Received: 10 May 2022 | Revised: 14 July 2022 | Accepted: 15 July 2022

DOI: 10.1002/pro.4401

METHODS AND APPLICATIONS

Randomized gates eliminate bias in sort-seq assays

Brian L. Trippe[1,2,3] | Buwei Huang[3,4] | Erika A. DeBenedictis[3,4] | Brian Coventry[3,4] | Nicholas Bhattacharya[2,5] | Kevin K. Yang[2] | David Baker[3,4,6] | Lorin Crawford[2]

[1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Microsoft Research New England, Cambridge, Massachusetts, USA

[3]Institute for Protein Design, University of Washington, Seattle, Washington, USA

[4]Department of Biochemistry, University of Washington, Seattle, Washington, USA

[5]Department of Mathematics, University of California Berkeley, Berkeley, California, USA

[6]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA

Correspondence

Brian L. Trippe, Massachusetts Institute of Technology, Cambridge, MA, USA.
Email: blt2114@columbia.edu

Lorin Crawford, Microsoft Research New England, Cambridge, MA, USA.
Email: lcrawford@microsoft.com

Funding information

David and Lucile Packard Foundation; National Science Foundation

Review Editor: Aitziber Cortajarena

Abstract

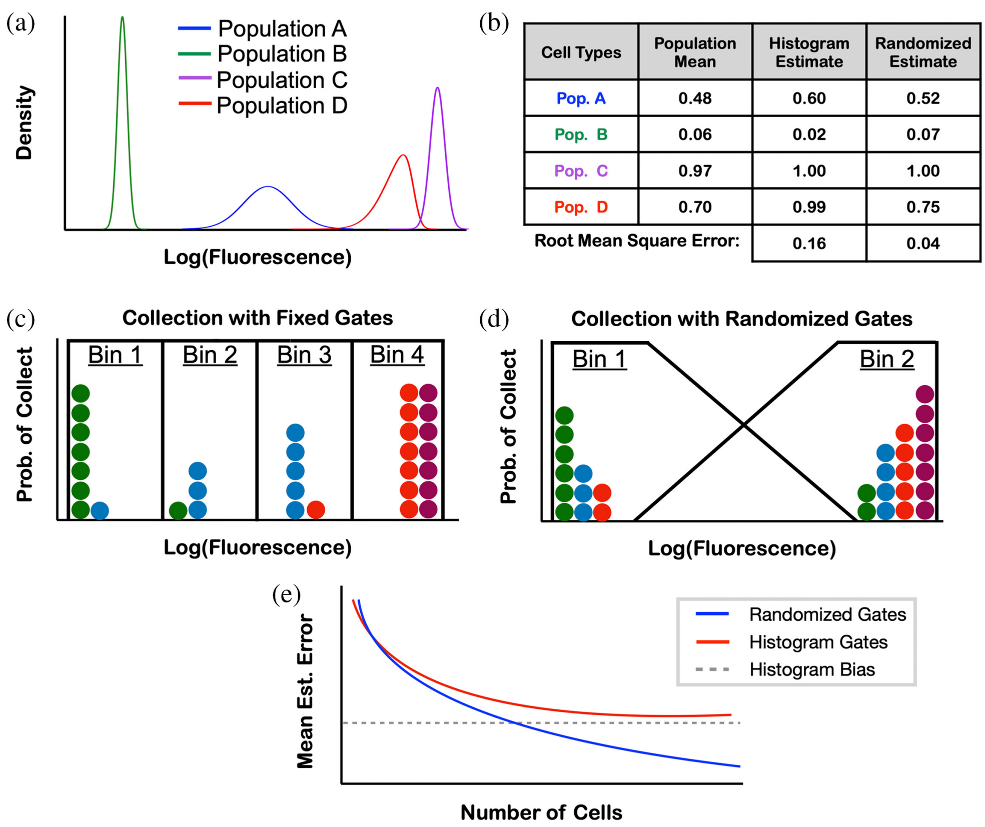
Sort-seq assays are a staple of the biological engineering toolkit, allowing researchers to profile many groups of cells based on any characteristic that can be tied to fluorescence. However, current approaches, which segregate cells into bins deterministically based on their measured fluorescence, introduce systematic bias. We describe a surprising result: one can obtain unbiased estimates by incorporating randomness into sorting. We validate this approach in simulation and experimentally, and describe extensions for both estimating group level variances and for using multi-bin sorters.

KEYWORDS

fluorescence activated cell sorting, high-throughput screening, multiplexed measurements, sort-seq assays, statistical methods

1 | INTRODUCTION

Quantitative, multiplexed assays relying on fluorescence activated cell sorting (FACS) followed by high-throughput sequencing are critical to modern biology and molecular engineering because they enable construction of large scale datasets connecting sequence to function. For example, these "sort-seq" assays are widely used to profile the strength of protein–protein binding interactions via yeast display.[1–4] In particular one (i) synthesizes a *library* of 10^4 to 10^5 DNA sequences encoding proteins that may bind to a target of interest; (ii) transforms the library into yeast such that each putative binder is expressed on the surface of a population of cells; (iii) incubates cells with fluorescently labeled target protein; (iv) physically separates 10^6 to 10^8 cells based on



Cell Types	Population Mean	Histogram Estimate	Randomized Estimate
Pop. A	0.48	0.60	0.52
Pop. B	0.06	0.02	0.07
Pop. C	0.97	1.00	1.00
Pop. D	0.70	0.99	0.75
Root Mean Square Error:		0.16	0.04

FIGURE 1 Schematic overview of randomized gates. (a) Distributions of log fluorescence for different cell populations and (b) their hypothetical true and estimated means. (c) An example of histogram approach with deterministic collection into four bins and (d) an example of randomized collection approach with two bins. (e) Estimated means of the randomized gating scheme are more accurate than the histogram approach as the number of collected cells increases

binding affinity by FACS; and finally, (v) quantifies the prevalence, and thereby binding affinity, of each library member by high throughput sequencing. Due to biological and technical variability, there is a distribution over (log) fluorescence for each library sequence, and the challenge is to estimate the means of each of these distributions (Figure 1a,b). For example, for binding interactions, this mean fluorescence relates directly to biophysical quantities of interest including dissociation constants and binding energies.[4–6]

In previous work, cells are deterministically segregated into one or more collection tubes (referred to as "bins") based on their measured fluorescences, and the mean fluorescence of each population is estimated from the histogram of observed sequence counts in each bin (Figure 1c). Peterman and Levine[7] compare the error associated with different strategies for collecting and analyzing such data, and they show that average squared error is the sum of contributions from *bias* and *variance* (e.g., Hastie et al.,[8] Chapter 7.3). The variance arises from experimental noise and variability across cells, and it can be reduced by increasing the number of cells screened. The bias arises from the discretization of the space of log fluorescence into bins (Figure 1b,c); for example, narrow distributions can be sorted all into the same bin but have means as different as the bin width. Moreover, this bias poses reproducibility challenges; the direction and magnitude of bias depends on how the bins are chosen, but this choice is subjective and commonly depends on variable experimental conditions. Because even the most sophisticated FACS machines can sort cells into at most six bins, resolution is limited. This low resolution limits the value of sort-seq data in quantitative analyses, for instance, by prohibiting computation of precise binding energies. This challenge has spurred much work on how to effectively reduce histogram bias.[5–7,9] One common approach seeks to overcome the resolution limits of histograms by assuming fluorescence is log-normally distributed for each population and using maximum likelihood estimation to estimate moments.[2,9–11] However, on real data, this assumption is violated and the resulting estimates can have greater bias than the naive approach (Figure S1).

2 | RESULTS AND DISCUSSION

In this work, we show that the bias generated using histograms can be eliminated altogether by incorporating randomness into FACS collection strategies with as few as two bins (Figure 1d), thereby obtaining arbitrarily accurate estimates with many cells (Figure 1e). To do this, we take a statistical approach. We consider a population of cells that pass through a 2-bin sorter, each with log fluorescence F independently and identically distributed according to a density function p_F. Our target of interest is the mean log fluorescence, $\boldsymbol{\mu}_F = \int f p_F(f) \mathrm{d}f$. Let B denote the bin (either 1 or 2) into which a cell is collected, and let Y_1 and Y_2 be the counts of cells in Bins 1 and 2 after sorting, respectively. In multiplexed sort-seq assays, we obtain Y_1 and Y_2 for thousands of populations, and our goal is to accurately estimate the mean of each population simultaneously.

For standard binning, a *gate* is chosen for each bin that defines the range of values F for which cells are collected into that bin; so, the bin B is deterministic once F is measured (e.g., as in Figure 1c). We instead consider *randomized gates* which define for each bin the probability of collecting a cell at each fluorescence (as in Figure 1d) and rely on pseudo-random numbers to determine the bin. For estimating population means, when the fluorescence measurements fall between lower and upper bounds L and U, one first sorts using randomized gates such that for any f on the interval $[L,U]$,

$$\mathbb{P}(B=1|F=f) = 1 - \frac{f-L}{U-L} \text{ and } \mathbb{P}(B=2|F=f) = \frac{f-L}{U-L}. \tag{1}$$

The counts are then combined into an empirical estimate of $\boldsymbol{\mu}_F$ as $\widehat{\boldsymbol{\mu}} = (U-L) \cdot Y_2/(Y_1+Y_2) + L$.

While one might expect introducing randomness to decrease precision by introducing additional noise, $\widehat{\boldsymbol{\mu}}$ is directly informative to the mean fluorescence. In particular, $\widehat{\boldsymbol{\mu}}$ is an *unbiased* estimate of the true population mean in the sense that the average value we would expect for $\widehat{\boldsymbol{\mu}}$ if we repeated the sort-seq experiment many times is equal to $\boldsymbol{\mu}_F$ (see Theorem 1 in Section 4).

This unbiasedness theorem guarantees that, in contrast to the histogram approach, we can get arbitrarily accurate estimates by screening a larger numbers of cells (Figures 1e and S2). More precisely, recalling that the mean squared error (MSE) is the sum of the bias squared and the variance,[8] unbiasedness implies that the error of $\widehat{\boldsymbol{\mu}}$ is dictated solely by its variance. Moreover, $\widehat{\boldsymbol{\mu}}$ allows a transparent trade-off between the number of cells sorted per population and the precision of the estimates; notably, with as few as 400 cells, a 95% confidence interval for $\boldsymbol{\mu}_F$ will cover at most 10% of the range from L to U (Section 4).

2.1 | Randomized gates provide superior accuracy to histograms in simulation

We used a simulation study to explore the implications of unbiasedness on estimation accuracy with the randomized gate approach relative to the standard histogram approach. In this study, we simulated fluorescence of 250 cells from log-normal distributions with different means and variances (Figure 2a). We then simulated sorting these cells based on their fluorescence either with four deterministic gates of equal width or with two randomized gates as dictated by Equation (1). For the deterministic gates, we constructed histograms and computed estimates of the mean fluorescence as the average of the bin centers weighted by the fraction of cells they contained; and for the randomized gates, we estimated the mean as $\widehat{\boldsymbol{\mu}}$. Figure 2b,c report the performance of these estimates in terms of MSE, along with their bias and variance components. As expected, the randomized gates approach has negligible bias except for broad distributions violating the conditions of our theorem (Section 4).

With even as few as 250 cells per population, the MSE of the histogram approach is dominated by bias. Accordingly, the unbiased randomized approach typically provides more accurate estimates. Notably, 250 cells is fewer than is the typical in sort-seq assays; with larger samples, more pronounced improvements are obtained (Figure S2). Because the histogram estimates are systematically biased toward bin centers, they can however be more accurate for narrow distributions with means near bin centers (Figure 2b).

2.2 | Experimental implementation via shifting gate thresholds

We next tested our approach experimentally. Current FACS software does not support randomized gate programming, so we devised an experimental approximation in which we manually changed the gating threshold 20 times during sorting at regular intervals (Section 4). We tested this procedure in the context of a binding assay using yeast display.[12] We synthesized DNA encoding four mini-protein binders to the SARS-COV-2 receptor binding domain (RBD) with a range of binding affinities.[1] While the value of this approach is greatest for highly multiplexed assays with many thousands of sequences, we chose this small number so that we could also test each binder easily in serial. We separately transformed and expressed each design in yeast and then incubated the populations with RBD. Both the target and binders were fluorescently labeled, and we considered the log ratio of target to binder fluorescence as an expression normalized proxy for binder strength.[6] We measured each sample on a Sony SH800 cell sorter separately, recording the binding signal for each binder (Figure 3a). We then pooled the samples together and sorted 1,000,000 cells, collecting 50,000 cells at each of the 20 thresholds (Section 4).

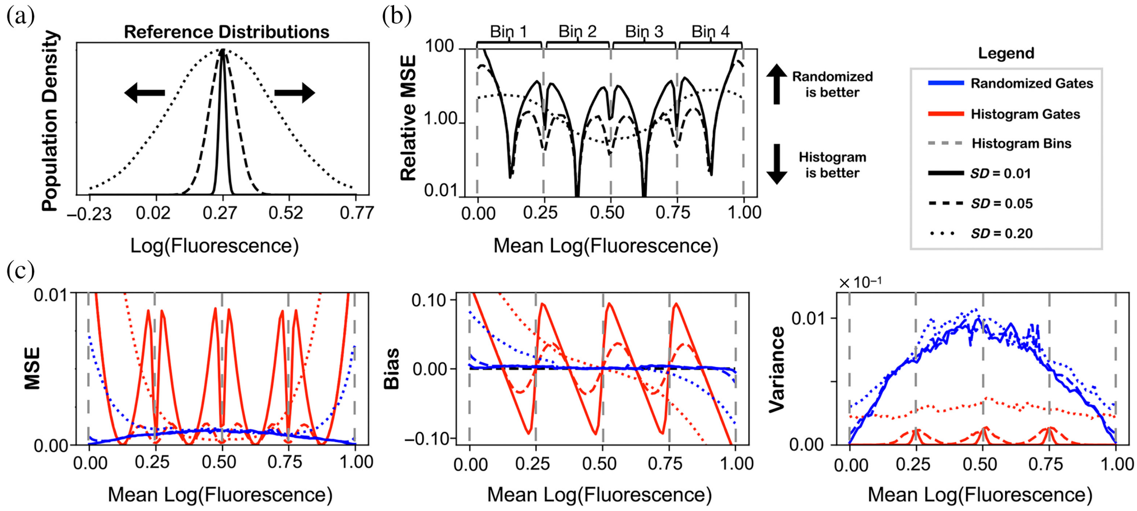


FIGURE 2 Simulation study reveals improved estimation properties obtained with randomized gates as compared to histograms. (a) Fluorescence values of cells are drawn independently from log-normal distributions with different scales and with varied means, where the black arrows represent simulated changes of the means. (b) The relative performances of estimates from histograms and randomized gates across a range of mean log fluorescences in terms of mean squared error (ratios greater than 1 reflect lower error with randomized gates and ratios below 1 reflect lower error with histograms). (c) The mean squared error (left) decomposed into bias (center) and variance (right) for both estimates. All points are the average across 200 replicates, each with $N = 250$ cells

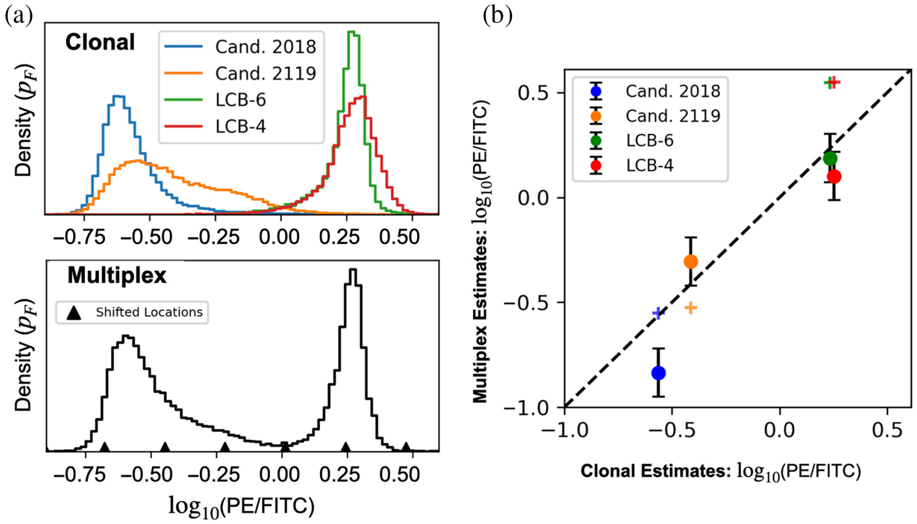


FIGURE 3 Agreement of binding signal of de novo designed binding proteins measured via yeast display in multiplex with ground truth values obtained in clonal yeast. (a) Distributions of samples measured clonally by flow cytometry, and distributions of pooled samples during sorting with a shifting gate boundary. Black triangles represent 6 of the 20 stopping points for the shifted gate. (b) Agreement of clonal and multiplexed binding signal. The x-axis is measured by flow cytometry while the y-axis is a multiplexed measurement by next-generation sequencing. Error bars represent size of the steps used when shifting the threshold

The multiplexed measurements largely recapitulate the ground-truth clonal measurements (Figure 3b), with the exception of design candidate 2018, for which the multiplexed estimate is below the clonal one. We suspect this is due to dissociation of some of the target protein in the time between the clonal and multiplexed measurements; kinetics experiments suggest dissociation occurs rapidly for this design.[1]

3 | CONCLUSION

In the supplementary note, we additionally describe two extensions of this idea. First, because the differences in the variability of fluorescence across each population is often of interest (in addition to mean fluorescence), we show how to extend the approach to estimate the variance for each population and validate this approach in simulation (Figure S4). Second, we describe how to effectively take advantage of sorters that sort into more than two bins simultaneously to obtain more accurate estimates. We view these contributions as a starting point for future work of using randomness to obtain precise, multiplexed estimates.

We have shown how to obtain precise, multiplexed estimates in sort-seq experiments with a simple strategy that incorporates randomness. With as few as two randomized gates, this mathematical technique allows one to collect more accurate data than one could previously obtain with four or six bin sorters. Moreover, this greater accuracy is attained with less sophisticated hardware and less downstream experimental effort. While we have emphasized studies of binding affinity, we believe our strategy is applicable to a wider range of applications of sort-seq assays including studying transcriptional regulation[10,13] and protein stability,[11] and building datasets for protein design.[14] Widespread implementation of randomized gates in FACS and community adoption of this strategy, will greatly simplify and improve sort-seq assays by eliminating a common bias in this ubiquitous assay. We believe this will allow FACS to play a more central role in screening settings, for construction of reliable datasets for machine learning models in bio-design applications, and for building datasets for quantitative models in biology more generally.

4 | MATERIALS AND METHODS

4.1 | Unbiasedness of estimates from randomized gates

The advantage of the randomized gates presented in Equation (1) is that the resulting counts in each bin (Y_1 and Y_2) may be combined as $\widehat{\boldsymbol{\mu}} = (U-L)\cdot Y_2/(Y_1+Y_2)+L$ to estimate $\boldsymbol{\mu}_F$ without bias. We make this statement precise and present a theorem that guarantees when this is the case.

For an estimator $\widehat{\theta}$ of a fixed estimand θ, the estimator's *bias* is the expected value of its error $\mathbb{E}[\widehat{\theta}-\theta|\theta]$ conditioned on that particular value of θ. An estimator is called *unbiased* if, regardless of the value of the estimand, the bias is equal to zero—that is, if $\mathbb{E}[\widehat{\theta}|\theta]=\theta$ for every θ. Theorem 1 states that this property holds for $\widehat{\boldsymbol{\mu}}$.

Theorem 1. *(*Unbiasedness with randomized gates*).* If the support of p_F is bounded between L and U, then $\widehat{\boldsymbol{\mu}}$ is an unbiased estimator of mean fluorescence. That is, $\mathbb{E}[\widehat{\boldsymbol{\mu}}]=\boldsymbol{\mu}_F$.

Proof. We begin by rewriting the probability that a cell is collected into Bin 2 to expose the connection between this quantity and $\boldsymbol{\mu}_F$:

$$
\begin{aligned}
\mathbb{P}(B=2) &= \int_L^U p_F(f)\mathbb{P}(B=2|F=f)df \\
&\quad //\text{via law of total probability support assumption} \\
&= \int_L^U p_F(f)(f-L)/(U-L)df \\
&\quad //\text{by Equation (1)} \\
&= (\boldsymbol{\mu}_F-L)/(U-L).
\end{aligned}
$$

If $N = Y_1+Y_2$ total cells are collected, then the count in the second bin is distributed as $Y_2\,|\,N \sim \text{Binomial}((\boldsymbol{\mu}_F-L)/(U-L),N)$ and has mean $\mathbb{E}[Y_2]=N\cdot(\boldsymbol{\mu}_F-L)\;/(U-L)$. Accordingly, for any N total number of cells, $\mathbb{E}[\widehat{\boldsymbol{\mu}}|Y_1+Y_2=N]=N\cdot(\boldsymbol{\mu}_F-L)/(Y_1+Y_2)+L=\boldsymbol{\mu}_F$. When N is random as well, then by the law of iterated expectation, $\mathbb{E}[\widehat{\boldsymbol{\mu}}]=\mathbb{E}[\mathbb{E}[\widehat{\boldsymbol{\mu}}|Y_1+Y_2=N]]=\boldsymbol{\mu}_F$ as desired.

Notably, this theorem holds for any distribution p_F satisfying the support condition and does not require any parametric assumptions such as log-normality.

4.2 | Trade-off between number of cells sorted and precision of estimates

The relative simplicity of the estimate $\widehat{\boldsymbol{\mu}}$ leads to a transparent trade-off between the precision and scale of the experiment. Recalling that $Y_2\,|\,N\sim\text{Binomial}((\boldsymbol{\mu}_F-L)/(U-L),N)$, the variance of $\widehat{\boldsymbol{\mu}}$ is

$$
\text{Var}[\widehat{\boldsymbol{\mu}}|N]=\frac{(U-L)^2}{N^2}\text{Var}[Y_2]=\frac{(U-L)^2}{N}\mathbb{P}(B=1)\mathbb{P}(B=2).
$$

To construct a confidence interval for $\boldsymbol{\mu}_F$, we can therefore first approximate the standard error of $\widehat{\boldsymbol{\mu}}$ by $\frac{U-L}{\sqrt{N}}\frac{\sqrt{Y_1Y_2}}{N}$, and appeal to approximate normality of the Binomial distribution for moderate to large N to report $\boldsymbol{\mu}_F=\widehat{\boldsymbol{\mu}}\pm2\frac{U-L}{\sqrt{N}}\frac{\sqrt{Y_1Y_2}}{N}$ with 95% confidence. Because $\sqrt{Y_1Y_2}/N$ can be at most 1/2 (if $Y_1 = Y_2$), the size of this interval is at most $2(U-L)/\sqrt{N}$. Therefore, to estimate $\boldsymbol{\mu}_F$ to within one tenth of the range with high confidence, at most $N = 400$ cells are needed, since in this case $2(U-L)/\sqrt{N}=(U-L)/10$.

For scale, commercial machines sort on the order of 10,000 cells per second, and typical assays sort tens of millions of cells divided amongst many populations. Thus, a library of 100,000 populations could be screened to high precision with on the order of 1 hr of sorting time.

4.3 | Simulation details

In the simulations depicted in Figure 2, we compare against the standard approach of using a histogram to estimate $\boldsymbol{\mu}_F$. Consider a K bin histogram. For each bin k, if the range of fluorescences collected is from lower bound l_k to upper bound u_k, then $\mathbb{P}(B=k|F=f)=\mathbf{1}[l_k \le f < u_k]$. The histogram estimate then corresponds to combining the resulting counts as

$$\widehat{\boldsymbol{\mu}}_{\text{Hist}}=\sum_{k=1}^{K}\frac{Y_k}{N}\left(\frac{u_k+l_k}{2}\right).$$

In order to use the unbiased estimator, both in simulation and in practice, we must slightly extend the randomized gate definition proposed in Equation (1). In particular, Theorem 1 assumes that the support of the fluorescence density p_F is bounded between L and U (i.e., that for $F \sim p_F, \mathbb{P}[L \le F \le U]=1$). In practice, this may not be the case. But, as previously stated, Equation (1) returns negative "probabilities" outside of this range. Therefore, we propose to "clip" the collection probabilities at the boundaries, and instead define

$$\mathbb{P}(B=1|F=f)=\left(1-\tfrac{f-L}{U-L}\right)_{\dagger} \text{ and}$$
$$\mathbb{P}(B=2|F=f)=\left(\tfrac{f-L}{U-L}\right)_{\dagger}$$

where † denotes clipping between zero and one such that, for a scalar x, $(x)_{\dagger}=\max(\min(x,1),0)$. This ensures that $\widehat{\boldsymbol{\mu}}$ is well-defined, but gives up unbiasedness in situations where the support assumption of Theorem 1 is violated. This bias is apparent, for example, at the right and left sides of the left panel of Figure 2c.

4.4 | Experimental approximation of randomized gates with shifting thresholds

Because current FACS software does not support randomized gate programming, we devised an experimental approximation in which we manually changed the gating threshold 20 times during sorting at regular intervals. Specifically, we use a gate that collects all cells with fluorescence above a threshold into Bin 2 and those below the threshold into Bin 1, and we shift that threshold over the course of the collection from the lower limit L to the upper limit U. In theory, this approach exactly recovers Equation (1) in the limit that the threshold is shifted continuously from L to U at a constant rate. This is because for a cell with fluorescence f between L and U, the probability that it is collected into Bin 2 is the fraction of the experimental time during which the threshold is below f, which is $(f-L)/(U-L)$. This approximation does not, however, account for possible changes in the distribution, p_F over time. Such changes occur in binding assays, for example, when nontrivial labeled target protein dissociates over time. This challenge is a disadvantage of the approximation relative to randomized gates that could in theory be implemented into sorters.

4.5 | Yeast display and deep sequencing

EBY100 yeast cells expressing each of the four mini-protein binders were grown in C-Trp-Ura media. Binder protein expression was induced by replacing the growing buffer with SGCAA and incubating at 30°C for 24 hr.[15] The induced cells were labelled with 250 nM biotinylated receptor binding domain target protein, washed twice with PBSF (PBS + 1% BSA), then labelled again with anti-c-Myc fluorescein isothiocyanate (FITC) and streptavidin-phycoerythrin (SAPE). The experiments were performed on a Sony SH800 cell sorter. Sixty thousand cells were recorded for each binder to reflect the individual distribution of baseline PE signal intensity. In the shifting gate experiment, a square area (AreaTotal) with side length (L) was predetermined at the SH800 collection panel. The area was divided into 2 separate collection gates, Gate1 and Gate2 (corresponding to Bin 1 and Bin 2 in Equation (1)). Gate2 was in an isosceles right triangle and started with a small area in the right-bottom corner of AreaTotal and Gate1 took up the remaining. The yeast cells were run through the SH800 and each cell went into either the Gate2 or Gate1 collection tube if its log PE/FITC signal was in the range of AreaTotal. All other cells were discarded. After collecting 50,000 cells, the cell flow was paused, Gate2 was shifted both leftwards and upwards for $L/10$ and cell flow continued. Because the proprietary software for operating the sorter allowed setting gate positions only through a point and click graphical user interface (rather than numerically), we measured out gate increments by pixel distance on the display using a ruler. The above shifting process repeated 19 times for a total of 20 collections. The cells collected in Gate1 and Gate2 were then grown, and 1×10^7 cells from each gate were barcoded and the sequences for each cell were determined by Illumina next-generation sequencing.[11] The number of cells collected by each gate for each population was estimated from the proportion of sequencing reads attributed to each population and the number of cells collected into the gates.

Because the number of cells collected by each gate was not made directly available through the proprietary software, we estimated this from the raw exported data.

In particular, we imported the data using the FlowCal python package[16] and computationally implemented the gates and filters (including for forward and backward scatter).

4.6 | Sensitivity of maximum likelihood inference to non-normality of real data

Likelihood-based inference is a common strategy used with the intent to circumvent the resolution limitation of the histogram approach.[2,9–11] However, this approach can fail on real data. In particular, existing likelihood methods rely on the assumption that for each of the cell populations the fluorescence values are log normal distributed, $\log F \sim \mathcal{N}(\mu, \sigma^2)$ where the mean log fluorescence $\mu = \boldsymbol{\mu}_F$ is the target of inference and σ^2 is the typically unknown variance of the population.

We evaluate performance of maximum likelihood inference in this situation with simulations using data sub-sampled from a flow cytometry dataset of binding signal of a computationally designed mini-protein binder to ActRII. Data were collected using yeast display as previously described except with the addition of a supplemental binding protein, protein A, the binding signal $\log(\text{FITC}/\text{PE})$ was recorded for approximately 10,00,000 cells. The distribution of this signal is highly non-Gaussian (Figure S1A).

We first compared the performance of the maximum likelihood approach (described in greater detail below) to the randomized approach on downsampled datasets with $N = 250$ cells with the same set-up described in Figure 2. As in the earlier simulations, the randomized approach provides improved MSE across most simulation conditions (Figure S1B). This improvement is again explained by estimation bias, which is mitigated by the randomized approach (Figure S1C). Though one might expect the benefit of maximum likelihood would appear for larger sample sizes (e.g., due to the asymptotic efficiency of maximum likelihood estimation in theory), this is not the case. In fact, due to the bias of maximum likelihood, the relative improvement of the randomized approach is larger at $N = 1{,}000$ cells (Figure S1D). Moreover, Figure S1E demonstrates that the maximum likelihood approach does not empirically provide more accurate estimates even under correct specification (with fluorescences sampled as in Figure 2a).

4.7 | Maximum likelihood estimation

To estimate $\boldsymbol{\mu}_F$, likelihood-based approaches consider the counts in each of K bins (Y_1, Y_2, ..., Y_K), since the measured fluorescence values cannot be disambiguated when multiple populations are sorted in multiplex. These counts follow a multinomial distribution as

$$Y_1, Y_2, \ldots, Y_K \sim \text{Mult}\left(\boldsymbol{\pi}\left(\mu, \sigma^2\right), N\right),$$

where $N = \sum_{k=1}^{K} Y_k$ is the total number of cells sorted into any bin and $\boldsymbol{\pi}(\mu, \sigma^2) = (\pi_1, \pi_2, \ldots, \pi_K)$ are the normalized bin probabilities. In particular, if for each bin k the range of fluorescences collected is from lower bound l_k to upper bound u_k, then

$$\pi_k = \frac{\Phi\left(\frac{u_k - \mu}{\sigma}\right) - \Phi\left(\frac{l_k - \mu}{\sigma}\right)}{\sum_{k'=1}^{K} \Phi\left(\frac{u_{k'} - \mu}{\sigma}\right) - \Phi\left(\frac{l_{k'} - \mu}{\sigma}\right)},$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal. The log likelihood function is then

$$\log\, p\left(Y_1, \ldots, Y_K; \mu, \sigma^2\right) = \log N! - \sum_{k=1}^{K} \log Y_k! + \sum_{k=1}^{K} Y_k \log \pi_k,$$

where the dependence of each π_k on μ and σ^2 is left implicit. The maximum likelihood approach is to return μ that maximizes this expression,

$$\widehat{\boldsymbol{\mu}}_{\text{MLE}} = \arg\max_{\mu} \left[\max_{\sigma^2 > 0} \log p\left(Y_1, \ldots, Y_K; \mu, \sigma^2\right)\right].$$

This optimization problem is not analytically tractable, and its constraints and non-convexity pose challenges for local, gradient-based optimizers. So we instead solve the optimization approximately with a grid search.

AUTHOR CONTRIBUTIONS

Brian L. Trippe: Conceptualization (equal); formal analysis (equal); funding acquisition (equal); investigation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Buwei Huang:** Data curation (equal); formal analysis (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Erika A. DeBenedictis:** Data curation (equal); formal analysis (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Brian Coventry:** Data curation (equal); formal analysis (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Nicholas Bhattacharya:** Formal analysis (equal); investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Kevin K. Yang:** Conceptualization (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); software (equal); supervision (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **David Baker:** Data curation (equal); investigation (equal); project administration (equal); resources (equal); supervision (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Lorin Crawford:** Conceptualization (equal); funding acquisition (equal); investigation (equal); methodology (equal); project administration (equal); supervision (lead); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

We would like to thank Sarah Kate Nyquist for helpful conversations and suggestions. Brian L. Trippe would like to acknowledge support from the National Science Foundation Graduate Research Program. Lorin Crawford is supported by a David & Lucile Packard Fellowship for Science and Engineering. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

ORCID

David Baker https://orcid.org/0000-0001-7896-6217
Lorin Crawford https://orcid.org/0000-0003-0178-8242

REFERENCES

1. Cao L, Goreshnik I, Coventry B, et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science. 2020;370(6515):426–431.
2. Cao L, Coventry B, Goreshnik I, et al. Design of protein-binding proteins from the target structure alone. Nature. 2022;605:551–560.
3. Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci. 2010;107(20):9158–9163.
4. Starr TN, Greaney AJ, Hilton SK, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182(5):1295–1310.
5. Adams RM, Mora T, Walczak AM, Kinney JB. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. Elife. 2016;5:e23156.
6. Reich L, Dutta S, Keating AE. SORTCERY—A high-throughput method to affinity rank peptide ligands. J Mol Biol. 2015;427(11):2135–2150.
7. Peterman N, Levine E. Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics. 2016;17(1):1–17.
8. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. New York: Springer, 2001.
9. de Boer CG, Ray JP, Hacohen N, Regev A. MAUDE: Inferring expression changes in sorting-based CRISPR screens. Genome Biol. 2020;21:1–16.
10. Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. Nat Genet. 2019;51(12):1664–1669.
11. Rocklin GJ, Chidyausiku TM, Goreshnik I, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017;357(6347):168–175.
12. Boder ET, Dane Wittrup K. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol. 1997;15(6):553–557.
13. Sharon E, Kalma Y, Sharp A, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012;30(6):521–530.
14. Biswas S, Kuznetsov G, Ogden PJ, Conway NJ, Adams RP, Church GM. Toward machine-guided design of proteins. bioRxiv. 2018;337154.
15. Chevalier A, Silva D-A, Rocklin GJ, et al. Massively parallel de novo protein design for targeted therapeutics. Nature. 2017;550(7674):74–79.
16. Castillo-Hair SM, Sexton JT, Landry BP, Olson EJ, Igoshin OA, Tabor JJ. Flowcal: A user-friendly, open source software tool for automatically converting flow cytometry data from arbitrary to calibrated units. ACS Synth Biol. 2016;5(7):774–780.
17. Luenberger DG. Optimization by vector space methods. New York: John Wiley & Sons, 1969.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Trippe BL, Huang B, DeBenedictis EA, Coventry B, Bhattacharya N, Yang KK, et al. Randomized gates eliminate bias in sort-seq assays. Protein Science. 2022;31(9):e4401. https://doi.org/10.1002/pro.4401