First, we consider factors that may skew results or require optimization in the calculation of relative rate constants from substrate ratios. Then, we consider factors affecting the workup and measurement of the substrate ratios themselves by Illumina sequencing.

The magnitude of k_{rel} is independent of the distribution of substrate mole fractions in the initial precursor RNA population

One key factor apparent from inspection of Eq. (1) is that this expression is valid for any initial values of S_1 and S_2. Accordingly, the observed k_{rel} values measured by HTS-Kin should necessarily be independent of the individual concentrations of each individual substrate in the randomized pre-tRNA population. To test this in the application of HTS-Kin, we calculated the apparent mole fraction for each substrate variant in the initial substrate pool using its number of sequencing reads and dividing by the total number of reads for all substrate variants and then compared these values with the calculated k_{rel} for that substrate. As shown in Fig. 2A, a density plot of the observed k_{rel} plotted versus the mole fraction in the initial substrate population clearly shows that the two distributions are uncorrelated. In addition, a comparison of the ratio of high-throughput sequencing reads of mutant substrate to wild type in the starting material to the observed rate constant also reveals no correlation between these two parameters, as expected (Fig. 2B). In contrast, the change in the ratio of Illumina sequence reads for each substrate variant at a specific fraction of reaction relative to the ratio in the initial substrate population necessarily defines the magnitude of the observed k_{rel} calculated by Eq. (3). In Fig. 2C, the change in Illumina sequencing reads over the course of the reaction is plotted versus the magnitude of the calculated k_{rel} value to illustrate this fact. Thus, these results are consistent with principles of alternative substrate kinetics introduced above and described in more detail elsewhere [20].

Optimization of reaction kinetics and choice of internal reference for calculation of k_{rel}

Two additional aspects of the application of internal competition kinetics to calculate k_{rel} that are self-evident in Eq. (3) are the selections of the fraction of reaction (f) and the reference substrate (essentially S_1 from Eq. (1)). For the application of HTS-Kin to RNase P specificity, the genomically encoded leader sequence for the pre-tRNA$^{\mathrm{Met}}$ served as the reference substrate. For ease of interpretation, the use of a wild-type sequence as the reference has the obvious advantage that the absolute magnitude of k_{rel} reflects the fold difference in observed rate constant from a biologically relevant standard. Note, however, that the genomically encoded reference may or may not be the optimal substrate with respect to enzyme processing.

It follows that substrate variants with fast rate constants will exhibit a large change in substrate ratio relative to the reference (R_i) per unit time, whereas slower reacting species will result in only small changes in the observed ratios. A disadvantage is that precision of every k_{rel} measurement depends on the level of error in the measurements of the wild-type reference substrate (S_1 and $S_{1,0}$). The contribution of this error to the calculated k_{rel} value may limit precision of rate constant measurements that are significantly different from the reference. To address this potential limitation, 21 different reference substrate variants spanning a wide range from fast to slow processing by RNase P were each used to calculate the k_{rel} of all 4096 substrate variants. The k_{rel} determined for a single substrate variant using each of the 21 different reference substrates was averaged, and a standard deviation was calculated. References that produced a k_{rel} outside of the standard deviation for any substrate were eliminated. The remaining 15 reference substrates were used to calculate an average k_{rel} for each pre-tRNA, and a plot of this analysis is shown in Fig. 3A. The results clearly show a high correlation with k_{rel} determined using the wild-type reference. Error bars on the plot of k_{rel} values determined using multiple references provide an estimate of maximum uncertainty from using a single reference.

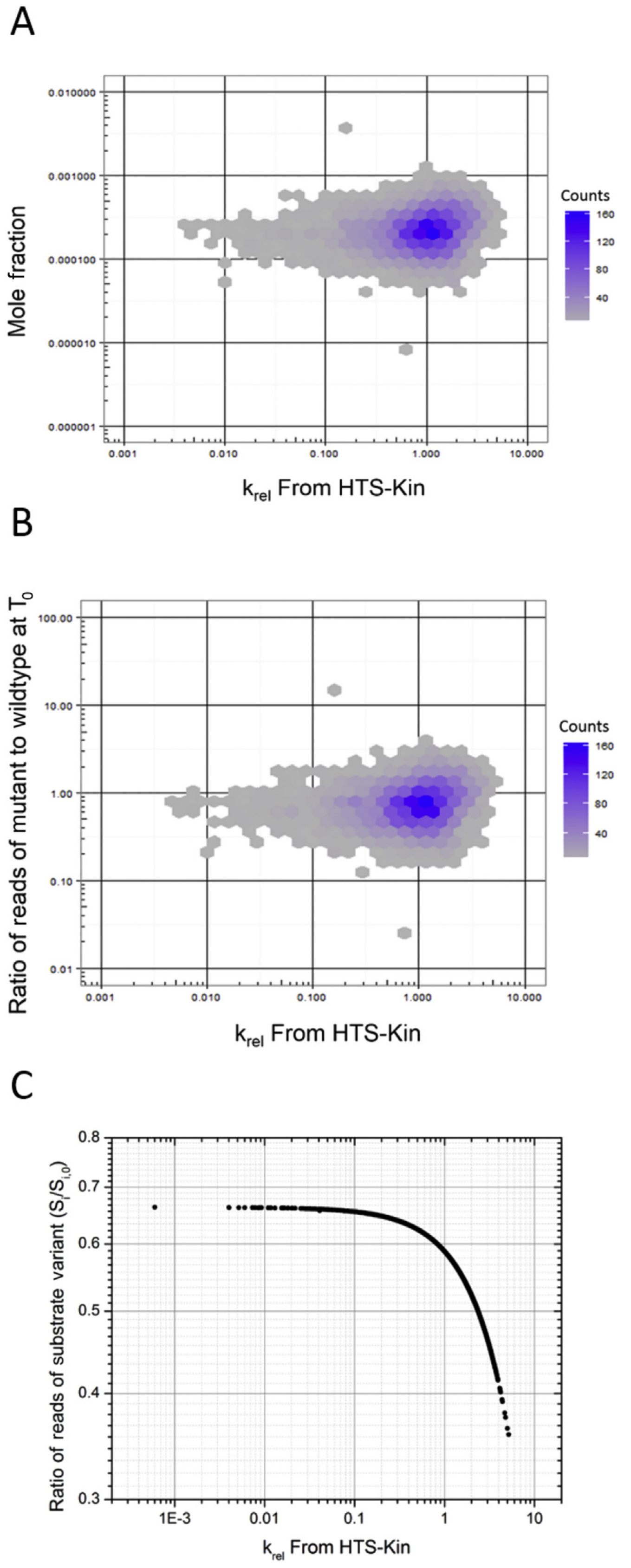

Fig.2. Analysis of the dependence of the observed k_{rel} on the distribution of substrate mole fractions in the initial precursor RNA population. From a single HTS-Kin reaction (Experiment 1) with pre-tRNA substrate variants randomized in the 5′ leader at N(−6 to −1), various aspects from Illumina sequencing are compared with the calculated k_{rel}. (A) Density plot of the mole fraction of each substrate variant (calculated as the ratio of number of reads of that substrate to that for all substrates at T_0) compared with its calculated k_{rel}, with the number of substrates in a particular area on the graph indicated by the shade of blue. (B) Density plot of the raw number of reads of each substrate variant in the starting material compared with the calculated k_{rel} of that substrate. (C) Ratio of raw reads of a substrate variant at a defined time in the reaction to that in the starting material (Illumina reads for S_n at 12% reacted/Illumina reads for S_n at 0% reacted) shows an exponential decrease with increasing k_{rel}. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As noted previously [25,26], a second factor that is integral in optimizing the range of effects that can be measured by HTS-Kin is the selection of appropriate time points for calculation of the expected range of k_{rel} values. This is apparent from Eq. (3), where the fraction of reaction is used to calculate each k_{rel}. The primary consequence of choice of inappropriate time points is illustrated in Fig. 3B, where affinity distributions calculated from samples taken at different time points in the same reaction are compared. The affinity distribution determined from an early time point provides the greatest range in k_{rel} values, whereas at later points it contains higher levels of substrate conversion exhibit compression in the range of observed k_{rel} values, as discussed previously [20,26]. The basis for this effect is illustrated in the inset of Fig. 3B with simulated kinetics for RNAs with different rate constants. At very early points in the reaction, only the fastest substrates will be processed, making calculation of k_{rel} for the vast majority of substrates highly error prone because their concentrations have changed little over this short time. Conversely, calculating k_{rel} from late points in the reaction provides a poor measure of processing rates because the fastest substrates are nearly consumed to completion, making the measurement of their k_{rel} inaccurate. At these later times in the reaction, the substrate ratios approach values reflect incomplete reactivity of the initial RNA population due to misfolding or other chemical differences. In addition, substrates with slower rate constants are afforded sufficient time to reach similar fractions of reaction to their faster counterparts. As a result, the observed k_{rel} values become artificially faster, as discussed previously [25,26].

As shown in Fig. 3C and D, we investigated the effect of varying the fraction of substrate reacted (f), from approximately 0.1 to 0.5, on the determined k_{rel} for each substrate variant. In this experiment, a single HTS-Kin reaction containing the same randomized pre-tRNA pool was sampled at several time points. The observed fraction of substrate reacted was determined for each time point, and affinity distributions were calculated. Fig. 3C and D shows the comparison of the k_{rel} distributions obtained for $f = 0.12$ versus the distributions obtained at $fs = 0.23$ and 0.54, respectively. A clear difference is observed in the range of k_{rel} values calculated using the substrate populations from later time points compared with $f = 0.12$. The range of k$_{\mathrm{rel}}$ values decreases dramatically from 1000-fold at $f = 0.12$–100-fold at 0.23 and just over 10-fold at 0.54. This compression in the calculated k_{rel} values is clearly shown in an overlay of the histograms representing the individual affinity distributions for the three experiments (Fig. 3A). The data further demonstrate that sampling at early time points at low substrate conversion provides the greatest accuracy. However, gains in the increase in signal to noise for slower reacting substrates achieved by sampling at later time points is more than offset by a large increase in systematic error affecting the entire affinity distribution.

Reliability of first-strand cDNA synthesis and quantitative PCR for Illumina library preparation

After the substrate RNA population is isolated from different times in the reaction, it must be converted to cDNA using reverse

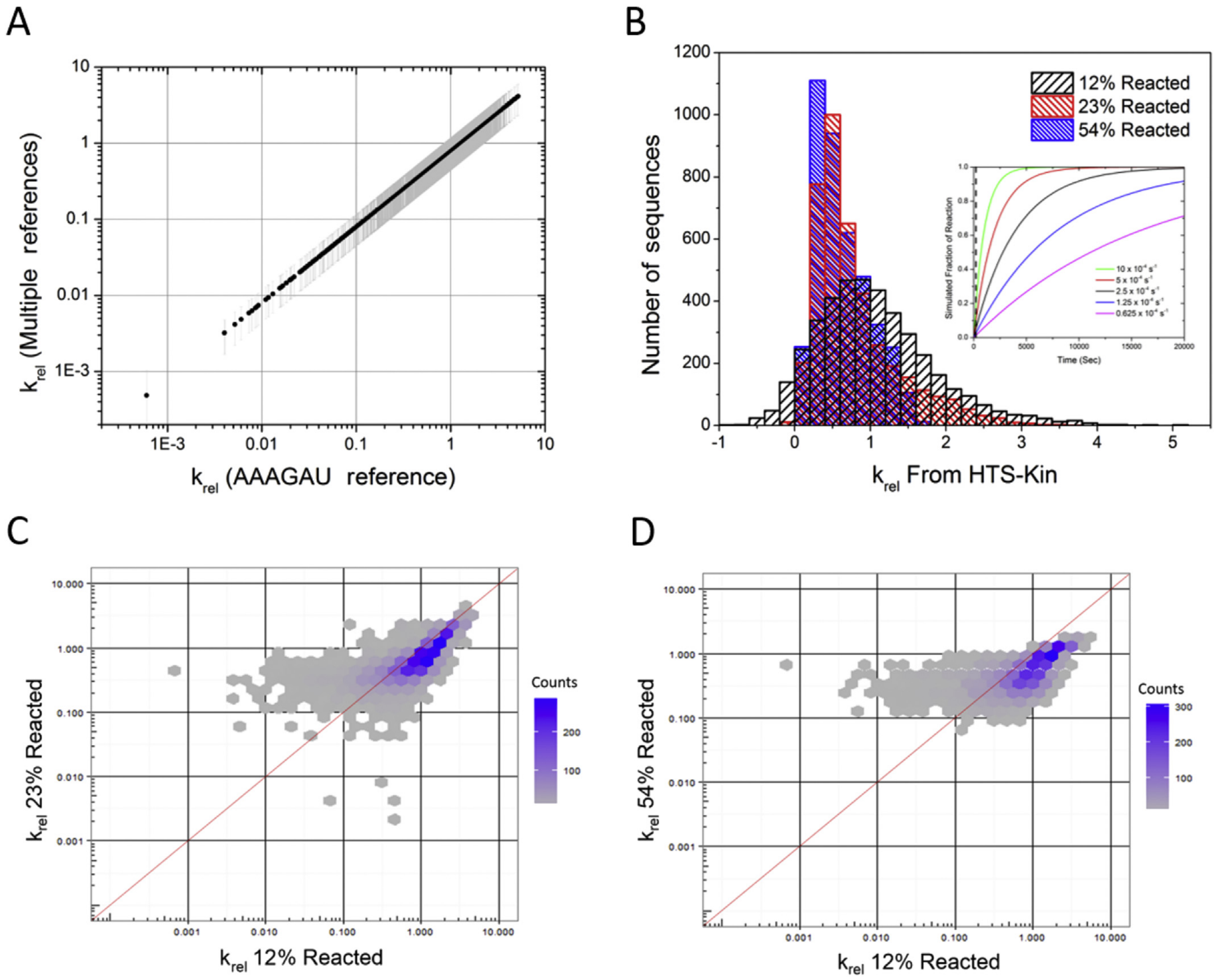

Fig.3. Optimization of reaction kinetics and choice of internal reference for calculation of k_{rel}. (A) Calculation of k_{rel} from HTS-Kin data from Experiment 1 using the wild-type 5′ leader (AAAGAU) as a reference compared with using 15 5′ leader variants spanning a range of k_{rel} as references in combination. The results of a single HTS-Kin reaction of pre-tRNA$^{\mathrm{Met}}$N(−6 to −1) with RNase P are investigated for their processing rate at increasing fractions of total substrate reacted. Inset: Simulation of the reaction progress of substrate variants with a range of rate constants. The simulated rate constants are indicated in the legend, and an identified optimal time for isolation and calculation of k_{rel} by HTS-Kin is indicated by the dashed black line. (B) Affinity distributions of the number of substrates with an indicated k_{rel} at various times in the same HTS-Kin reaction indicated in the legend. (C,D) Comparison of the k_{rel} determined from the same HTS-Kin reaction at different time points is shown as a density plot. Compression of rate constants is observed strongly at late times in the reaction.

transcription followed by PCR to generate the library for Illumina sequencing. During PCR, Illumina adapters are added to the cDNA corresponding to each RNA substrate as well as unique barcodes in order to distinguish reaction time points to allow for multiplexing. Previous analytical studies of RNA quantification using Illumina sequencing showed that the majority of error is the result of library preparation or poor choice of PCR primers [13,14]. The accuracy of k_{rel} in turn relies on the accuracy of measuring changes in the abundance of substrate variants over time. Therefore, it is essential to amplify the library under conditions where these differences are accurately preserved; thus, later amplification steps of HTS-Kin must be carefully considered and performed.

Several studies have aimed at achieving a quantitative understanding of various artifacts introduced by PCR that are relevant to HTS-Kin. For instance, template concentration, bias against high GC templates, template switching, and polymerase errors may contribute to errors in downstream steps [27–29]. These previous studies indicated that this bias and these errors can be minimized by using the minimum number of amplification cycles required to form products and defining the optimal template concentration in the PCR. Another consideration is the importance of testing for differential amplification of different barcoded primers because this can introduce amplification and subsequent sequencing bias for barcodes containing structure [30–32]. In our own experience, inaccurate results in one instance during preliminary experiments were traced to this effect. This consideration is tested by validation of all barcoded primers used for amplification by RT-PCR (reverse transcription PCR) or qPCR (quantitative PCR).

To diminish to the greatest extent possible the types of error during PCR amplification listed above, we determined the minimum number of PCR cycles necessary to achieve an identifiable cDNA product. Differences in the amount of pre-tRNA substrate remaining at different time points in the reaction were accounted for by normalizing the amount of template RNA used in the reverse

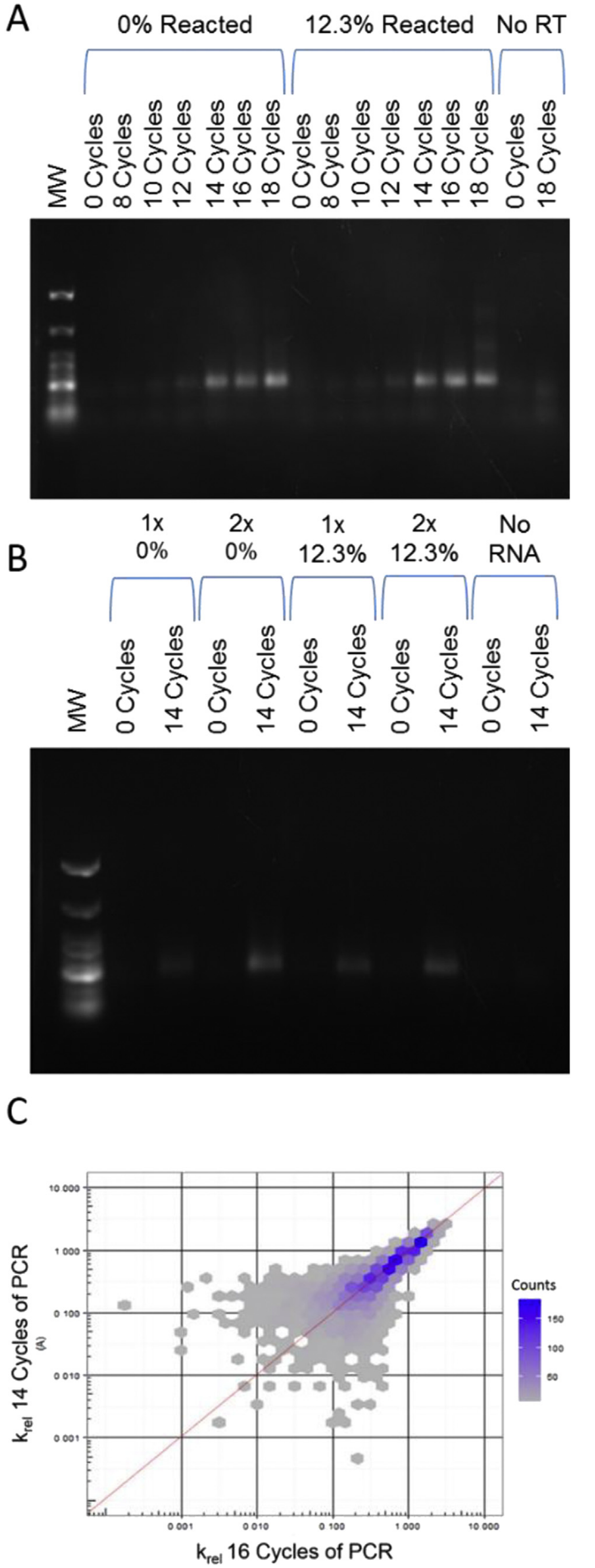

Fig.4. Analysis of the reliability of first-strand cDNA synthesis and quantitative PCR for Illumina library preparation. Substrate cDNA synthesis and amplification from RNase P HTS-Kin reactions from Experiment 2 with pre-tRNA^{Met}N(-6 to -1) are shown. (A) A 1% agarose gel showing the results of semi-quantitative PCR performed on the first-strand synthesis template from HTS-Kin reactions at different times in the reaction. The reaction time and number of PCR cycles are indicated at the top of the gel. RT, reverse transcription. (B) A 1% agarose gel showing linearity in the first-strand cDNA synthesis by reverse transcriptase. PCRs containing 1 or 2 times the amount of first-strand cDNA template from substrate populations at 0 and 12.36% reacted in HTS-Kin were performed for 0 and 14 cycles, and a control reverse transcription reaction was included in which no substrate RNA was added. (C) Comparison of the k_{rel} determined for the same HTS-Kin reaction in which the same first-strand cDNA from the substrate population was amplified for 14 or 16 cycles depicted as a density plot.

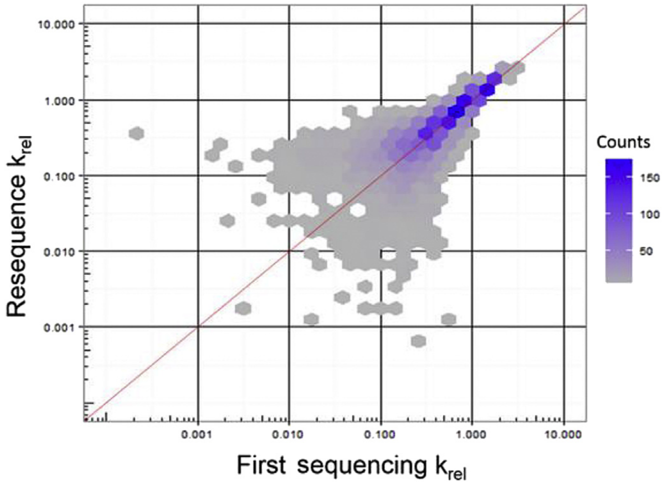

Fig.5. Illumina sequencing errors contribute to imprecision in measurement of low k_{rel} values. Resequencing was performed on the cDNA created from the second experimental replicate of HTS-Kin performed on RNase P processing of pre-tRNA^{Met}N(-6 to -1). The k_{rel} values determined from both Illumina sequencing runs of the same samples are plotted in a density plot where the number of points in a given area of the graph are indicated by the color and key at the right and show significant error in k_{rel} determination for slow reacting substrates. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this article.)

transcription reaction. We used semi-quantitative PCR to identify the linear range of amplification for each residual substrate population (Fig. 4A). To combat variations in PCR that would diminish the variation in the substrate population, we selected 14 cycles as the first number of PCR cycles for which a definable cDNA product band was observed.

In addition, inaccuracy in the construction of the Illumina sequencing library may arise if the amount of PCR products is not proportional to the concentration of input cDNA. To minimize this possibility, we ensure that the amount of DNA produced at the chosen number of PCR cycles is dependent on the amount of the first-strand cDNA product used as template. To demonstrate this, we performed PCRs for 14 cycles for reactions containing a 2-fold difference in the amount of first-strand cDNA synthesis products used as template. As shown in Fig. 4B, an approximately 2-fold increase in the amount of PCR product is detected by agarose gel electrophoresis in reactions containing a proportional increased cDNA template.

Nonetheless, it is possible that despite optimization there is nonlinear amplification of individual sequences even within the linear range for PCR amplification of the total population, which could be a potential source of error in the determination of k_{rel} values by HTS-Kin. To test this directly, we determined the observed k_{rel} from samples in which the same cDNA template was amplified for 14 versus 16 cycles of PCR, which are both in the apparent linear range of PCR amplification. In Fig. 4C, the k_{rel} values measured for all substrate variants in these two samples are compared. The k_{rel} values are highly correlative, in particular for the fastest reacting substrate variants. Significantly greater differences are observed in the k_{rel} values for slower reacting species.

Because the samples compared in Fig. 4C are from the same reaction, the observed error for the slow reacting species could be due to errors in downstream Illumina sequencing steps. As discussed above, the slowest reacting species will undergo the smallest change in concentration over the reaction; therefore, these data will exhibit the greatest sensitivity to stochastic measurement errors in the determination of these values. The high degree of correspondence for the vast majority of the population demonstrates the robustness of the method so long as attention is paid to whether linearity is maintained with respect to template concentration and PCR amplification.

Robustness of Illumina sequencing for reproducible determination of k_{rel} values

An unknown level of error may come from the variability between Illumina sequencing runs due to variation in flow cell, sample handling, or the instrument itself. Error from these sources can be minimized by pooling samples from different HTS-Kin reactions and different time points in the same Illumina flow cell lane using unique barcodes and combining these with other users' samples or a control sample. The reported error rate for Illumina HiSeq 2000 is 0.26%, the lowest reported for major high-throughput sequencing platforms [33]. Although we used data from the Illumina Hi-Seq 2500 in these studies, a similarly high level of fidelity is expected. Systematic miscalling of a particular nucleotide in the cDNA has been investigated and quantified, and there are various approaches to correcting these errors [34,35]. However, because of the large number of sequence reads (500–1500) obtained for most substrate variants, it is not necessary to apply them in HTS-Kin.

To estimate the error introduced in the Illumina sequencing step of the procedure, we compared the rate constants calculated from two sequencing runs on the same cDNA sample. Fig. 5 shows a plot of the two k_{rel} data sets obtained from the two separate sequencing runs. Inspection of the data shows that the substrate variants with the slowest k_{rel} have the greatest difference between measurements. Because the samples were not prepared separately for each run, we attribute this error directly to the variability of the high-throughput sequencing. Hence, Illumina sequencing appears to limit the ability to detect small changes in concentration of the slowest substrates over the short term in the RNase P reaction. Nonetheless, the data reveal highly robust reproducibility of the calculated k_{rel} values, demonstrating that for the majority of sequences the error introduced by Illumina sequencing is minimal.

Evaluation of experimental error

Optimally, analytical methods should provide data with sufficient precision such that the principal source of error is due to differences between experimental trials. We quantified the magnitude of experimental error between replicate HTS-Kin experiments. RNase P reactions were performed with the same pre-tRNA$^{\mathrm{Met}}$N(−6 to −1) population in triplicate and time points taken to achieve similar fractions of reaction, and the k_{rel} values were

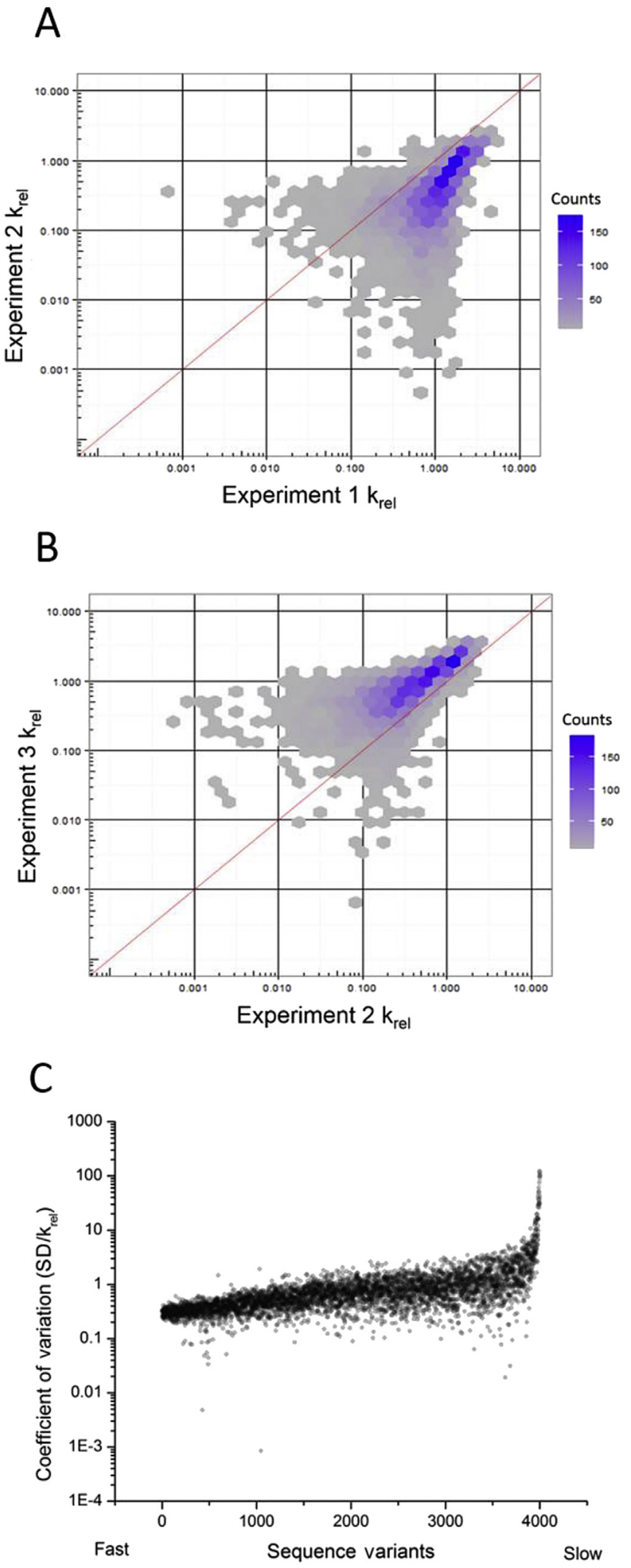

Fig.6. HTS-Kin replicates show reproducible determination of substrate variant k_{rel} except for slowest reacting substrates. (A,B) Comparison of the k_{rel} determined from replicate HTS-Kin experiments shown as a density plot, with the color indicating the number of points in that portion of the graph as shown by the legend. There is good correlation between replicates except for substrates processed with very slow k_{rel}. (C) The standard deviation (SD) in k_{rel} from three replicate HTS-Kin reactions of RNase P with pre-tRNA$^{\mathrm{Met}}$N(−6 to −1) was calculated, and the ratio of substrate k_{rel} to its SD was plotted. Substrates with high error are indicated by a ratio greater than 1. The coefficient of variation ratio (SD/k_{rel}) is compared with the observed processing rate, and substrate variants are aligned from fast to slow reacting. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this article.)

determined for each substrate variant using Eq. (3). The variation among the three individual experiments is visualized by plotting the resulting affinity distributions. In Fig. 6A, the affinity distributions for Experiments 1 and 2 are compared, and in Fig. 6B the data for Experiments 2 and 3 are compared. Both plots demonstrate strong correlation among the three data sets given that the majority of substrate variants are processed with very similar observed k_{rel} values between replicate experiments. Deviation from this trend is observed for substrates with very slow k_{rel} values that show the least correlation between replicates. As described above, this is due in large part to the relatively small changes in these substrates' concentration over the short time of the reaction that are in turn limited by error in the quantification of RNA levels by Illumina sequence reads.

The average k_{rel} and standard deviation calculated for each substrate variant was used to calculate the coefficient of variation (CV = standard deviation/average). The CV for each substrate variant was then plotted versus the magnitude of its average k_{rel} value. As shown in Fig. 6C, the substrate variants with the fastest k_{rel} values are measured with the greatest precision. As expected based on the plots shown in Fig. 6A and B, the CV for each substrate variant increases as k_{rel} decreases. The error increases sharply only for substrate variants with k_{rel} values that are 50- to 100-fold slower than the reference. However, the majority (75%) are measured with CV < 1, and the fastest 50% of sequence variants are measured with higher precision (CV < 2).

Conclusions

The analyses shown here provide strong support for the interpretation that the primary source of error for most k_{rel} values determined by HTS-Kin arises due to experiment-to-experiment variation. Importantly, the reproducibility between experiments for the majority of substrates shows a CV less than or equal to 1 for k_{rel} values spanning two orders of magnitude. For systems with a greater range of rate constants, the reproducibility is expected to be even better. However, for the slowest reacting substrate variants, an additional source of error becomes significant. The application of internal competition kinetics requires the measurement of the change in the ratio of the abundance of a particular RNA at the start of the reaction and at a specific time point. For slow reacting sequences, this change in RNA concentrations is small and falls below the range that can be reproducibly measured by Illumina sequencing. This effect is not significantly amplified by experimental error, but it limits accurate measurement at the lowest k_{rel} values. In sum, a carefully performed HTS-Kin experiment will include benchmarks using the procedures outlined here. Namely, the fraction of reaction should be carefully chosen to provide the greatest range in rate constants and an appropriate reference substrate identified that lies near the center of the rate constant distribution. Any analytical method used to quantify the change in substrate or product ratios can be applied; however, the error within these measurements necessarily impacts the measured k_{rel}; therefore, its precision must be investigated. In addition, the method of preparation of the RNA substrates for high-throughput sequencing, be it amplification to cDNA by PCR or ligation, introduces its own bias that can be handled to an extent by appropriate determination of the linear range of this amplification. The main error for substrates with slow k_{rel} comes from errors in precision of Illumina sequencing, and this should be investigated for other forms of quantification. HTS-Kin provides reproducible determinations of k_{rel} values for RNase P processing reactions, and these principles are likely to hold for many analogous in vitro RNA processing reactions.

Acknowledgments

This work was supported by U.S. National Institutes of Health (NIH) grants GM056740 and GM096000 to M.E.H., GM118088 to E.J., and T32 GM008056 to C.N.N. We thank members of the Harris and Jankowsky labs for thoughtful discussions on the HTS-Kin procedure and D. Anderson and N. Molyneaux for computational support in analyzing these data.

References

[1] R. Parker, H. Song, The enzymes and control of eukaryotic mRNA turnover, Nat. Struct. Mol. Biol. 11 (2004) 121–127.
[2] I. Wohlgemuth, C. Pohl, J. Mittelstaet, A.L. Konevega, M.V. Rodnina, Evolutionary optimization of speed and accuracy of decoding on the ribosome, Philos. Trans. R. Soc. Lond. B Biol. Sci. 366 (2011) 2979–2986.
[3] H.S. Zaher, R. Green, Fidelity at the molecular level: lessons from protein synthesis, Cell 136 (2009) 746–762.
[4] D. Wichtowska, T.W. Turowski, M. Boguta, An interplay between transcription, processing, and degradation determines tRNA levels in yeast, Wiley Interdiscip. Rev. RNA 4 (2013) 709–722.
[5] S. Lin, R.I. Gregory, MicroRNA biogenesis pathways in cancer, Nat. Rev. Cancer 15 (2015) 321–333.
[6] E. Jankowsky, M.E. Harris, Specificity and nonspecificity in RNA–protein interactions, Nat. Rev. Mol. Cell Biol. 16 (2015) 533–544.
[7] A. Cornish-Bowden, Enzyme specificity: its meaning in the general case, J. Theor. Biol. 108 (1984) 451–457.
[8] A. Fersht, Structure and Mechanism in Protein Science, W. H. Freeman, San Francisco, 1998.
[9] S. Cha, Kinetics of enzyme reactions with competing alternative substrates, Mol. Pharmacol. 4 (1968) 621–629.
[10] J.D. Buenrostro, C.L. Araya, L.M. Chircus, C.J. Layton, H.Y. Chang, M.P. Snyder, W.J. Greenleaf, Quantitative analysis of RNA–protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes, Nat. Biotechnol. 32 (2014) 562–568.
[11] M. McKeague, R.S. Wong, C.D. Smolke, Opportunities in the design and application of RNA for gene expression control, Nucleic Acids Res. 44 (2016) 2987–2999.
[12] Y. Chen, G. Varani, Engineering RNA-binding proteins for biology, FEBS J. 280 (2013) 3734–3754.
[13] R. Choudhury, Y.S. Tsai, D. Dominguez, Y. Wang, Z. Wang, Engineering RNA endonucleases with customized sequence specificities, Nat. Commun. 3 (2012) 1147.
[14] U.P. Guenther, L.E. Yandek, C.N. Niland, F.E. Campbell, D. Anderson, V.E. Anderson, M.E. Harris, E. Jankowsky, Hidden specificity in an apparently nonspecific RNA-binding protein, Nature 502 (2013) 385–388.
[15] C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science 249 (1990) 505–510.
[16] A. Zykovich, I. Korf, D.J. Segal, Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing, Nucleic Acids Res. 37 (22) (2009) e151.
[17] J.M. Tome, A. Ozer, J.M. Pagano, D. Gheba, G.P. Schroth, J.T. Lis, Comprehensive analysis of RNA–protein interactions by high-thoughput sequencing–RNA affinity profiling, Nat. Methods 11 (2014) 683–688.
[18] X. Guo, F.E. Campbell, L. Sun, E.L. Christian, V.E. Anderson, M.E. Harris, RNA-dependent folding and stabilization of C5 protein during assembly of the *E. coli* RNase P holoenzyme, J. Mol. Biol. 360 (2006) 190–203.
[19] L.E. Yandek, H.C. Lin, M.E. Harris, Alternative substrate kinetics of *Escherichia coli* ribonuclease P: determination of relative rate constants by internal competition, J. Biol. Chem. 288 (2013) 8342–8354.
[20] V.E. Anderson, Multiple alternative substrate kinetics, Biochim. Biophys. Acta 1854 (2015) 1729–1736.
[21] W.W. Cleland, P.F. Cook, Enzyme Kinetics and Mechanism, Garland, New York, 2007.
[22] W.W. Cleland, The use of isotope effects to determine enzyme mechanisms, Arch. Biochem. Biophys. 433 (2005) 2–12.
[23] D. Herschlag, The role of induced fit and conformational changes of enzymes in specificity and catalysis, Bioorg. Chem. 16 (1988) 62–96.
[24] W.W. Cleland, Enzyme mechanisms from isotope effects, in: A. Kohnen (Ed.), Isotope Effects in Chemistry and Biology, CRC/Taylor & Francis, Boca Raton, FL, 2006.
[25] D.L. Kellerman, K.S. Simmons, M. Pedraza, J.A. Piccirilli, D.M. York, M.E. Harris, Determination of hepatitis delta virus ribozyme N(−1) nucleobase and functional group specificity using internal competition kinetics, Anal. Biochem. 483 (2015) 12–20.
[26] H.-C. Lin, L.E. Yandek, I. Gjermeni, M.E. Harris, Determination of relative rate constants for in vitro RNA processing reactions by internal competition, Anal. Biochem. 467 (2014) 54–61.
[27] J.M. Kebschull, A.M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets, Nucleic Acids Res. 43 (21) (2015) e143.

[28] K. Kennedy, M.W. Hall, M.D. Lynch, G. Moreno-Hagelsieb, J.D. Neufeld, Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles, Appl. Environ. Microbiol. 80 (2014) 5717–5722.

[29] C. Brandariz-Fontes, M. Camacho-Sanchez, C. Vila, J.L. Vega-Pla, C. Rico, J.A. Leonard, Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results, Sci. Rep. 5 (2015) 8056.

[30] M. Pawluczyk, J. Weiss, M.G. Links, M. Egana Aranguren, M.D. Wilkinson, M. Egea-Cortines, Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples, Anal. Bioanal. Chem. 407 (2015) 1841–1848.

[31] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, Y. Wang, Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes, BMC Genom. 16 (2015) 589.

[32] R. D'Amore, U.Z. Ijaz, M. Schirmer, J.G. Kenny, R. Gregory, A.C. Darby, M. Shakya, M. Podar, C. Quince, N. Hall, A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling, BMC Genom. 17 (2016) 55.

[33] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences, and Illumina MiSeq sequencers, BMC Genom. 13 (2012) 341.

[34] X. Yang, S. Aluru, K.S. Dorman, Repeat-aware modeling and correction of short read errors, BMC Bioinform. 12 (Suppl. 1) (2011) S52.

[35] M. Schirmer, R. D'Amore, U.Z. Ijaz, N. Hall, C. Quince, Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data, BMC Bioinform. 17 (2016) 125.