Published online 27 March 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 12* ***e67***
https://doi.org/10.1093/nar/gkab199

Kinetic sequencing (*k*-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters

Yuning Shen[1], Abe Pressman[2], Evan Janzen[1,3] and Irene A. Chen[1,3,4,*]

[1]Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA 93106, USA, [2]Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA, [3]Program in Biomolecular Sciences and Engineering, University of California, Santa Barbara, CA 93106, USA and [4]Department of Chemical and Biomolecular Engineering, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA

Received December 10, 2020; Revised February 15, 2021; Editorial Decision March 08, 2021; Accepted March 13, 2021

ABSTRACT

Characterizing genotype-phenotype relationships of biomolecules (e.g. ribozymes) requires accurate ways to measure activity for a large set of molecules. Kinetic measurement using high-throughput sequencing (e.g. *k*-Seq) is an emerging assay applicable in various domains that potentially scales up measurement throughput to over 10^6 unique nucleic acid sequences. However, maximizing the return of such assays requires understanding the technical challenges introduced by sequence heterogeneity and DNA sequencing. We characterized the *k*-Seq method in terms of model identifiability, effects of sequencing error, accuracy and precision using simulated datasets and experimental data from a variant pool constructed from previously identified ribozymes. Relative abundance, kinetic coefficients, and measurement noise were found to affect the measurement of each sequence. We introduced bootstrapping to robustly quantify the uncertainty in estimating model parameters and proposed interpretable metrics to quantify model identifiability. These efforts enabled the rigorous reporting of data quality for individual sequences in *k*-Seq experiments. Here we present detailed protocols, define critical experimental factors, and identify general guidelines to maximize the number of sequences and their measurement accuracy from *k*-Seq data. Analogous practices could be applied to improve the rigor of other sequencing-based assays.

INTRODUCTION

Determining the genotype-phenotype relationships for any large set of biomolecules requires a high-throughput method. For catalytic nucleic acids, this requires measuring the activity of each unique sequence in a diverse population. Ideally, methods to accomplish this would: (a) yield accurate activity measurements for individual sequences, (b) achieve high throughput to cover a large number of variants in sequence space and (c) be adaptable to different ribozymes (and deoxyribozymes).

High-throughput sequencing (HTS) provides the ability to address these goals. The large amount of sequencing data (~10^8 reads) can allow high accuracy count data for many sequences in a high throughput, parallelized format. If reacted and unreacted molecules can be separated from each other and sequenced independently, HTS can quantify the extent of a reaction for millions of different sequences simultaneously. Since nucleic acid sequences act as their own 'barcodes', using sequencing as the assay avoids the need to isolate and test each unique sequence individually. Such a method can measure the activity of each sequence in a population of functional molecules, at multiple time points, substrate concentrations, or other variable conditions. HTS-based kinetic measurements have been proposed and demonstrated with nucleic acids, including catalytic DNA (1), catalytic RNA (2–5), substrate RNA ('HTS-Kin') (6), RNA aptamers (7) and transcription factor (TF) binding DNA (8). In these studies, approximately 10^3~10^6 unique sequences are measured, depending on the experimental design. Similar approaches have also been developed for proteins, notably an assay of ligand binding affinities through mRNA display (9), 'deep mutational scanning', in which the phenotype of fitness is assayed for many mutants by deep sequencing (10), and a large-scale measurement of dose-response curves (11). However,

[*]To whom correspondence should be addressed. Tel: +1 310 825 046; Fax: +1 310 206 4107; Email: ireneachen@ucla.edu
Present address: Abe Pressman, Cellular Engineering Group, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA.

the development of these massively parallel measurements also raises important questions about experimental design, normalization, sequencing errors, and measurement accuracy and precision. To date, there is a relative lack of critical study of the theoretical and experimental effects of such variables on the outcome of high-throughput measurements. The present work examines these issues for the case of ribozymes and develops appropriate methodology to address these concerns for high-throughput measurement of genotype-phenotype relationships.

Here, we focus on kinetic sequencing (*k*-Seq), a recently reported method for quantification of kinetics in a mixed pool of sequences (2). Advantages of *k*-Seq include absolute, rather than relative (6) measurements, as well as the lack of requirement for specialized instrumentation (4,7). A general schema of *k*-Seq is described as follows (Figure 1): an input pool is designed containing sequences of interest (e.g. candidate ribozymes). Aliquots of the input pool are reacted under different experimental conditions, such as different substrate concentrations or different time points. Then, reacted and unreacted molecules are separated. Each pool is converted to a DNA library and prepared for sequencing. Absolute measurement of reacted (or unreacted) quantities is also performed to allow normalization. Reads generated from HTS are subjected to quality control and de-replicated to generate a 'count' table of the copy number for each sequence detected in a sample. Count data are normalized to absolute abundance and fit to the appropriate kinetic model to estimate the rate constants and other parameters of interest.

Multiple issues could potentially limit the applicability of *k*-Seq and related methods. Kinetic measurements require properly chosen experimental conditions (e.g. substrate concentrations or time points) for a sufficient dynamic range. For a heterogeneous pool where different sequences have different optimal conditions, the conditions chosen will represent a compromise for some sequences. For example, in a single time-point (reacted versus unreacted) experiment determining enzyme kinetics over 4096 RNA substrates (6), the choice of reaction time would be optimal for either highly active RNA substrates or less active ones, but not all. Previous work determining kinetics for ribozymes by *k*-Seq also showed that characterization of less active ribozymes was limited when the kinetic model parameters (rate constant *k* and maximum amplitude of reaction *A*) could not be independently estimated due to model identifiability problems (2). While the number of experimental conditions could be increased to address this problem, with HTS this potential solution would quickly become prohibitively expensive in time and resources. Therefore, it is important to rigorously understand how the choice of conditions would affect the estimation of kinetic parameters and trustworthiness of measurements on each sequence.

Another consideration unique to HTS-based kinetic measurements derives from the inevitability of sequencing errors in the data. Sequencing error might misidentify a molecule as a nearby sequence variant, and subsequently change the quantification of both the true and incorrect sequences. This could be particularly problematic when one sequence is present in high abundance relative to others, and thus creates a relatively large number of misidentified reads that confound the quantitation of related sequences. It is important to note that this problem cannot be solved by increasing the number of replicates or sequencing depth, since the number of erroneous reads rises in proportion to the number of reads (a systematic bias rather than random noise).

A final concern is assessing the accuracy and precision of *k*-Seq measurements. Using discrete count data (number of reads of a particular sequence) as the approximation of a sequence's relative abundance in the sample introduces some complexity in assessing measurement accuracy, particularly at low counts where large stochastic variation exists. One approach taken in earlier works is to limit the library size to thousands of sequences which are each present in high copy number (high coverage of sequences) (1,6). However, this requirement might unnecessarily restrict the applicability of *k*-Seq and related methods. When extending the approach to larger libraries or libraries with uneven coverage (e.g. a 'doped' variant pool or pool obtained from *in vitro* selection), high error may be associated with measurements of sequences with low counts (3,4,9). While one may simply exclude sequences with counts lower than a cutoff value, it is not obvious how to choose such cutoffs. Instead, it would be desirable to estimate the uncertainty (e.g. confidence intervals) on fitted parameter values for each sequence, given the experimental scenario of a low number of replicates. This approach would maximize the return of accurate information from a *k*-Seq experiment.

In this work, we performed a *k*-Seq experiment on a newly designed pool of variants based on ribozymes previously isolated from *in vitro* selection (2). These ribozymes react with an activated amino acid substrate (biotinyl-Tyr(Me)-oxazolone, or BYO) to produce aminoacyl-RNA, and are characterized by pseudo-first order kinetics. Coupled with theoretical and simulation studies, we systematically characterize model identifiability, accuracy, and precision of estimation from HTS data. We discuss key factors to consider when optimizing experimental design for *k*-Seq experiments. Lastly, we present a Python package of analysis tools for users undertaking *k*-Seq experiments. Incorporation of these techniques and lessons into HTS-based kinetic measurements should improve rigorous quantitative inference from similar experiments.

MATERIALS AND METHODS

Pseudo-first order kinetic model

The fraction of BYO reacted with ribozymes during the experiments is small compared to the initial concentration of BYO and to BYO hydrolysis in aqueous solution (2). With this approximation, we modeled the kinetics of the reaction using a pseudo-first order rate equation. Fixing the reaction time and varying the initial BYO concentration was experimentally expedient, as BYO degradation could be included as a constant coefficient (α):

$$f_{ij} = A_i \left(1 - e^{-\alpha t k_i c_j}\right) \qquad (1)$$

where f_{ij} is the reacted fraction for sequence i in sample j with initial BYO concentration c_j. The kinetics for sequence

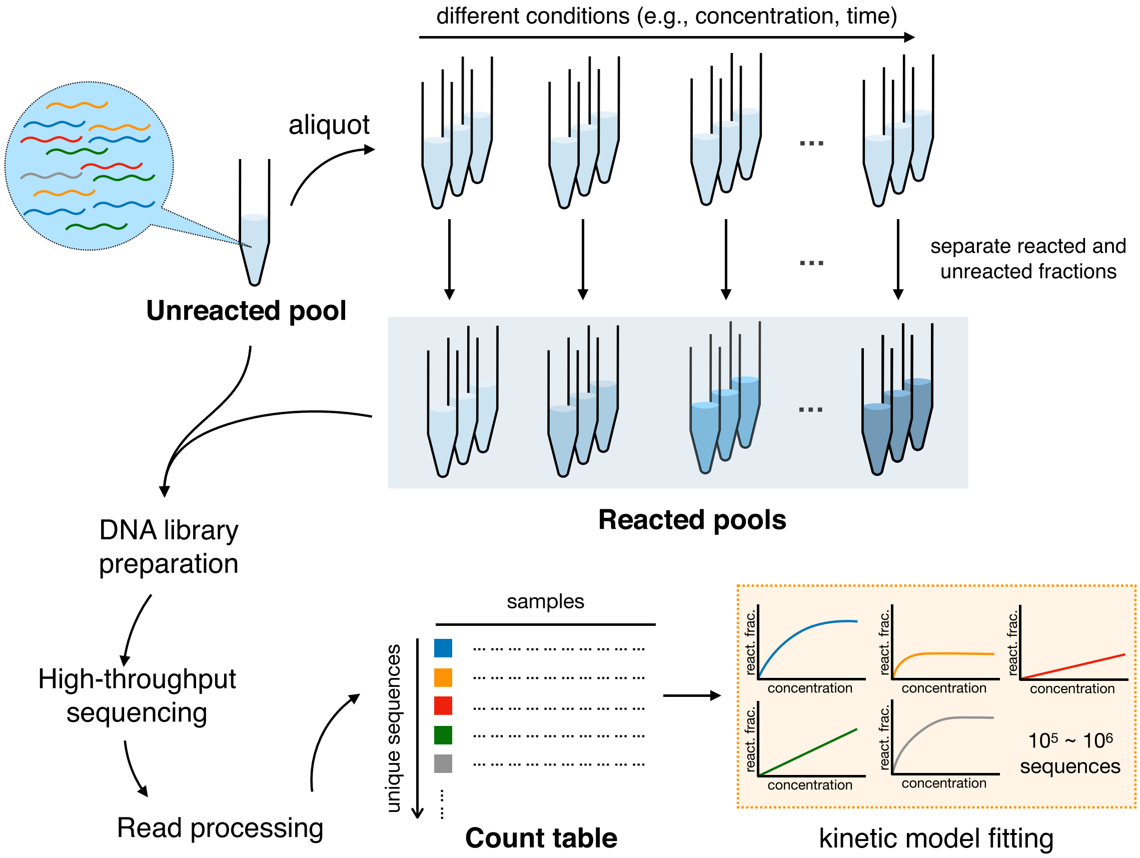

Figure 1. General scheme of k-Seq experiment and analysis. A heterogeneous input pool containing nucleic acids is reacted at different experimental conditions (e.g. different substrate concentrations or different reaction time). Reacted and unreacted molecules are separated and either (or both) of these fractions is prepared for high-throughput sequencing. The reads from DNA sequencing are processed to obtain a count table for each unique sequence across samples, normalized by a standard, and abundances across samples are fit into a kinetic model to estimate parameters (e.g. rate constants). react. frac. = reacted fraction.

i are characterized by A_i (the maximum amplitude of reaction) and k_i (the rate constant). A constant reaction time $t = 90$ min was used, corresponding to the degradation coefficient $\alpha = 0.479$, as measured in (2). The product $k_i A_i$ was used as a combined activity measure and represents the initial rate of the reaction.

Model identifiability for different k and A

Depending on the BYO concentrations, the pseudo-first order kinetic model may be characterized as practically unidentifiable (12) where k and A are sensitive to noise and cannot be separately estimated. We evaluated this effect for each sequence using bootstrapping (resampling data with replacement; see below for details). We designed two metrics to score the model identifiability from bootstrapping results: σ_A and $\gamma = \log_{10} \frac{\sigma_k \times \mu_A}{\sigma_{kA}}$, where σ_A, σ_k, σ_{kA} are the standard deviations for A, k and the product kA, and μ_A is the mean value of A, from bootstrapped samples for each sequence. The score γ is the ratio of standard deviations for the rate constant k over the product kA, scaled by estimated A. If k and A are well-estimated independently, the ratio would be close to 1 and γ would be close to 0; if k and A are not estimated independently such that k has larger variance than kA in estimation, γ would be larger than 0. While bootstrapping can capture the noise in the experimental measurements through resampling, we also separately examined the convergence of fitting through 20 independent runs with random initial values of k and A sampled from Uniform(0, 10) min^{-1}M^{-1} and Uniform(0, 1) respectively, using the original data (no resampling). The convergence of fitting was evaluated by the range of fitted A values ($\Delta A = A_{\max} - A_{\min}$) and used as a third candidate metric for model identifiability (higher $\Delta A \sim$ less identifiable).

To study model identifiability with controlled variables, we created a simulated reacted fraction dataset containing 10 201 (101^2) sequences with the $\log_{10}$ of true values of k and A on a 101-by-101 grid across the region with $\log_{10} k \in [-1, 3]$ and $\log_{10} A \in [-2, 0]$. The true reacted fraction f_{ij}^0 for sequence i in sample j with initial BYO concentration c_j was calculated based on the pseudo-first order rate equation (Eq. 1). To model the effect of measurement noise, we added a Gaussian error term err$_{ij}$ on the true reacted fraction f_{ij}^0 with the variance equal to ϵf_{ij}^0, where the ϵ is the relative error of the reacted fraction:

$$f_{ij} = f_{ij}^0 + \mathrm{err}_{ij} = A_i\left(1 - e^{-\alpha t k_i c_j}\right) + \mathrm{err}_{ij}$$
$$\mathrm{err}_{ij} \sim \mathrm{Normal}\left(0, \epsilon f_{ij}^0\right) \quad (2)$$

where f_{ij} is the observed reacted fraction. We chose $\epsilon =$ 0, 0.2, 0.5, 1.0 to evaluate the effect of measurement error. Negative values of f_{ij} were reassigned to be zero. These simulated reacted fractions were then used to estimate k_i and A_i for each simulated sequence using least-squares fitting as described below.

To study whether experimental and sequencing effort would be best spent extending the substrate concentration range vs. performing additional replicates, we simulated reacted fraction data for three different sets of BYO concentrations: (i) standard set: 2, 10, 50, and 250 μM with triplicates, for 12 samples in total, as done previously to analyze a pool after *in vitro* selection (2); (ii) additional replicates: 2, 10, 50 and 250 μM with four replicates each, for 16 samples in total and (iii) extended substrate range: 2, 10, 50, 250 and 1250 μM with triplicates, for 15 samples in total, as done in the variant pool experiment reported here.

Sequence variant pool for aminoacylation assay

Four DNA libraries were obtained from Keck Biotechnology Laboratory, with the sequence 5′-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-[central variable region, length 21]-TTCACTGCAGACTTGACGAAGCTG-3′ (nucleotides upstream of the transcription start site are underlined). For each library, the central region corresponded to a 21-nucleotide variable sequence, based on a ribozyme family wild-type sequence (S-1A.1-a, S-1B.1-a, S-2.1-a or S-3.1-a previously identified in (2), see Supplementary Table S1 for sequences), with partial randomization at each position (specified to be 91% of the wild-type nucleotide and 3% of each base substitution at all 21 positions, i.e. a 'doped' pool). 60-90 μg RNA was transcribed from 500 ng template DNA using HiScribe T7 RNA polymerase (New England Biolabs) and purified by denaturing polyacrylamide gel electrophoresis (PAGE) with 0.5× TBE buffer, as previously described (2). An equimolar mixture of these four RNA libraries (the variant pool) was prepared for the *k*-Seq experiment.

k-Seq experiment on the mixed pool of variants

Reactions were carried out in triplicates at 2, 10, 50, 250 and 1250 μM BYO for 90 min, following the incubation, RNA recovery and reverse-transcription protocols used in (2). Briefly, in each 50 μL *k*-Seq reaction, 2 μg total RNA (1.7 μM) was reacted with BYO in the aminoacylation buffer (100 mM HEPES, pH 7.0, 100 mM NaCl, 100 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$) for 90 min. The reactions were stopped using Bio-Spin P-30 Tris desalting columns (Bio-Rad) to remove unreacted substrates and placed on ice. Reacted sequences were isolated by pull-down with Streptavidin MagneSphere paramagnetic beads (Promega) at a volume ratio of 1:1 and eluted with 95% formamide/10 mM EDTA for 5 min under 65°C. 10% of eluted RNA was taken to measure the total RNA amount using Qubit and qPCR (see below). A 'spike-in' RNA was added as an alternative quantification method (see below). RNA was prepared for sequencing by reverse transcription and PCR (RT-PCR), with primers complementary to the fixed sequences flanking the variable region (GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC, forward;CAGCTTCGTCAAGTCTGCAGTGAA, reverse). DNA from each of 15 samples was barcoded and pooled together in equal proportions. A reverse-transcribed unreacted sample was added at three times the total amount of DNA of one reacted sample to have similar total sequencing depth with each set of BYO concentration triplicates. Pooled DNA was sequenced on an Illumina NextSeq 500 with 150 bp paired-end run (Biological Nanostructures Laboratory, California NanoSystems Institute at UCSB), using a high output reagent kit expected to produce > 400 million reads. To confirm that the concentration of BYO was in excess during the aminoacylation reaction, sequencing of the lowest BYO concentration reaction (2 μM) showed that the fraction of reacted RNA was $\sim 2\times 10^{-3}$. Given the initial RNA concentration of 1.7 μM, the amount of BYO consumed by aminoacylation was < 1% of the initial concentration.

Quantitation of total amount of RNA per sample

We used two methods to quantify the absolute amount of RNA in *k*-Seq samples. Method 1 measured the amount of RNA in the samples after elution using Qubit or qPCR. For reactions carried out at 250 and 1250 μM, 10% of the RNA recovered after elution was quantified with an Invitrogen Qubit 3.0 fluorometer. If the recovered RNA was below the limit of detection by Qubit, quantitation was done by reverse-transcription-qPCR (same PCR primers as in the previous section) using a Bio-Rad C1000 thermal cycler with CFX96 Real-Time PCR block, SuperScript III RTase (Invitrogen), Phusion® High-Fidelity Polymerase (New England Biolabs), and SYBR Green (Bio-Rad, 0.5× standard concentration) (Supplementary Figure S1). Method 2 used an internal standard (spike-in sequence) and data were normalized using sequencing results. A control spike-in sequence (5′-GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC-AAAAACAAAAACAAAAACAAA-TTCACTGCAGACTTGACGAAGCTG-3′, promoter underlined) was added to each sample before reverse transcription. 0.04, 0.2, 1, 2 and 2 ng of spike-in RNA was added to samples with 2, 10, 50, 250, 1250 μM BYO concentration respectively. 10 ng of spike-in RNA was added to the unreacted pool sample. The total RNA recovered (Q_j) in sample j was calculated as

$$Q_j = \frac{N_j - n_{sj}}{n_{sj}} \times q_{sj} \tag{3}$$

where N_j is the total number of reads in sample j, n_{sj} is the total reads of sequences within 2 edit distance (i.e. number of substitutions, insertions, or deletions) of the spike-in sequence in sample j, and q_{sj} is the quantity of spike-in sequence added to sample j after the reaction.

Processing of *k*-Seq reads

FASTQ files of de-multiplexed paired-end Illumina reads were processed using EasyDIVER (13) to count the number of reads of each unique sequence in each sample. The forward and reverse reads were joined using PANDAseq

(14) with the options '-a' to join the paired-end reads before trimming and 'completely_miss_the_point:0' to enforce absolute matching in the overlapped variable region (any pairs with a disagreement between forward and reverse reads were discarded), thus minimizing sequencing errors. After joining, forward and reverse primers were trimmed by PANDAseq using 'CTACGAATTC' (forward) and 'CTGCAGTGAA' (reverse) adapter sequences. Next, multiple lanes for the same sample were combined and reads were de-replicated to give unique sequences and counts. The generated count files were analyzed using the 'k-seq' python package (https://github.com/ichen-lab-ucsb/k-seq). We collected all detected sequences in unreacted and/or reacted samples and discarded those that were not 21 nucleotides long, with ambiguous nucleotides ('N'), or within an edit distance of 2 from the spike-in sequence. The absolute amount (ng) for sequence i in sample j were quantified using total RNA recovered Q_j and number of reads:

$$q_{ij} = \frac{n_{ij}}{N_j - n_{sj}} \times Q_j \tag{4}$$

The reacted fractions for sequences in reacted samples were further calculated as the ratio to the absolute amount in the unreacted pool (q_{0j}):

$$f_{ij} = \frac{q_{ij}}{q_{0j}} \tag{5}$$

To be considered analyzable for fitting, a sequence needed at least one non-zero value among reacted samples as well as a non-zero count in the unreacted sample; non-analyzable sequences were discarded.

Experimental coverage of mutants in the variant pool

Coverage of mutants in the variant RNA pool was analyzed from sequencing results of the unreacted pool. Unique sequences were classified into family centers (Hamming distance $d = 0$), single mutants ($d = 1$), double mutants ($d = 2$), triple mutants ($d = 3$), and others ($d \geq 4$) based on their Hamming distance (number of substitutions) to the nearest family center (S-2.1-a, S-1A.1-a, S-1B.1-a or S-3.1-a). The coverage fraction for a certain class of mutants was calculated by dividing the number of unique sequences detected by the number of possible sequences in each class ($4C(L, d)3^d$ where C is the combination operator to select d elements from a set of size L, where $L = 21$ is the length of the variable region, and the factor 4 is the number of families in the pool).

Point estimation of model parameters

Model parameters k and A for each sequence were estimated using least-squares fitting on reacted fractions with different initial BYO concentrations. Least-squares fittings were performed using the 'optimize.curve_fit' function of the 'SciPy' package in python with 'trust region reflective' (trf) method. The initial values of k and A were sampled from uniform distribution between 0 and 1. The bounds $[0, 1]$ were applied on A values and $[0, +\infty)$ on k values. The tolerances for optimization termination (ftol, xtol, gtol) were kept as default (10^{-8}). Optimal k, A determined from all sample points for a sequence were reported as point estimates.

Simulated count data based on the experimental pool

To best resemble the experimental pool and conditions, we simulate the k-Seq pool dataset from experimentally measured values. We sampled $M = 10^6$ sequences with parameters (p_{i0}, k_i, A_i) estimated in the k-Seq experiment on the mixed variant pool, where p_{i0} is the relative abundance for sequence i in the unreacted pool (normalized to 1 after sampling) and k_i, A_i are point estimates from fitting. We used the pseudo-first order rate equation (Eq. 1) to calculate the reacted fraction f_{ij} for each sequence with initial BYO concentrations from the extended substrate range and triplicate samples. The relative abundance in the simulated reacted samples was $p_{ij} = \frac{p_{i0} f_{ij}}{\sum_{r=1}^{M} p_{r0} f_{rj}}$. We then used the multinomial distribution Multinomial(N_j, p_{1j}, p_{2j}, ..., p_{Mj}) (where $N_j = 40M$, yielding a similar mean count per sequence to that observed in the experimental pool), to model the process of sequencing by sampling N_j reads for each sample (unreacted and reacted). To simulate total RNA recovered, we sampled from Normal(μ_j, $0.15\mu_j$), where $\mu_j = \sum_{i=1}^{M} p_{i0} f_{ij}$ is the total RNA amount reacted in the mixed pool reaction and 0.15 is the relative error. The value of the relative error was based on the relative standard deviation calculated from quantification using the spike-in or by direct RNA amount quantification (Supplementary Figure S2).

Uncertainty estimation using bootstrapping

The uncertainty of estimation was assessed using bootstrap sampling of the relative residuals. Let f_{ij} be the reacted fraction for sequence i in reacted sample j, and $\hat{f}_{ij}$ be the fitted value from point estimation. For each sequence, we calculated the relative residual as $r_{ij} = \frac{f_{ij} - \hat{f}_{ij}}{\hat{f}_{ij}}$. Each bootstrapping process resampled the relative residuals for sequence i (with replacement) to the same sample size ($\hat{r}_{i1}, \hat{r}_{i2}, \cdots, \hat{r}_{iJ}$), then applied the resampled relative residuals to $\hat{f}_{ij}$ with proper scaling (i.e. $(1 + \hat{r}_{ij})\hat{f}_{ij}$) as bootstrapped data points. Least-squares fitting was performed on each set of bootstrapped data points for which k, A, and kA values were recorded. Sample mean, standard deviation (s.d.), median, and estimated 95% confidence interval (CI-95, as mean ± 1.96 s.d. or [2.5-percentile, 97.5-percentile]) on k, A and kA were calculated from bootstrapped results for each sequence.

We performed least-squares fitting on data from the mixed pool of variants (reported in this paper), data from the previously published selection (2), the simulated reacted fraction dataset (reported in this paper), and the simulated pool dataset (reported in this paper). Bootstrapping was performed for 100 re-samples for each sequence for uncertainty estimation. To compare the performance of bootstrapping, we also applied the triplicates method, used previously (2), to the simulated pool dataset, with each replicate in a BYO concentration assigned to one of three series.

Each of the simulated triplicate series was fitted separately to calculate the standard deviation in estimating k, A and kA for each sequence.

‘k-seq’ python package for data analysis

The k-Seq analysis pipeline is open-source and given in a python package, ‘k-seq’, for future practitioners. The package includes pre-processing of count data from raw FASTQ files using EasyDIVER (13), filtering sequences and samples, quantification of individual sequences using total nucleic acid amount or the spike-in sequence, kinetic model fitting with bootstrapping, data simulation, as well as data analysis and visualizations. The package contains both pipeline scripts for standard data analysis and python modules to quickly build customized pipelines. Read preprocessing, sequence quantification, and auxiliary data simulation are included in the ‘data’ submodule. k-Seq fitting with bootstrapping is implemented in the ‘estimate’ module. The package in its current form uses pseudo-first order kinetics, but it can be used to fit other reaction kinetics by modifying the fitting equation (Supplementary Text S1). See GitHub (https://github.com/ichen-lab-ucsb/k-seq) for examples and usages.

RESULTS

Model identifiability depends on kinetic coefficients, experimental conditions and measurement error

We used the simulated reacted fraction dataset to evaluate the effect of kinetic coefficients, measurement error, and experimental conditions on model identifiability for pseudo-first order kinetics, specifically if k and A can be separately estimated. We first evaluated the model identifiability qualitatively, for sequences selected from 6 regions over the parameter space of $\log_{10}(k)$ from -1 to 3, $\log_{10}(A)$ from -2 to 0, and $kA > 0.1\,\text{min}^{-1}\text{M}^{-1}$ (Supplementary Figure S3-S9). In each region, sequences fitted in repeated fitting or bootstrapping were sampled for visual evaluation of the separability of k and A, under various measurement error rates ϵ. As summarized in Supplementary Table S2, sequences with higher k, A values and lower ϵ are more likely to be separable.

To quantify separability for individual sequences, we calculated three metrics: ΔA or the range of A across repeated fittings (no resampling); σ_A, or the standard deviation of A from bootstrapping samples; and γ, a measure of how noisy the separate estimation of k and A is compared to estimating the product kA. ΔA was able to identify sequences with numerically unstable fitting results which have small k, A values, but was not able to identify sequences whose fitting optima were sensitive to noise in the data. Almost all the sequences from the 6 selected regions each converged to a uniform optimum in repeated fitting, and the convergence was insensitive to the level of noise. In contrast, bootstrapping results account for noise in the data, and the optima from fitting subsampled data points did not always converge, providing more comprehensive separability information than convergence of multiple fittings. By comparing the distribution of each metric for sequences in the selected regions, both σ_A and γ reflected the trend of model identifiability observed by examining individual curves: higher metric value corresponded to less separable parameters of a sequence (Supplementary Figure S10). In practice, we found σ_A and γ aligned equally well with human intuitions (Supplementary Figure S11) and with each other (Supplementary Figure S12) in evaluating model identifiability from variant pool k-Seq data.

Using metric γ, we further assessed the effects of experimental conditions and measurement error on model identifiability (Figure 2). Model identifiability depends on the true k and A values, choice of substrate concentrations, and the level of measurement error. Controlling the experimental design and measurement error, k and A were more separable for sequences with higher k and A value. Comparing the sequences for a given kA, model identifiability appears to be more dependent on k than A, especially for cases with lower measurement error (e.g. relative error ≤ 0.5). To assess the effect of experimental conditions, we compared the case of adding one more replicate to each reaction (4 versus 3) to adding a higher concentration of BYO (1250 μM) in triplicate. Despite having more samples, adding another replicate did not change the region of sequences with identifiable models. However, adding a higher concentration of substrate shifted the boundary of separability on kA values by a factor of ~10, effectively increasing the dynamic range of the k-Seq assay (Figure 2A). Additionally, the difficulty of separating k and A increased when the measurements were noisier (Figure 2B). Sequences with separable parameters at low measurement error (e.g. $kA \sim 10\,\text{min}^{-1}\text{M}^{-1}$ and relative error < 0.2) became non-separable when the measurement error was large (e.g. relative error $= 1.0$).

While the ability to estimate k and A separately is of general interest for kinetic measurements, we found they could not be estimated separately for most of sequences within a Hamming distance of 2 to the family centers in the variant pool (Supplementary Figure S11). Thus, for the purpose of analyzing accuracy and uncertainty for k-Seq over a wide range of activities (analysis below), we focus on the estimation of the product kA.

k-Seq of ribozyme mutants: data pre-processing

We conducted the k-Seq experiment on a multiplexed sample containing mixed pools of variants of ribozymes S-1A.1-a, S-1B.1-a, S-2.1-a and S-3.1-a, using the expanded experimental conditions evaluated above (2–1250 μM substrate). A known amount of an unrelated RNA sequence (the ‘spike-in’ sequence) was added to each reaction to aid absolute quantitation. After demultiplexing the reads, we obtained 39 151 684 paired-end reads in the unreacted sample and a mean of 13 057 929 (s.d. = 4 359 249) paired-end reads in reacted samples (Supplementary Figure S13A). Around 90-92% of the reads were successfully joined in each sample (Supplementary Figure S13B, Supplementary Table S3). Dereplication, removal of reads with length not equal to 21, removal of reads with ambiguous nucleotides (‘N’), and removal of the spike-in sequence reads yielded a count table of the number of reads for each unique sequence detected in each sample. On average 87.9% (s.d. = 1.1%) of total reads were preserved in the samples.

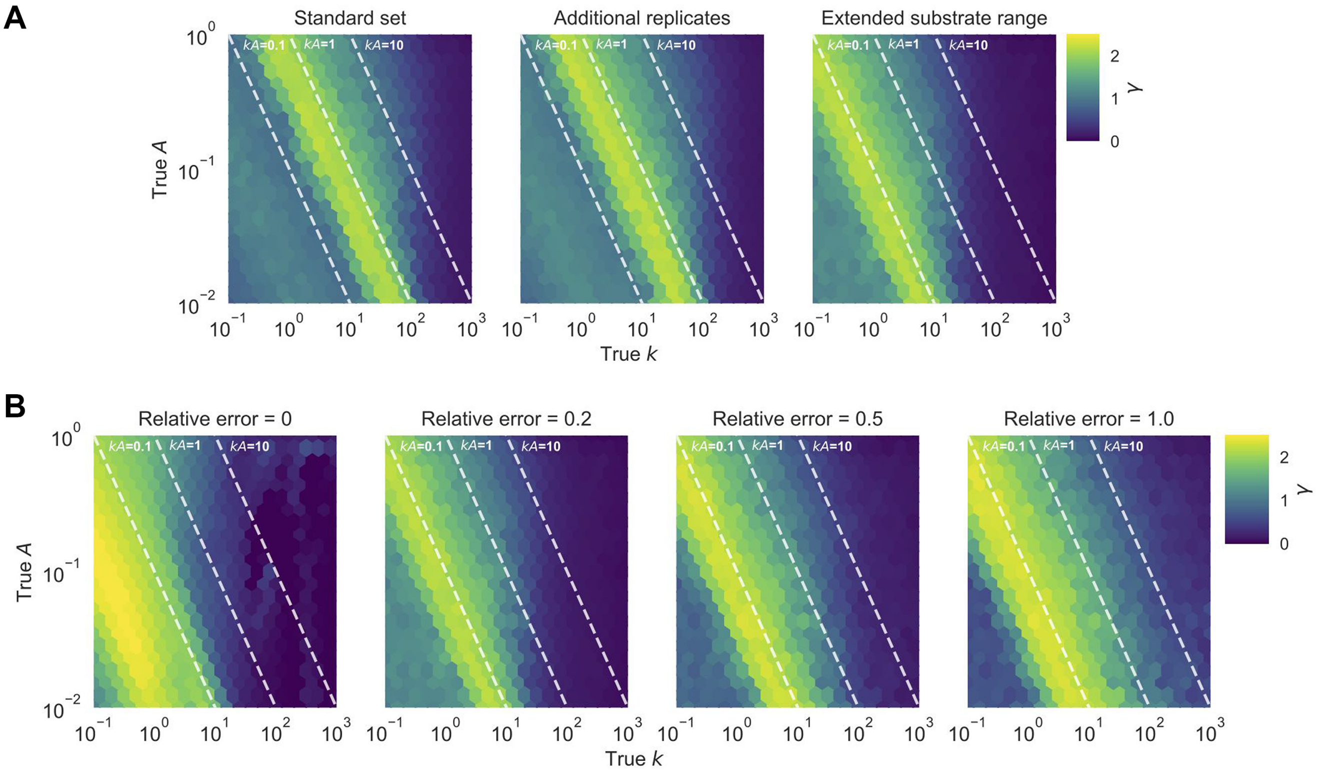

Figure 2. Effect of experimental factors on model identifiability to separately estimate k and A. Identifiability was evaluated using metric γ, based on the simulated effects of (**A**) choice of BYO samples (with relative error = 0.2) and (**B**) relative error (using the BYO series of the extended substrate range). Reacted fractions for 10 201 (101^2) simulated sequences with true k, A in the parameter space shown in the figure were fit to the pseudo-first order model, and γ values for each sequence were calculated from 100 bootstrapped samples. Higher values of γ indicate that k and A are less separable. (A) Choosing a wider range of BYO concentration is more effective in improving the region of identifiable data compared to adding more replicates of the same BYO concentrations. (B) With higher measurement error, k and A become increasingly difficult to estimate separately.

In principle, in order to calculate the reacted fraction with at least one non-zero value for fitting, a sequence must be detected in the unreacted sample (denominator) and in at least one of the reacted samples (numerator). Using this initial criterion, 764 756 valid unique sequences were considered to be analyzable for least-squares fitting to a pseudo-first order kinetic model, which comprised 77.9% of total reads in the unreacted sample and an average of 87.7% (s.d. = 0.6%) of total reads among reacted samples (Supplementary Figure S13B, Supplementary Table S3).

Absolute quantitation of sequence concentration

While the relative abundance of each sequence in a particular reaction sample can be calculated by dividing read counts by the total reads in each sample, calculation of the reacted fraction of each sequence requires comparing the absolute quantity of each sequence in each sample to the quantity of that sequence in the unreacted sample. This can be done by measuring the absolute RNA quantity in each sample. We compared two methods: (i) spiking in a sequence at a known concentration into each sample, providing a conversion between the number of reads and absolute concentration in each sample; or (ii) measurement of the total absolute RNA concentration of each sample by Qubit or qPCR. Sample quantitation by both methods agreed well with each other (Figure 3), with both having comparable

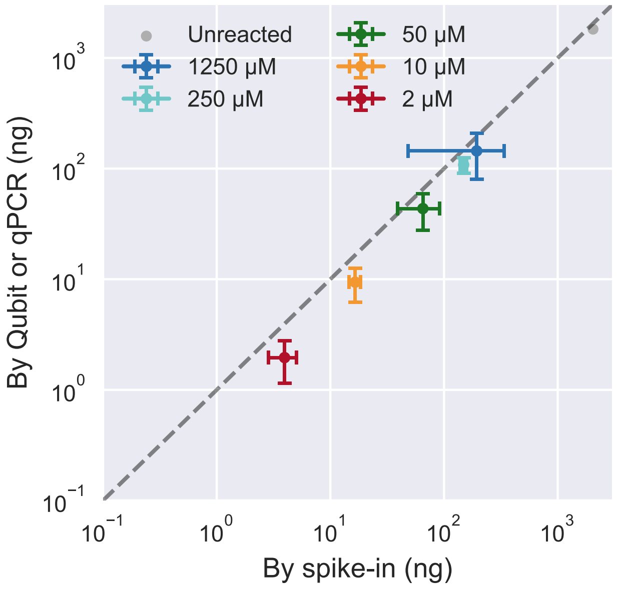

Figure 3. Comparison of RNA quantitation methods for k-Seq. Total RNA amount quantified for samples incubated with different BYO concentrations, determined by spike-in method vs. direct quantification using Qubit or qPCR, correlates well (Pearson's r = 0.999, P-value = 2.39 × 10^{-21}) and with comparable relative standard deviation (Supplementary Figure S2). Error bars show standard deviations calculated from triplicates for reacted samples.

relative standard deviation among triplicates (Supplementary Figure S2). Further analysis was done based on the second method for quantitation, since the first method is disadvantageous in reducing the HTS reads available for other sequences and requiring additional bioinformatic steps.

Distribution of ribozyme mutants in the variant pool

For each chosen ribozyme (S-1A.1-a, S-1B.1-a, S-2.1-a, S-3.1-a), a variant pool was chemically synthesized such that each position had the wild-type identity with 91% probability, with non-wild-type residues being equally probable (i.e. 3% each). In theory, each variant pool should contain ∼14% wild-type ($d = 0$), ∼0.45% of each single mutant ($d = 1$), and ∼0.015% of each double mutant ($d = 2$), or a ratio of ∼0.033 for the abundance of a d-th order mutant to a ($d -$ 1)-th order mutant (Supplementary Text S2, S3). The wild-type probability was selected to maximize the relative abundance of double mutants (Supplementary Figure S14). We sequenced the unreacted pool and categorized each sequence read according to ribozyme families (1A.1, 1B.1, 2.1, 3.1) and the Hamming distance to the family center. Sequencing results confirmed that the mixed variant pools followed the design (Figure 4A). The variant pools contained at least ∼1000 reads per sequence for $d \leq 2$ (up to double mutants, see Figure 4A and Table 1), and a mean of 39.7 reads per sequence for $d = 3$. Thus, the unreacted pool showed good coverage of analyzable sequences for $d \leq 3$ (Table 1).

Sequencing errors of a common sequence can spuriously inflate the apparent counts of a related sequence, a particularly acute problem if the pool is uneven and a small number of sequences (e.g. wild-type sequences) are very highly represented. Consequently, sequencing errors could potentially confound k-Seq results and lead to incorrect estimation of the kinetic coefficients. This may be a particular problem for less abundant or less active sequences that are closely related to more abundant and active sequences (e.g. resulting in estimation of parameters being biased toward those of the abundant sequence). There are two distinct effects related to this problem. First, the number of reads observed for a given sequence is lower than the true number, due to erroneous reads that are assigned to other sequences. Second, the number of reads observed for a sequence is inflated by the contribution of erroneous reads arising from related sequences. The combination of the two effects could change the observed sequence abundances substantially at high sequencing error rate (Figure 4B). In our variant pool, containing a 9% mutation rate, a sequencing error rate of 1% could cause >10% of reads for a mutant ($d \geq 1$) to be the result of sequencing error from its neighbors (Supplementary Figure S15). On the other hand, the most abundant sequences in the doped pool, wild-type sequences ($d = 0$), were least affected by the sequences from its neighbors. This problem can be mitigated by decreasing the error rate. If the sequence length is small enough to be covered by paired-end sequencing, requiring absolute matching of the overlapped region between the paired-end reads of a single sequence during joining should result in compounded fidelity (e.g. from an error frequency of 1% to 0.01%). With a sequencing error rate of 0.01%, the fraction of spurious reads from neighboring sequences was reduced to <0.5% for up to quadruple mutants ($d \leq 4$) without significant loss of reads during joining (Supplementary Figure S13B, Supplementary Table S3).

Accuracy of *k*-Seq estimation of model parameters and uncertainty

To evaluate the accuracy of k-Seq for estimating kinetic model parameters, sequences with read counts in unreacted and reacted pools were simulated using parameters (k and A) estimated from the variant pool k-Seq experiment as the 'ground truth'. The data set was constructed to simulate an experiment in which ribozymes were reacted with the extended BYO concentration series in triplicate. These data were fitted according to the pseudo-first order kinetic model to estimate k and A. We expected that sequences having a low number of counts in the data set would show reduced estimation accuracy. To characterize this effect, we plotted the ratio of the point estimate of kA to the true kA against the average number of counts across the simulated samples (Figure 5A). We found that the relative error in estimation for sequences with high mean counts (>1000) was <10% and <2-fold for a mean count around 100. However, the error increased substantially as the mean count decreased <100. Thus, sequences with low mean counts (especially <100), either from low abundance in the input pool or low abundance in the reacted pool (due to low activity), were susceptible to high error in estimation. Meanwhile, very high mean counts (e.g. >10 000) would not substantially benefit the measurement, as other sources of experimental error would likely be greater (2,15). Thus, the results indicate that >1000 mean count would be favorable for estimation, with >100 counts being acceptable if a 2-fold error in estimation is tolerated.

We also compared the k-Seq-estimated kA for selected sequences with an independent experimental measurement, namely gel-shift assays reported in (2). For seven unique sequences measured in both techniques, the results were well-correlated (Pearson's $r = 0.835$, P-value $= 1.94 \times 10^{-2}$ and Spearman's $\rho = 0.750$, P-value $= 5.22 \times 10^{-2}$; Supplementary Figure S16). This level of correlation is similar to that seen between the two k-Seq experiments (see section on Precision of k-Seq estimation).

While the above analysis demonstrated the accuracy of point estimation, a reliable quantification of uncertainty is required to assess the precision in estimating from real data when the ground truth is unknown. We therefore explored the accuracy of uncertainty quantification using bootstrapping. Bootstrap resampling ($n = 100$) was used to estimate the 95% confidence intervals (CI-95) in two ways: first, using mean and standard deviation of estimated kA (mean ± 1.96 s.d., assuming a normal distribution), and second, using the 2.5-percentile to 97.5-percentile confidence intervals (normal distribution not assumed), for sequences in the simulated pool dataset that were analyzable (602 246 sequences in total). A sensible evaluation of CI-95 estimation is the fraction of sequences with true kA value included in the estimated CI-95. If the estimation were correct, the CI-95 would include the true value for roughly 95% of sequences. We found bootstrapping gave an accurate CI-95

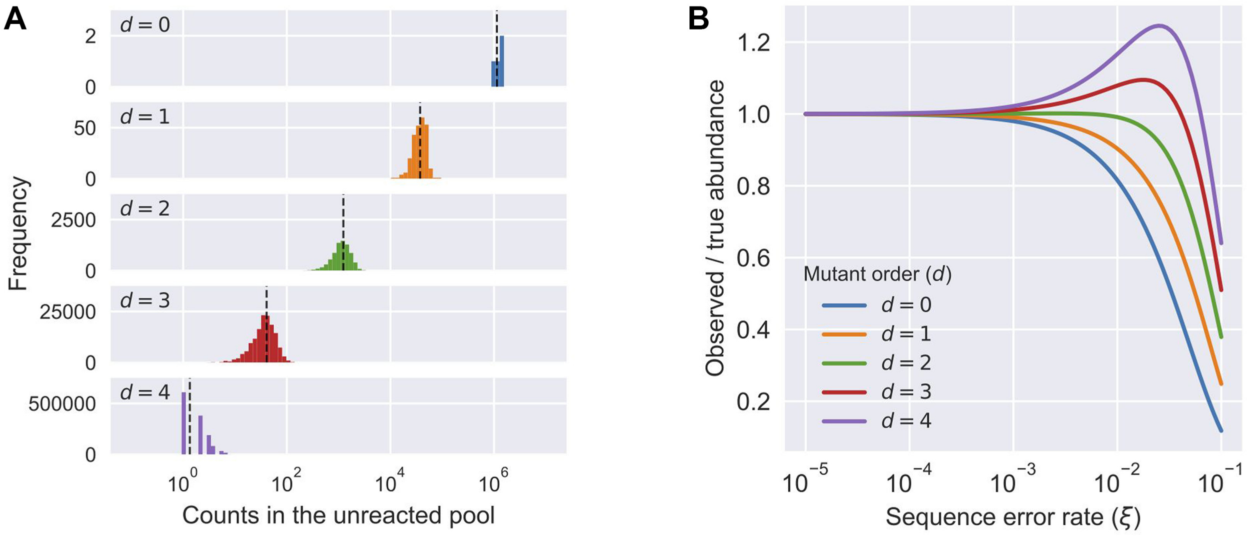

Figure 4. Distribution of mutants in the pool and the effect of sequencing error. (**A**) Relative abundance (counts) of sequences in the unreacted pool (four ribozyme families, total number of reads = 32 931 917), categorized by Hamming distance to its nearest family center. Observed abundance of different classes was similar to the expected number of counts (black dashed line). (**B**) The effect of different levels of sequencing error (ξ) to the expected observed abundance as the ratio to the true abundance for mutants with different orders (d) in a variant pool with 9% mutation rate. Due to the mixed effects of losing counts from being misidentified to a neighboring sequence and gaining counts from the misidentification of a neighboring sequence, the observed abundance for a sequence would either decrease (d = 0, 1) or first increase then decrease (d = 2, 3, 4) as the sequencing error increases. See Supplementary Text S3 for calculation details.

Table 1. Coverage of local sequence space in the variant pool containing four ribozyme families. N/A = not applicable. s.d. = standard deviation. The calculation of expected counts in the unreacted pool does not include effects of sequencing error

Order of mutants (d)	# of unique sequences	# of analyzable sequences	Fraction of possible unique sequences that were analyzable	Mean counts in the unreacted pool (s.d.)	Expected counts in the unreacted pool
0	4	4	1.000	1 243 500 (151 356)	1 136 125
1	252	252	1.000	37 599.9 (10 607.1)	37 455
2	7560	7560	1.000	1198.3 (454.8)	1234
3	143 640	143 482	0.999	39.7 (18.9)	40.7
4	1 939 140	590 115	0.304	1.7 (1.3)	1.34
≥5	N/A	23 343	N/A	0.7 (0.5)	N/A

estimation by either method. 96.5% of sequences included the true kA within the estimated CI-95 from 2.5-to-97.5-percentiles, and 96.4% did so when estimating CI-95 from the mean and standard deviation. Of note, the fractions of sequences with their true kA included in the CI-95 were relatively consistent regardless of their mean counts (Figure 5B) or true kA values (Supplementary Figure S17), indicating that these methods could be used broadly to quantify uncertainty for sequences across different abundance or activity values. For comparison, we also examined uncertainty estimation using the mean and standard deviation estimated from triplicates (mean ± 1.96 s.d.). In our simulated pool dataset, 83.5% of sequences had the true kA value included in the CI-95 estimated from triplicates, indicating that uncertainty is underestimated for a substantial fraction of sequences (i.e. overconfident in estimated values) (Figure 5B).

Precision of *k*-Seq estimation: experimental data set

The precision of *k*-Seq measurement for data from the variant pool experiment was evaluated in two ways. First, given the reasonable accuracy obtained by the bootstrapping procedure, we calculated the fold-range (97.5-percentile divided by 2.5-percentile) estimated from bootstrapping (n = 100). While there was a slight tendency for sequences with higher kA to have higher estimation precision (lower fold-range; Supplementary Figure S18) in each order of mutants, the precision was more evidently dependent on the mean counts value for sequences, both within and across groups (Figure 6A). All wild-type sequences and most single and double mutants had CI-95 spanning less than one order of magnitude (fold-range < 10). Triple mutants had generally lower precision, consistent with their lower counts.

Precision as measured above represents variation among replicates done in the same experimental batch but does not include variation between different *k*-Seq experiments. To understand the precision of estimates from independently designed and separately executed *k*-Seq experiments, we compared the results from the variant pool *k*-Seq reported here to a previously reported *k*-Seq assay from a selection pool (2). 2513 unique sequences were found for which the 2.5-percentile for kA, estimated from bootstrapping, was greater than the baseline kA of 0.124 $\text{min}^{-1}\text{M}^{-1}$ (measured in (2)) in both experimental data sets. Point estimates of kA

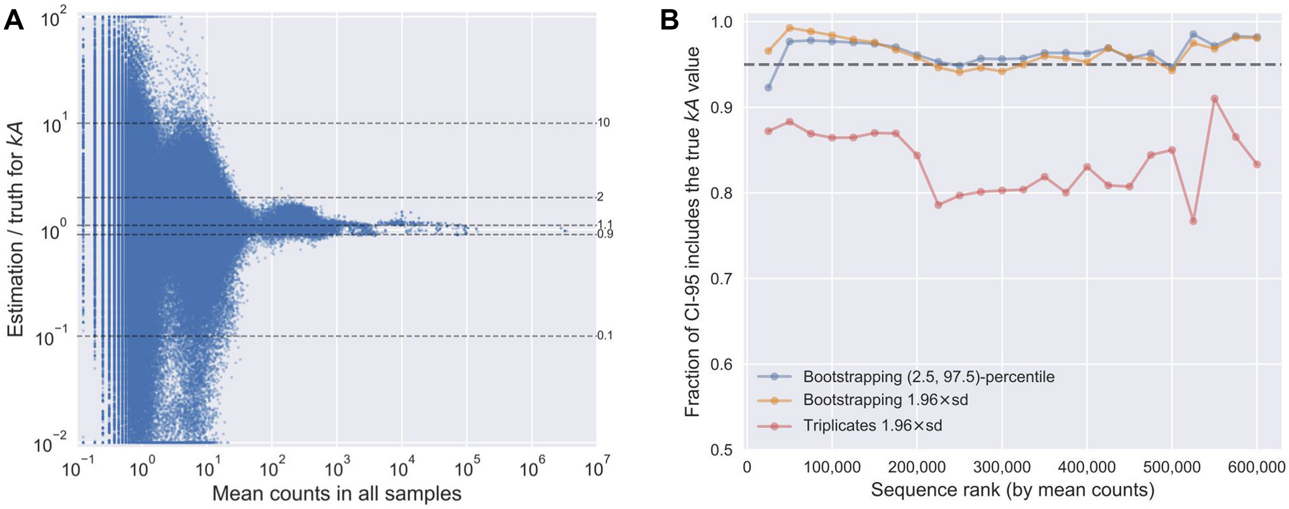

Figure 5. Accuracy of parameter estimation by k-Seq. (**A**) Dependence of accuracy (ratio of estimated kA to true kA) on mean counts across all simulated samples (including the unreacted pool sample). The dashed lines correspond to ratios as labeled. Ratios >100-fold or <0.01-fold are shown at the borders of the plot. (**B**) Fraction of sequences for which the CI-95, estimated using bootstrapping or using triplicates, includes the true kA values, for sequences with different mean counts across all samples. Sequences were ranked by mean counts (from highest to lowest) and binned in sets of 25 000 sequences. Each data point indicates the fraction of CI-95 that includes the true values in each bin.

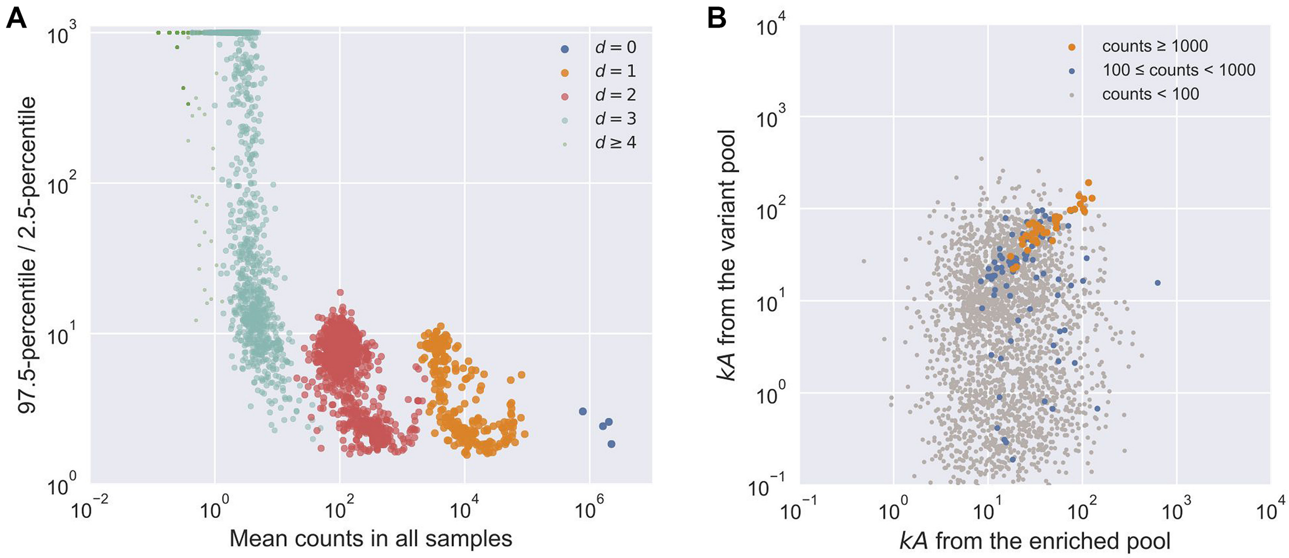

Figure 6. Precision of estimation by k-Seq. (**A**) Fold-range (97.5-percentile/2.5-percentile) of kA estimation depended on the mean counts. Increasing mean counts increases precision, as shown by the relationship of fold-range with mean counts across different orders of mutants. For $d \geq 2$, only 1000 sequences were randomly selected for visualization. (**B**) Alignment between estimated kA from two independently conducted experiments (experiment from (2), and the k-Seq experiment reported here). Only sequences with 2.5-percentile higher than baseline catalytic coefficient ($kA = 0.124\,\mathrm{min}^{-1}\mathrm{M}^{-1}$, reported in (2)) were included. Each point represents a sequence whose color reflects the minimum of mean counts (between two experiments).

for these sequences from the two experiments were compared to each other (Figure 6B). For sequences with sufficient counts (e.g. mean counts of at least 1000 in both experiments, corresponding to 39 sequences), the results from the two experiments were well correlated (Pearson's $r = 0.896$, P-value $= 1.20 \times 10^{-14}$; Spearman's $\rho = 0.864$, P-value $=$ 1.38×10^{-12}), indicating good repeatability of those measurements from different experiments. As expected, lowering the count threshold gave a decrease in repeatability: sequences with mean counts greater than 100 in both experiments (119 sequences) showed Pearson's $r = 0.183$ (P-value $= 4.68 \times 10^{-2}$) and Spearman's $\rho = 0.460$(P-value $=$ 1.44×10^{-7}). For analyzable sequences with mean counts <100 in either experiment, weak to no correlation was found (Pearson's $r = –0.0234$, P-value $= 2.52 \times 10^{-1}$; Spearman's $\rho = 0.0505$, P-value $= 1.34 \times 10^{-2}$).

DISCUSSION

In this work, we addressed several issues related to the rigor of inferring kinetic model parameters from k-Seq analysis, namely model identifiability, sequencing error, and estimation of uncertainty.

A model is not identifiable when the optimal set of parameters fitting the model cannot be uniquely determined due to noise in the data (12,16). For the ribozymes exhibiting pseudo-first order kinetics studied here, the parameters k (rate constant) and A (maximum reaction amplitude) could not be separately estimated when the data collected did not show saturation behavior, i.e. the data fell into the initial linear region of the curve. While it would be possible to adjust the substrate concentrations to mitigate the problem for individual sequences, it is impossible in k-Seq experiments to apply optimal conditions for each sequence due to pool heterogeneity. We previously reported the product kA as the measure of chemical activity (2). However, separate estimation of k and A is still an important goal. In general, we found that higher values of k or A and lower noise level yield better separability. Of the metrics (σ_A and γ) we calculated from bootstrapping results, both showed good performance scoring the separability of k and A. These metrics allows one to semi-quantitatively assess separability, in combination with experimental intuition, to determine which results can be reasonably reported as k and A separately vs. reported as the product kA.

Using a simulated data set, we studied how an experiment might be designed to increase the number of sequences in the separable region. Extending the substrate concentration range (to 1250 μM) expands the region of separable sequences by pushing the lower bound on k or A down by roughly one order of magnitude. By contrast, adding another replicate to each substrate concentration did not substantially change the separability map. Thus, we used the extended substrate range in the present k-Seq experiment to provide a wider dynamic range. Even so, a substantial fraction of sequences exhibited non-separable kA, so in practice the choice of parameters to be reported depends on the goals of the experiment (e.g. for maximum exploitation of the data, all of k, A, and kA could be reported along with σ_A or γ).

Experimental biases from non-specific retention of RNA and PCR amplification are two possible concerns for k-Seq-type experiments. In the aminoacylation ribozyme system used here, non-specific binding to the streptavidin beads was found to be small and thus should have little impact on RNA quantification (Supplementary Figure S19). For systems with substantial non-specific binding, a control experiment without substrate may be used for baseline subtraction to better estimate the amount of RNA isolated due to reaction. Differences in PCR amplification efficiency could potentially affect kinetic estimation. Since this bias is largely driven by extremes in GC content (17,18), some pool designs, such as that used here (GC content ranging from 42-46% for the wild-types and 40-48% for the double mutants), are less likely to exhibit substantial bias. For pools in which PCR bias is anticipated to be larger, we suggest the following experimental design. If the number of cycles is kept constant across samples, including the unreacted sample, then the factor by which PCR biases affect the true quantity should be constant across samples for a given sequence. Because the reacted fraction is determined by the ratio between the amount of the sequence in the reacted sample versus the unreacted sample, this sequence-specific factor should cancel out and thus minimize the effect of PCR bias. Another simple measure to reduce PCR bias is to minimize the number of PCR cycles. In our case, 10 or fewer cycles were used to recover RNA. Finally, another method to correct for PCR bias could also be applied, such as unique molecular identifiers (UMI) to track and deduplicate original sequences (19). Ultimately, k-Seq or similar assays should be validated by measurement of individual sequences using an appropriate orthogonal technique. Estimation of error in the orthogonal technique is also important to assess the agreement between the two techniques. The assay presented here was validated by gel-shift measurement of the kinetics for 10 individual sequences (Supplementary Figure S16) measured in (2). Ideally, a high Pearson correlation coefficient should be seen across the dynamic range of the assay. The Spearman correlation is relevant in cases where there is a nonlinear relationship between measurements in different assays (such as may be caused by a saturation effect in one assay), or in case the relative ranking of sequences is more important than the absolute measurements.

Using DNA sequencing counts to quantify the abundance of sequences has two consequences that need consideration: sequencing error that mis-identifies a sequence as a related sequence, and stochastic noise in measurement for sequences associated with lower counts. The first is a particular problem when sequences are close in the sequence space (e.g. mutational analysis on a variant pool) and the pool is quite uneven. In this case, sequencing errors can contribute a non-trivial portion of reads to less abundant neighbor sequences, effectively mixing the k-Seq measurement for these less abundant sequences with those of their neighbors. While model-based sequencing error correction could be attempted (20,21), this problem can also be circumvented by paired-end reads. By enforcing absolute matching during the joining process, the error rate of sequencing (e.g. ∼1% per base) can be substantially decreased (e.g. to 0.01% per base) if the entire sequence was read in both directions. While this decreases the number of reads that passed quality control (Supplementary Table S3), the benefit is important as misidentification from this level of sequencing error was essentially negligible for observed abundances (Supplementary Figure S15).

Low counts had a major effect on the accuracy of estimation of kinetic coefficients. In practice, we found that the average counts for a sequence across samples (i.e. mean counts) was a better guide for estimation accuracy compared to counts in the input pool (Figure 5A, Supplementary Figure S20). Accuracy was good for sequences above a certain threshold of mean counts (e.g. 100 reads per sample) but decreased markedly below this. At the same time, the benefit from large counts (e.g. 10 000 reads per sample) was marginal and other experimental factors likely contribute greater error. The activity of a sequence also affected the accuracy, as lower counts were found in the reacted samples for low activity sequences. Nevertheless, our simulated count data showed that low activity sequences ($kA < 1\ \mathrm{min}^{-1}\mathrm{M}^{-1}$) still yielded reasonable accuracy (∼2-fold error) if abundance in the initial pool was high enough (e.g. >1000 counts, Supplementary Figure S21). Thus, k-Seq can be effective for low activity sequences if they are abundant in the pool, and the accuracy of estimation could be improved by increasing sequencing depth or altering the library design. To check that low abundance in the unreacted pool does not cause the analysis to miss potentially active sequences, one may look for sequences with high mean counts in the reacted samples that do not appear in the unreacted sequence reads. The experimental design used here did not yield any such cases (Supplementary Table S3), but this check should be performed for other experiments when needed.

Estimating uncertainty is important for *k*-Seq experiments, but replicates are likely to be limited due to the expense associated with HTS. Bootstrapping simulates virtual replicates by resampling data (in this case, the relative residuals from fitting) with replacement to its original size. The results can be used to estimate population characteristics, such as confidence intervals for estimated model parameters. Indeed, bootstrapping results reflected the true 95% uncertainty level more appropriately than the standard deviation estimated from triplicate experiments, as the latter tended to underestimate the uncertainty. As seen for the accuracy of low count sequences, precision showed a steep drop-off when mean counts dropped <100 (Figure 6A), while large counts also did not significantly improve precision. Using bootstrapping instead of replicates also provided resampling data to calculate σ_A and γ for model identifiability analysis. As modern computational resources become cheaper and easier to access, bootstrapping, despite its higher computation cost, becomes more affordable. Therefore, while experimental replicates are valuable for controlling for some sources of error, we suggest that bootstrapping analysis is an excellent method for properly estimating errors and understanding model identifiability.

To maximize coverage, or the number of sequences with estimated model parameters having acceptable accuracy and precision, it is desirable to maximize the number of sequences satisfying a minimal count requirement without spending excessive sequencing resources on abundant sequences. In the present experiment, we had approximately full coverage for single and double mutants in each family, for which the measurement precision may be considered reasonable (fold-range < 10; mean counts > 10) (Figure 6A). While HTS technology enabled the kinetic measurement for large pools with high richness (number of unique sequences), the practical coverage for *k*-Seq is affected by pool evenness, as highly uneven pools may have many sequences with insufficient counts for precise estimation. Such pools may result from enrichment after selections or from variant pool synthesis of ribozyme variants exploring many mutants of a given wild type. For enriched pools from selection experiments, the pool evenness usually decreases during the selection. For doped pools, evenness is tuned by the ratio of wild-type nucleotides at each position. In the analysis of BYO-aminoacylation ribozymes presented here, the designed variant pool was more even than the selection pool from which these ribozymes were derived (Supplementary Figure S22); thus *k*-Seq analysis of selected ribozymes may be improved by designing a new pool rather than directly analyzing the selected pool itself. The design of the variant pool used here was optimized for maximum representation of double mutants (mutation rate $\eta = 0.09$), yielding >1000 reads per sequence at a sequencing depth of 10^7 reads per sample (Supplementary Text S2, Supplementary Figure S14). Increasing η would decrease the number of reads of double mutants while increasing the number of reads of triple and higher-order mutants. However, due to the very high number of possible triple mutants and limited sequencing depth ($\sim 10^7$) done at the time, triple mutants would have had <100 counts on average even at an η optimized for them. Other experimental designs may choose to explore higher-order mutants at a cost of lower estimation certainty, depending on the purpose of the investigation. Increasing sequencing depth (e.g. $>10^8$ reads, which is now readily available), would enable complete coverage of the triple mutants at reasonable counts per sequence (>100 and possibly >1000), such that a choice of higher η, such as 0.15, could give both full coverage of the triple mutants and good estimation of kinetics. Alternatively, different pool synthesis methods (e.g. the precise synthesis of large uniform oligo pools (22)) that allow increased evenness would allow greater sequence exploration.

In this work, we presented a model analysis pipeline to apply *k*-Seq to a pseudo-first order kinetic system with varying initial substrate concentrations. This pipeline can be readily applied to any kinetic model with a closed integral form, including first-order kinetics with varying time and second-order kinetics (Supplementary Text S1). Pool composition and the resulting sequencing coverage are critical parameters of the *k*-Seq measurement. Greater evenness in the samples may be achieved in the near future by emerging pool synthesis technologies as well as by improvements in sequencing technology to achieve greater total sequencing depth. With bootstrapping, data with low counts can be assessed with estimated uncertainty rather than discarded. On the other hand, systematic errors due to sequencing errors cannot be assessed by replicates or bootstrapping analysis; instead, effort would be well-spent reducing sequencing errors, and the degree to which results might be biased by the resulting error rate should be kept in mind when interpreting the data. Attention to these issues is important for fulfilling the promise of kinetic sequencing and related techniques for providing unprecedented insights into genotype-phenotype relationships.

DATA AVAILABILITY

'k-seq' is an open-source Python package used for analyzing data in this work, available in the GitHub repository (https://github.com/ichen-lab-ucsb/k-seq). Sequencing data, simulated datasets, *k*-Seq results, and code for data analysis are available at Dryad Data Repository (https://doi.org/10.25349/D9T02R).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Use was made of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific

Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara.

FUNDING

Simons Collaboration on the Origins of Life [290356FY18]; US National Aeronautics and Space Administration [NNX16AJ32G]; National Institutes of Health New Innovator Program [DP2GM123457]; Camille Dreyfus Teacher-Scholar Program. Funding for open access charge: National Institutes of Health [DP2GM123457].

Conflict of interest statement. None declared.

REFERENCES

1. Dhamodharan,V., Kobori,S. and Yokobayashi,Y. (2017) Large scale mutational and kinetic analysis of a self-hydrolyzing deoxyribozyme. *ACS Chem. Biol.*, **12**, 2940–2945.
2. Pressman,A.D., Liu,Z., Janzen,E., Blanco,C., Müller,U.F., Joyce,G.F., Pascal,R. and Chen,I.A. (2019) Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J. Am. Chem. Soc.*, **141**, 6213–6223.
3. Kobori,S. and Yokobayashi,Y. (2016) High-throughput mutational analysis of a twister ribozyme. *Angew. Chem. Int. Ed.*, **55**, 10354–10357.
4. Andreasson,J.O.L., Savinov,A., Block,S.M. and Greenleaf,W.J. (2020) Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. *Nat. Commun.*, **11**, 1663.
5. Yokobayashi,Y. (2020) High-throughput analysis and engineering of ribozymes and deoxyribozymes by sequencing. *Acc. Chem. Res.*, **53**, 2903–2912.
6. Niland,C.N., Jankowsky,E. and Harris,M.E. (2016) Optimization of high-throughput sequencing kinetics for determining enzymatic rate constants of thousands of RNA substrates. *Anal. Biochem.*, **510**, 1–10.
7. Tome,J.M., Ozer,A., Pagano,J.M., Gheba,D., Schroth,G.P. and Lis,J.T. (2014) Comprehensive analysis of RNA-protein interactions by high-throughput sequencing–RNA affinity profiling. *Nat. Methods*, **11**, 683–688.
8. Le,D.D., Shimko,T.C., Aditham,A.K., Keys,A.M., Longwell,S.A., Orenstein,Y. and Fordyce,P.M. (2018) Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E3702–E3711.
9. Jalali-Yazdi,F., Huong Lai,L., Takahashi,T.T. and Roberts,R.W. (2016) High-throughput measurement of binding kinetics by mRNA display and next-generation sequencing. *Angew. Chem. Int. Ed.*, **55**, 4007–4010.
10. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–807.
11. Tack,D.S., Tonner,P.D., Pressman,A., Olson,N.D., Levy,S.F., Romantseva,E.F., Alperovich,N., Vasilyeva,O. and Ross,D. (2020) The genotype-phenotype landscape of an allosteric protein. bioRxiv doi: https://doi.org/10.1101/2020.07.10.197574, 11 July 2020, preprint: not peer reviewed.
12. Daly,A.C., Gavaghan,D., Cooper,J. and Tavener,S. (2018) Inference-based assessment of parameter identifiability in nonlinear biological models. *J. R. Soc. Interface*, **15**, 20180318.
13. Blanco,C., Verbanic,S., Seelig,B. and Chen,I.A. (2020) EasyDIVER: a pipeline for assembling and counting high-throughput sequencing data from in vitro evolution of nucleic acids or peptides. *J. Mol. Evol.*, **88**, 477–481.
14. Masella,A.P., Bartram,A.K., Truszkowski,J.M., Brown,D.G. and Neufeld,J.D. (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, **13**, 31.
15. Pressman,A., Moretti,J.E., Campbell,G.W., Müller,U.F. and Chen,I.A. (2017) Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res.*, **45**, 8167–8179.
16. Hines,K.E., Middendorf,T.R. and Aldrich,R.W. (2014) Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach. *J. Gen. Physiol.*, **143**, 401–416.
17. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
18. Laursen,M.F. (2017) Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. *Front. Microbiol.*, **8**, 8.
19. Kivioja,T., Vähärautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
20. Callahan,B.J., McMurdie,P.J., Rosen,M.J., Han,A.W., Johnson,A.J.A. and Holmes,S.P. (2016) DADA2: high-resolution sample inference from illumina amplicon data. *Nat. Methods*, **13**, 581–583.
21. Xulvi-Brunet,R., Campbell,G.W., Rajamani,S., Jiménez,J.I. and Chen,I.A. (2016) Computational analysis of fitness landscapes and evolutionary networks from in vitro evolution experiments. *Methods*, **106**, 86–96.
22. Almagro,J.C., Pedraza-Escalona,M., Arrieta,H.I. and Pérez-Tapia,S.M. (2019) Phage display libraries for antibody therapeutic discovery and development. *Antibodies*, **8**, 44.