Single-cell Bayesian deconvolution

Gabriel Torregrosa,[1] David Oriola,[2] Vikas Trivedi,[2, 3] and Jordi Garcia-Ojalvo[1]
[1] *Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, 08003 Barcelona, Spain*
[2] *EMBL Barcelona, Dr. Aiguader 88, 08003 Barcelona, Spain*
[3] *EMBL Heidelberg, Developmental Biology Unit, 69117 Heidelberg, Germany*

Flow cytometry enables monitoring protein abundance and activity at the single-cell level in a high-throughput manner, through the use of fluorescent labeling. Given the significant levels of autofluorescence emitted by cells at the spectral ranges used by this technique, removing the corresponding background signal is necessary for a correct assessment of cellular biochemistry. Existing methods of autofluorescence removal usually require dedicated monitoring resources, such as additional fluorescence channels or laser sources, which are costly and not universally accessible. Here, we have developed a computational method that enables autofluorescence subtraction without requiring dedicated measurement resources. The method uses a non-parametric Bayesian approach to deconvolve the target signal distribution from independent measurements of labeled and unlabeled cells readily available in a typical experiment. The distributions are approximated by mixtures of gamma functions, and the target distribution is obtained by sampling the posterior distribution using Markov chain Monte Carlo and nested sampling approaches. We tested the method systematically using synthetic data, and validated it using experimental data from mouse embryonic stem cells.

I. INTRODUCTION

The inherent stochasticity of biological processes leads to substantial heterogeneity even among genetically identical cells in the same environment[1–3]. The degree to which this heterogeneity affects, or even dictates, cellular decision making in most situations is still an open question. This issue is of paramount importance in processes such as mammalian development, where hundreds (if not thousands[4]) of distinct cell types (cell states) emerge from a small number of identical undifferentiated cells[5–7]. Identifying the molecular mechanisms underlying these cell-fate decision programs and their interplay with cellular heterogeneity[8,9] requires a rigorous quantification of cellular states across large numbers of cells.

Flow cytometry enables monitoring the distributions of abundances and activities of selected proteins for thousands of cells at a time, using fluorescently labeled markers. Cells, however, have a non-negligible amount of autofluorescence in the emission spectrum of most fluorescent probes (500 to 700 nm). This leads to a background noise that must be subtracted from the total signal emitted by fluorescently labeled cells[10], in order to adequately relate the signal distribution provided by the cytometer to the mechanisms regulating the expression and/or activity of the protein of interest. Several standard methods exist for addressing this issue on a cell-by-cell basis, but they require the use of dedicated measurement resources, such as an additional fluorescence channel outside the emission spectrum of the fluorophore being used, to directly measure autofluorescence[11,12], or a second laser system, to provide an independent measurement of both signal and autofluorescence[13]. The effectivity of such solutions is limited, beyond cost or accessibility issues, because there is no guarantee that autofluorescence from another channel, or from another excitation source, is a good proxy for autofluorescence in our channel of interest.

Commonly, rather than dedicating measurement resources to assess autofluorescence, control measurements of unlabeled cells are used to set a baseline of the signal coming from the naturally present autofluorescent components in the cell. This background distribution provides information about the resolution achievable by the technique. This procedure, however, does not lead to a quantitative determination of the distribution of the signal coming *exclusively* from the fluorescent probe. Such quantitative assessment would require deconvolving the fluorescence distribution obtained in labeled cells from the one produced by unlabeled cells. Here we propose a non-parametric Bayesian approach to this deconvolution problem, applicable to one-dimensional measurements. The method is robust and efficient, requiring cell numbers not larger than those typically considered in standard flow cytometry runs, and gives natural confidence intervals of the target distributions, which makes it attractive for a variety of applications.

There is an extensive statistics literature addressing the additive deconvolution problem[14]. A common set of deconvolution methods are kernel-based approaches, such as those relying on Fourier transforms[15–21], which use the fact that in Fourier space, a deconvolution is simply the product of two functions. Two problems arise from such methods that limit their applicability in practical cases. First, Fourier transforms (and other methods that use orthogonal local basis such as wavelets[22]) are not positive defined. Consequently, kernel-based methods lead to deconvolved pseudo-distributions with artificial features, which are hardly interpretable for practical applications. Second, these methods usually lead to point estimates, and therefore do not provide native confidence intervals (i.e. without applying additional statistical approximations) that allow us to assess the quality of the inferred target distribution.

A second class of deconvolution approaches are likelihood-based methods[23], which estimate the unknown target distribution using maximum likelihood approaches. As in the case of kernel-based methods, these approaches provide us with point estimates, and usually assume exact knowledge of the noise distribution. Finally, a third class of methods involve Bayesian inference[24–26], which does not require complete knowledge of the noise distribution and naturally provides confidence intervals of the estimates obtained. So far, however, these Bayesian methods have been applied to repeated measurements of the same individual entities that are being monitored (in our case, cells), which is not a realistic possibility in standard flow cytometry. Our semi-parametric Bayesian approach does not require repeated measurements and retains all the above-mentioned advantages of Bayesian methods. We have implemented the procedure in a Python package available in GitHub (https://github.com/dsb-lab/scBayesDeconv).

The work is structured as follows. First, we introduce the Bayesian approach and discuss sampling methods for exploring the posterior distribution generated from the model. Second, we validate our method using synthetic datasets with known target distributions, and compare its results to other existing methods. Third, we further test our method in real flow-cytometry data of mouse embryonic stem cells undergoing differentiation. To that end, we begin by artificially convolving this data with an *ad hoc* noise distribution, to check the robustness of the deconvolved distributions. We then treat our cells with a low concentration of a fluorescent dye (which masks the real flow-cytometry signal and acts as an external noise), and validate our method by deconvolving the noise coming from the dye and comparing it to the control case were the dye was not added. We conclude by discussing the limitations of the method and possible ways to improve it in future work.

II. MATERIALS AND METHODS

A. Theoretical definition of the problem

Consider a population of cells containing a fluorescent marker that labels the abundance or activity (e.g. phosphorylation state) of a protein of interest. Flow cytometry measurements provide us with the distribution $p_c(C)$ of total fluorescence signal C emitted by each individual cell in the population. This signal has two components: the fluorescence T emitted exclusively by the target fluorophore that reports on the protein of interest, and the autofluorescence ξ emitted by cellular components *other* than our fluorescent label:

$$C = T + \xi \qquad (1)$$

If these two components are independent of one another, the distribution $p_c(C)$ of the measured signal takes the form of a *convolution* of the distributions of T and ξ:

$$(p_T * p_\xi)(C) := \int_0^\infty p_T(C-\xi)\, p_\xi(\xi)\, d\xi \qquad (2)$$

Similarly to p_c, the distribution p_ξ of the autofluorescence ξ can be measured in the flow cytometer by using unlabeled cells that are otherwise identical to the labeled ones. On the other hand, the probability distribution of T, p_T, cannot be measured directly. Our goal is to extract (deconvolve) the distribution p_T from the measured distributions p_c and p_ξ, considering that we only have a finite set of samples (cells) of C and ξ.

In what follows, we first introduce the way in which we describe the distributions involved in the problem. Next we define the posterior distribution given by the model and the data, and finally we discuss the methods used to explore the parameter space for the deconvolution problem.

1. Mixture model

The vast majority of real datasets (including those generated in flow cytometry) result from complex combinations of variables that cannot be explained in general with simple distributions. In order to adapt flexibly to such conditions, we use mixtures of probability distributions as our basis set. Any function can be arbitrarily well approximated using an adequate choice of basis functions, provided enough components are included in the mixture. We can describe both the target and the noise distributions by independent mixtures of K components:

$$p(x|\boldsymbol{\phi}) = \sum_{\eta=1}^{K} \omega_\eta \mathcal{B}(x|\psi_\eta) \qquad (3)$$

where ω_η denotes the weight of each base $\mathcal{B}(x|\psi_\eta)$ in the mixture and ψ_η represents its parameters (e.g. the means and standard deviations in the case of normal distributions)[27]. The sets of ω_η and ψ_η are in turn represented by the vector $\boldsymbol{\phi}$. Under this description, the distribution of the observed signal C can be described as a superposition of all the convolutions between basis functions:

$$p_c(c|\boldsymbol{\phi}_T, \boldsymbol{\phi}_\xi) = \sum_{\eta=1}^{K_T} \sum_{\lambda=1}^{K_\xi} \omega_\eta^T \omega_\lambda^\xi (\mathcal{B} * \mathcal{B})(c|\psi_\eta^T, \psi_\lambda^\xi) \qquad (4)$$

where $(\mathcal{B}*\mathcal{B})(c|\psi_\eta^T, \psi_\lambda^\xi)$ represents the convolution of two basis distributions with parameters ψ_η^T and ψ_λ^ξ, respectively, and K_T and K_ξ denote the number of bases used in each of the two mixtures.

The choice of basis functions to describe our data is crucial. We propose two types of basis functions: normal and gamma distributions. Usually, normal distributions with unknown mean and variance have been shown to be flexible enough to represent datasets with high quality, requiring less components than more rigid methods[28]. On the other hand, gamma distributions can be more realistic when representing protein abundances[29]. Gamma distributions are thus a more natural choice to describe flow cytometry data, and might be able to capture the data with less components than normal distributions.

2. Posterior distribution and likelihood

Our goal is to extract, starting from samples of the distributions of total signal C and noise ξ, the parameters of the distribution of the target signal T. Defining the problem in terms of probability distributions leads very naturally to work with Bayesian methods. According to Bayes' rule, the posterior distribution that represents the probability of the parameters given the data is

$$p(\boldsymbol{\phi}_T, \boldsymbol{\phi}_\xi | \mathbf{c}, \boldsymbol{\xi}) \propto \Big[p(\mathbf{c}|\boldsymbol{\phi}_T, \boldsymbol{\phi}_\xi) p(\boldsymbol{\xi}|\boldsymbol{\phi}_\xi) \Big] \Big[p(\boldsymbol{\phi}_T) p(\boldsymbol{\phi}_\xi) \Big], \tag{5}$$

where $\mathbf{c} = \{c_i : i = 1 \dots N_c\}$ and $\boldsymbol{\xi} = \{\xi_i : i = 1 \dots N_\xi\}$ are the sets of observed samples of the total signal and the noise, respectively, with N_c and N_ξ representing the number of samples in each case.

The first bracket on the right-hand side of Eq. (5) is the likelihood function, which corresponds to the probability of the data given the parameters. Under the assumption of independent and identically distributed (iid) observations, this function is given by

$$\mathcal{L}(\boldsymbol{\phi}_T, \boldsymbol{\phi}_\xi) = \prod_{i=1}^{N_c} p_c(c_i | \boldsymbol{\phi}_T, \boldsymbol{\phi}_\xi) \prod_{j=1}^{N_\xi} p_\xi(\xi_j | \boldsymbol{\phi}_\xi) \tag{6}$$

As can be seen, the information about the noise parameters is contained in both datasets, while the information about the target parameters only appears in the convolved data.

As shown in Eq. (5) above, the posterior distribution can be estimated by multiplying the likelihood by the prior distribution of parameter values. Since the datasets required for a good deconvolution are large, the impact of the prior distribution should be negligible. Therefore the posterior landscape is effectively described by (6), and the prior distribution is only present in our approach for formal and computational reasons. In any case, since no prior information exists on the parameters that describe the noise and target mixture components, we impose vague priors over the plausible set of parameters based on sampling efficiency (see Supplementary Sec. S3).

3. Sampling methods

The posterior distribution is a complex multimodal (multi-peaked) object containing all the configurations of the target and noise distributions that are consistent with the observed data. A correct exploration of this distribution is essential to set bounds on the candidate models that explain the data. Our method relies on two sampling techniques that we introduce in what follows: Markov Chain Monte Carlo (MCMC)[28] and nested sampling[30].

MCMC sampling is a computational approach commonly used in Bayesian inference[28]. MCMC methods are fast for simple unimodal cases, but they have severe difficulties exploring complex multimodal distributions that contain multiple, possibly degenerate, peaks with deep valleys between them. This may be the case when the noise heavily dominates over the signal. Therefore, in the context of deconvolution, MCMC methods can only be optimal when deconvolving small amounts of noise.

Flow cytometry distributions usually require intensive exploration of parameter space. Nested sampling (NS) was developed as a method method for evidence sampling that gives posterior samples as a by-product[30–32]. Recent improvements in the NS approach have enabled an efficient and intensive exploration of high-dimensional and complex objects with degeneracies and multimodalities[33], as the ones that may be found in a deconvolution problem. The downside of this method is that it is generally more computationally expensive than MCMC methods even for simple cases, scaling with the size of the prior-distribution parameter space.

4. Analysis pipeline

Given the concepts and tools described above, the analysis of the data is performed as follows (Fig. 1). First, the distributions of the target and autofluorescence signals are assumed to be described by mixtures (top left panel in Fig. 1), as defined by Eqs. (3) and (4). The data measured experimentally correspond to the autofluorescence signal (noise) and the total signal of the labeled cells (which includes the autofluorescence), as shown in the top right panel of Fig. 1. The mixture assumption and the observed data samples allow us to construct the posterior distribution (Eq. (5) and middle panel in Fig. 1) from the likelihood function defined by Eq. (6) and the prior distributions discussed in the Supplementary Sec. S3. The posterior distribution is a function of the model parameters (weights of the mixtures and parameters of the basis functions). We next explore (sample) the posterior distribution in parameter space using Markov Chain Monte Carlo or Nested Sampling methods (bottom left panels in Fig. 1). These algorithms provide us with a representative sampling of the parameters of the target distribution, which allows us to compute an average of this distribution and its confidence interval (bottom right panel in the figure).

B. Materials

1. Synthetic data

For the target distribution we generated samples from three distributions: symmetric bimodal, asymmetric bimodal and skew symmetric distributions (Supplementary Fig. S1). This choice of distributions intends to capture

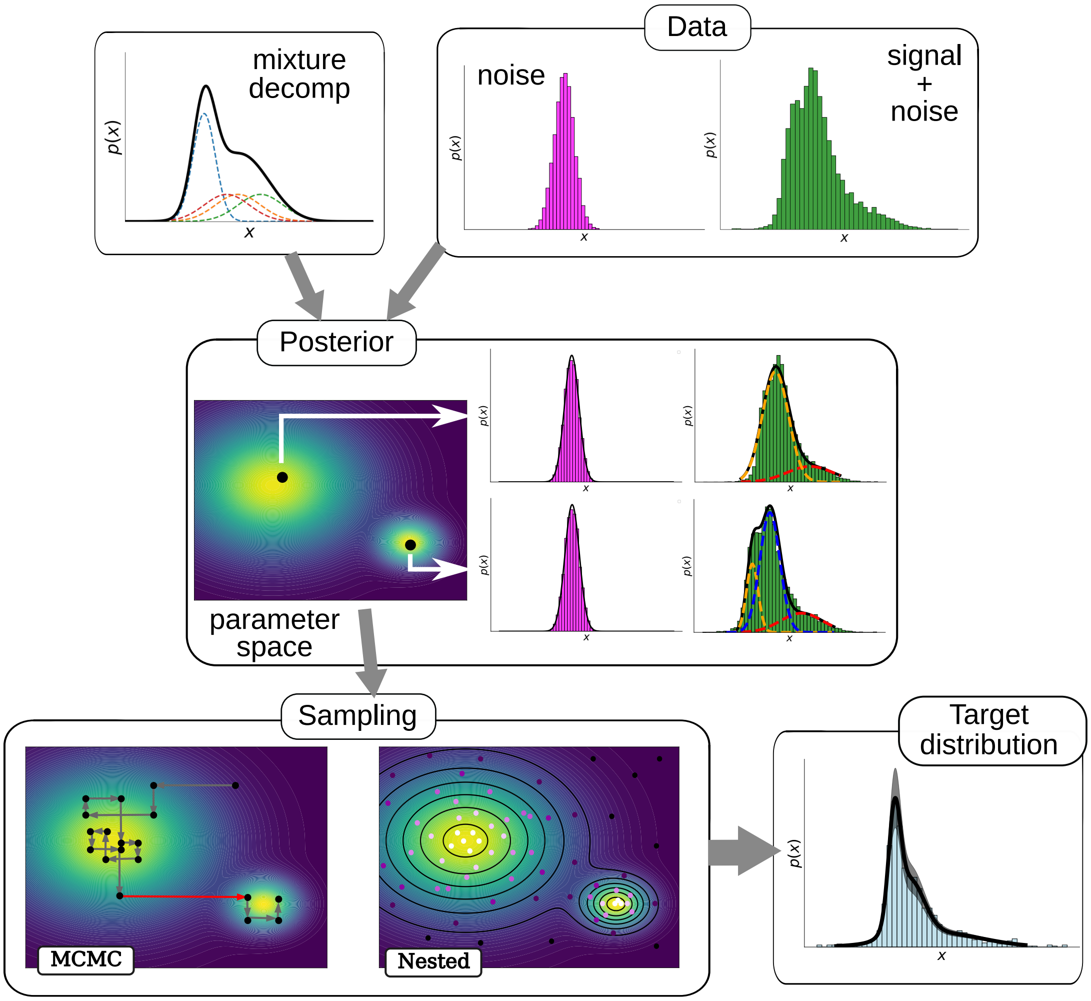

FIG. 1. Scheme of the deconvolution process. The signal and noise mixture distributions, together with the observed data (top row), define the posterior distribution over the parameter space of the mixture, Eq. (6) (middle row). This distribution can present multiple peaks, sometimes degenerate with respect to basis label exchange, each corresponding to a different mixture description of the observed and target distributions (two examples are shown in the right-hand-side of the middle row) The red arrow in the MCMC sampling plot (bottom left) represents a very unlikely jump between two peaks separated by a relatively wide probability valley.

the features present in real datasets, such as the presence of multiple peaks, different cluster sizes, and the generally non-Gaussian character of the data[25]. As for the noise, we generated a set of nine different noise datasets containing multiple peaks, skewness, and fat tails (Supplementary Fig. S1), in order to test the flexibility of the method against very dissimilar autofluorescence profiles. In order to check the impact of the noise strength, the convolutions between target and noise were generated at two signal-to-noise ratios (SNR). In our context, we define the SNR as the ratio between target and signal variances for the whole dataset. We chose a case with negligible noise (SNR = 10), and a difficult case were the noise is of the same magnitude as the signal (SNR = 1). We generated sample datasets of different sizes, with 100, 1000 and 10000 samples. The high range of these values is representative of typical single-cell flow cytometry experiments. The combination of the different target distribution types, noise distribution types, SNRs, and sample sizes generates a collection of 162 datasets with known ground truth.

2. *Flow cytometry*

E14 mouse embryonic stem cells containing a knock-in fluorescence reporter for the mesodermal transcription factor Brachyury, T/Bra::GFP were used[34]. Cells containing the T/Bra::GFP reporter were cultured in ES-Lif (ESL) medium (KnockOut Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1x Non-essential aminoacids (NEEA), 50 U/mL Pen/Strep, 1x GlutaMax, 1x Sodium Pyruvate, 50 μM 2-Mercaptoethanol and leukemia inhibitory factor (LIF)). Cells adhered to 0.1% gelatin-coated (Millipore, ES-006-B) tissue culture-treated 25 cm^2 (T25 Corning 353108) plates, and were passaged every second day, as previously described[35]. Cells were kept at 37°C with 5% CO$_2$, and were routinely tested and confirmed to be free from mycoplasma. Flow cytometry experiments were performed as follows: On day 1, cells were trypsinized and seeded into gelatin-coated 6-well plates (Corning, 353224) to a final density of $\sim 10^5$ cells/well in 3 mL ESL media. On day 2, the media was replaced by first washing twice with 3 mL DPBS$^{+/+}$ (Phosphate buffered saline containing Mg^{++} and Ca^{++}, Sigma, D8662) and then adding NDiff227 media (N2B27) (Takara Bio, #Y40002)[35] with the appropriate combination of Brachyury activators (3 μM CHIR99, Sigma, SML1046 and/or 25 ng/mL Activin A, Bio-Techne, 338-AC-010). The same procedure was followed on day 3. For the control conditions, the corresponding volume of dimethyl sulfoxide (DMSO) was added to the medium. Flow cytometry data was acquired on day 2 or day 3, depending on the type of experiment. For the extrinsic noise experiment, cells were incubated with 1 mL of 20 nM CellTracker Green (CMFDA, Thermofisher) for 3 min prior to trypsinization and flow cytometry. The data was acquired using an LSRIIb flow cytometer, with 2×10^4 cells being analyzed per condition. DAPI labelling was used to discard death cells and debris. Furthermore, cell doublets were discarded from the analysis. The readout of protein expression was obtained through the FITC-A channel while the PerCP-Cy5-5-A was used as a known noise source in the analysis. Measurements were extracted using the FACSDiva software and exported in a Python format for subsequent analysis.

C. Software

The software pipeline presented here, including MCMC and NS samplers for normal and gamma mixtures, has been implemented as a Python package (`scBayesDeconv`). The source code, with manual compilation and instalation instructions, as well as full documentation and a notebook with examples of use, can be obtained publicly from Github (https://github.com/dsb-lab/scBayesDeconv). The software is also available at PYPI and can be installed through the pip command. For the NS sampling method, the models are wrapped around the package Dynesty[33].

III. RESULTS

A. Synthetic data

First, we benchmark the ability of our method to recover the target distribution by using collections of synthetic datasets. In those datasets the target and noise distributions are known, so we can compare the result of the deconvolution against a ground truth. To do this, we applied our method to the synthetic data described in Sec. II B 1. For this test we employed normal distributions as basis functions, to avoid favoring our method over FFT approaches. Specifically, normal distributions are defined in the entire real axis and are not heavily skewed, which would be problematic for FFT-based methods[17]. To prove the robustness of the implementation, we ran the test using five components both for the noise and target distributions in the case of the MCMC algorithm, and three components in the case of the Nested Sampling algorithm (which is more computationally demanding). We also avoided checking the full convergence of the algorithm manually. We ran the algorithm in this sub-optimal conditions to avoid having to fine tune the specific parameters, which could lead to positive bias favoring our method in comparison with FFT-based approaches. We contrasted the results of our method with those of a specific FFT approach that does not require knowledge of the autofluorescence distribution[19]. Figure 2 shows a typical instance of the deconvolution performance of the two methods. Panel (a) shows in green the total (convolved) signal mimicking the output of a flow cytometry experiment. In this synthetic case, the signal is obtained by forward convolving a target distribution with the characteristics given above, shown in light blue in panels (b) and (c), with a noise (autofluorescence) distribution, shown in magenta in the inset of panel (a). The goal in this case is to recover (deconvolve) the ground-truth target distribution (panels b and c in Fig. 2) from the total and noise distributions (panel a). Figures 2(b,c) show that our Bayesian deconvolution method recovers reasonably well the target distribution as compared with the FFT-based method in this case. In particular, working in Fourier space leads to oscillatory components in the deconvolution, and correspondingly to artifactual negative values in the probability distribution. Additionally, the Bayesian deconvolution methods provides naturally a confidence interval, shown by the orange-shaded region in panel (b) of Fig. 2.

For benchmarking purposes, we compared the deconvolved distribution with the real target distribution, which is known in this case, using the mean integrated overlap (MIO), as defined in Supplementary Sec. S6. We preferred this measure to the mean integrated squared error (MISE), which is commonly used in the deconvolution literature for theoretical reasons[16,19,20,22], since the MIO measure is easier to interpret, as it corresponds directly to the absolute overlap of two probability distributions. We also avoid more common measures of distribution dissimilarity such as the Kolmogorov-Smirnov test, since such methods would underestimate the ability of FFT-based methods to converge to the ground-truth deconvolution, given that they lead to artifacts in the resulting distributions, as shown above.

Figure 3 compares the deconvolution efficiency of the FFT-based method and our single-cell Bayesian Deconvolution approach with MCMC sampling, in terms of the

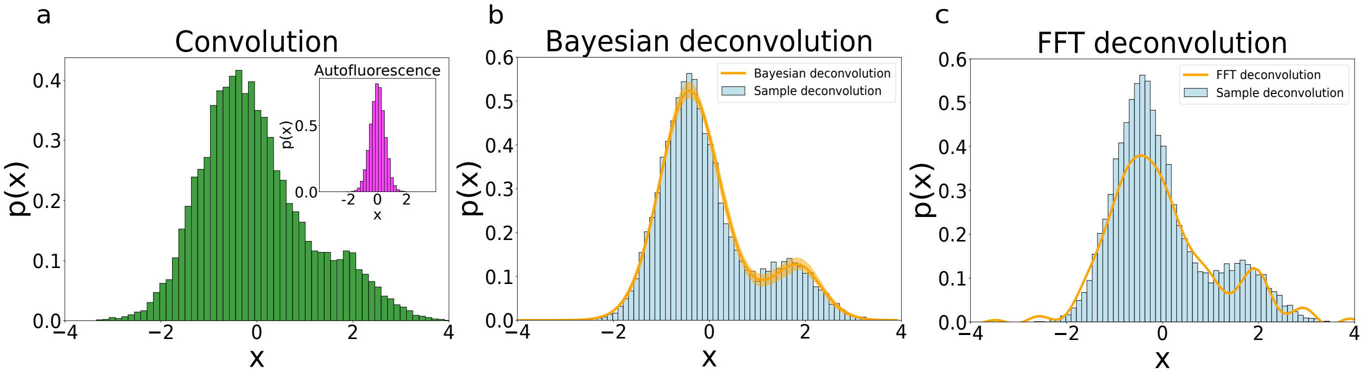

FIG. 2. Deconvolution instance for a target bimodal distribution (light blue in panels b and c) corrupted by a normally distributed noise with a SNS = 2 (magenta distribution in the inset of panel a), applying the Bayesian Deconvolution method (panel b) and the FFT method (panel c). The total distribution from which the noise is deconvolved is shown in green in panel (a). The orange-shadowed region in panel (b) depicts the confidence interval provided by the Bayesian deconvolution method. In this case, the noise distribution is a single normal function with mean $\mu^{\xi} = 0$ and standard deviation $\sigma^{\xi} = 0.5$, and the target is a mixture of two normal functions with means $\mu_1^T = -0.43$ and $\mu_2^T = 1.67$, standard deviations $\sigma_1^{\xi} = \sigma_2^{\xi} = 0.6$, and weights $\omega_1^T = 0.8$ and $\omega_2^T = 0.2$.

MIO measure that quantifies the similarity between the convolved distribution and the real one. A similar result is found for Nested Sampling (see Supplementary Fig. S2). According to its definition (see Supplementary Sec. S6), a MIO value of 1 corresponds to a perfect overlap, while the measure is 0 when two distributions do not overlap at all. As can be seen in Fig. 3, the single-cell Bayesian Deconvolution method outperforms the Fourier-based method in almost all the cases considered, the difference being more substantial for high levels of noise (small SNR, circles). In particular, the overlap is never below 0.7 for the Bayesian methods, while it can reach values near 0 in the FFT case depending on the type of distributions involved, particularly for low sampling numbers.

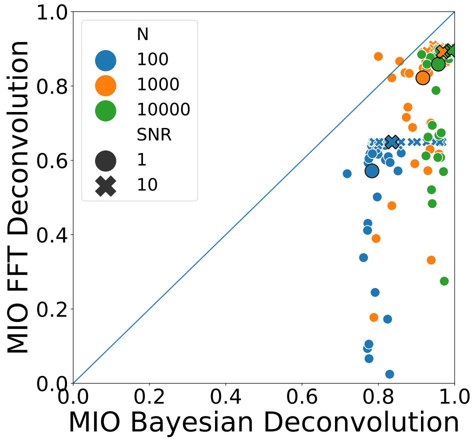

FIG. 3. Similarity between the deconvolved and ground-truth target distributions as expressed by the Mean Integrated Overlap (MIO) for the two deconvolution methods (x and y axis), at different sampling sizes for the synthetic datasets described in described in Sec. II B 1, with SNR = 1 (circles) and SNR = 10 (crosses). The large symbols represent the mean MIO over the different samples for each sample size and SNR.

Additionally, it is worth noting that while the single-cell Bayesian deconvolution method is able to reproduce almost perfectly the target distribution (MIO close to 1) for large enough dataset sizes, the Fourier-based method always saturates to a suboptimal level even as the number of samples increases, even for low noise levels (large SNR, crosses in Fig. 3). Qualitatively, this is due to the oscillatory features in the distribution produced by the noise during the deconvolution with the FFT method, as a consequence of the oscillatory nature of the Fourier basis functions. In general, it has been proved that in FFT methods the convergence to the actual distribution grows sublinearly with the sample size, with scaling dependencies that make the method unfeasible for practical purposes[16,19,25]. Since Bayesian deconvolution makes use of local basis functions, more parsimonious solutions can be obtained, preventing the degradation of the target distribution due to the noise.

B. An experimental dataset with an ad hoc noise distribution

Next we applied our method to real flow cytometry data using a noise distribution known *a priori*. Our goal was to mimick the conditions of a real dataset, while retaining an observable ground truth distribution. To that end, we studied the expression of the mesodermal gene Brachyury through a GFP reporter (T/Bra::GFP) in mouse embryonic stem cells (see Methods). Cells were treated for 24h with 3 μM CHIR99 and 25 ng/mL Activin A to upregulate Brachyury prior to flow cytometry on day 2 (see Methods). The signal of the GFP reporter was our target signal and it was acquired through the FITC-A channel. Note that this signal contains the fluorescence emitted by GFP together with the autofluo-

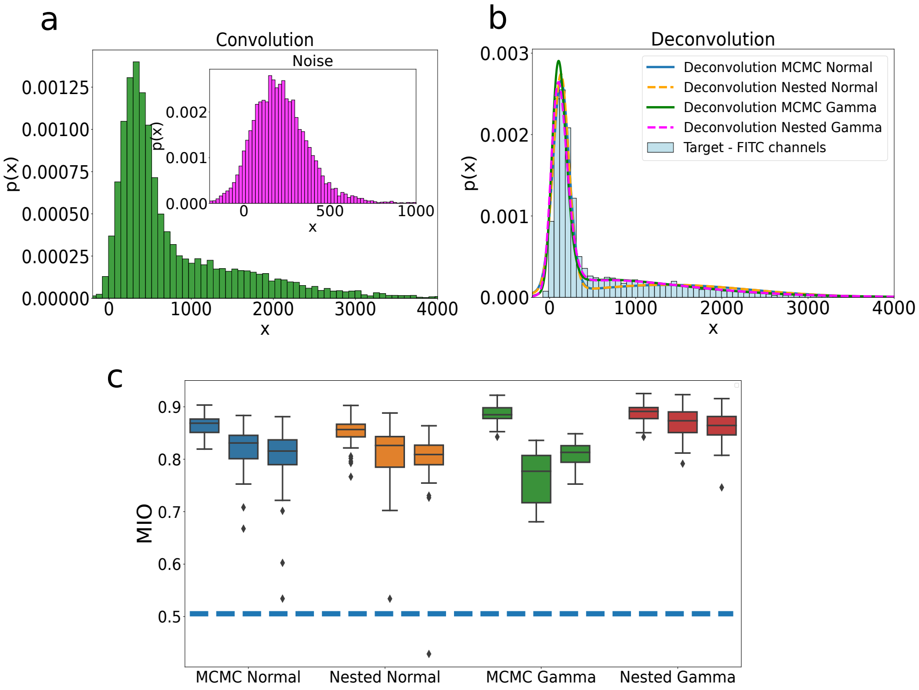

FIG. 4. Deconvolution of artificially convolved experimental data. a) Distribution of the total signal resulting from adding the GFP and red dye (PerCP-Cy5-5-A) signals for a population of 3000 cells (green bars). The "noise" signal corresponding to the dye channel is shown in magenta in the inset. b) Deconvolution of the green and magenta distributions shown in panel a, for both normal and gamma basis functions and MCMC and nested sampling methods, compared with the real target signal (light blue) measured in the GFP channel. c) Mean integrated overlap (MIO) between the deconvolved and ground-truth distributions for each of the four situations depicted in panel b. Each situation is divided in three different replicates with 3000 cells each. In each case the procedure was run 50 times, to produce a distribution of MIO values; outliers are represented as black diamonds.

rescence emitted by the cells at the GFP frequency. In that way, the signal collected in the GFP channel acts as our ground truth. Next we used the PerCP-Cy5-5-A channel as an artificial *ad hoc* noise (magenta bars in Fig. 4a, inset) and added it to the GFP signal in order to generate a 'convolved' distribution (green bars in Fig. 4a). Notice that the PerCP-Cy5-5-A and FITC-A channels correspond to different wavelengths. To avoid using twice the data from the PerCP-Cy5-5-A channel, which is contained in both the convolved dataset and the noise dataset, we used the PerCP-Cy5-5-A output from a different replica of the experiment. In addition, we were interested in checking the consistency of the deconvolutions between different experimental samples. For that purpose, we divided the data (consisting of 9000 cells in each replicate) in three subsets with 3000 samples each.

Figure 4(b) shows a typical deconvolution result for the case in which the target and noise distributions are described by mixtures composed of two and one components, respectively. The deconvolved distribution is compared to the original target distribution for both normal and gamma distributions, and for both MCMC and Nested Sampling. In the specific case shown in this plot, assuming gamma distributions for the deconvolution process leads to better results, irrespective of the sampling algorithm used. A systematic assessment of the efficiency of the method is shown in terms of the MIO in Fig. 4(c) As can be seen in the plot, the deconvolution procedure leads to a recovered target distribution that reproduces the original ground-truth distribution reasonably well in all cases, compared with simply ignoring the noise (indicated by the blue dashed horizontal line in the plot). In general, gamma distributions with nested sampling describe the data better than normal distributions. This is due to the highly skewed character of the flow cytometry data, which requires a higher number of normal mixture components to capture the distributions.

Additionally, the nested sampling solutions exhibit less fluctuations between subsamples than those generated by MCMC sampling. This is a consequence of the more accurate exploration of the posterior distribution enabled by the nested sampling in comparison with the MCMC method: MCMC trajectories have the tendency to become trapped in regions of parameter space with high posterior probability, when starting from random initial conditions, which can lead the algorithm getting stuck in suboptimal solutions from which it is hard to escape.

We recall that the deconvolution approach proposed here is based on the assumption that the target signal is independent of the auto-fluorescence. In the ad hoc experiment reported here, there is little (but non-negligible) correlation between the dye and GFP signals (Supplementary Fig. S3). The fact that our algorithm recovers the target signal successfully in this case indicates that the approach is robust to a certain degree of cross-talk between the two signals.

C. An experimental dataset with an external noise

Finally, we applied our method to an experimental dataset in which the noise distribution is unknown. To that end, we use again cells with the Brachyury reporter T/Bra::GFP in two media conditions (N2B27 supplemented with 3 μM CHIR99 and DMSO as control). At day 3, for each condition, one of the replicas was treated with 20 nM Green CMFDA dye and incubated for 3 min prior to flow cytometry, and the second one was incubated for 3 min with N2B27 (see Methods). Given that the dye incorporates in the cell cytoplasm and its emission spectrum is similar to the one of GFP, the dye acts as a noise source to the GFP signal coming from the Brachyury reporter. Consequently, we have the following four conditions, with their potential outcomes in terms of the signal measured in the FITC-A channel:

(c1) **DMSO**: Brachyury expression is minimal and the signal comes mainly from the intrinsic autofluorescence of the cells.

(c2) **DMSO+CMFDA**: Brachyury expression is minimal and the signal comes mainly from the CMFDA dye, plus intrinsic autofluorescence of the cells.

(c3) **N2B27+CHIR99**: Brachyury expression is upregulated, and thus the signal contains both the T/Bra::GFP reporter and the intrinsic autofluorescence of the cells.

(c4) **N2B27+CHIR99+CMFDA**: Brachyury expression is upregulated, and thus the signal contains both the T/Bra::GFP reporter, the signal from the CMFDA dye and the intrinsic autofluorescence of the cells.

The four signal distributions corresponding to these four conditions are shown in Supplementary Fig. S4. We note that the distribution of the signal coming only from the dye cannot be measured independently given the intrinsic autofluorescence of the cells. Moreover, the dye might be absorbed differently depending on the state of the cells (CHIR99 treatment), and the low concentration of the dye ($\sim$ nM) might lead to significant cell-to-cell variabilty on the absorbance of the dye. These features make it a suitable dataset to test the robustness of our method. We first used our method to deconvolve the signal of the dye from the total signal observed in the experimental conditions (c2), using condition (c1) to define the 'noise' distribution. To that end, we used nested sampling with gamma basis distributions, and considered two mixture components for the noise distribution and three for the target distribution. Again, we split the dataset in three subsamples, to quantify the consistency of the deconvolution. The results, shown in Fig. 5A, were consistent across replicates. Once the dye distribution was obtained from the deconvolution of conditions (c1) and (c2), we tested the consistency of our approach by considering the dye signal as the noise in condition (c4). Deconvolving the dye distribution from the one measured in (c4) should lead to the distribution obtained in condition (c3), which we can measure experimentally and thus serves as the ground truth in this case. The result can be seen in Fig. 5B. Even in this case, where we use a deconvolved dataset for a second deconvolution, we observe that the inferred distribution resulting from our method is in excellent agreement with the experimentally measured distribution in condition (c3) (MIO = 0.81).

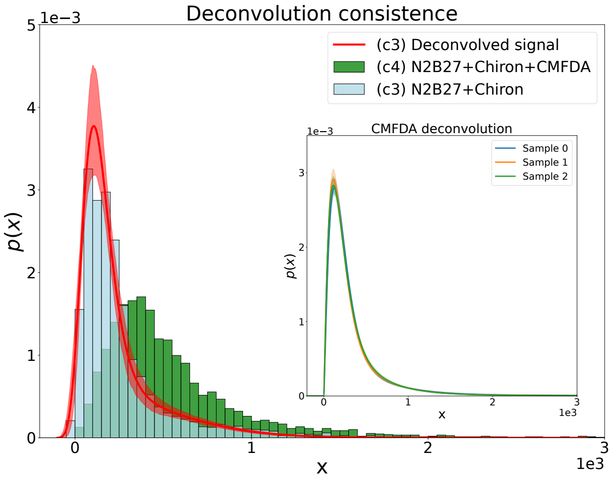

FIG. 5. Inferred distribution (red line) obtained by deconvolving the dye distribution shown in the inset from one of the distributions measured in condition (c4, green bars). The experimentally measured distribution in condition (c3) is shown with light blue bars. The inset shows the dye distribution deconvolved from experimental conditions (c1) and (c2), for the three subsets described in the text.

IV. DISCUSSION

In this paper, we propose a Bayesian approach to obtain flow-cytometry distributions of protein abundance or activity convolved with a known source of noise. The method, which relies on non-parametric Bayesian techniques, is freely available as a Python package (https://github.com/dsb-lab/scBayesDeconv) and can be used in a straightforward manner in a purely computational way, without the need of dedicated measurement channels or additional laser sources. It only requires measuring the fluorescently labeled and unlabeled cells and, unlike previously proposed deconvolution methods, it provides well-defined probability distributions, described by mixtures of basis functions.

We measure the quality of the results obtained with our method by comparing the deconvolved distributions with known (ground-truth) target distributions using synthetic data and ad hoc experiments with mouse embryonic stem cells. We argue that the use of local basis functions to describe all the distributions involved in the problem (both measured and unknown), and the corresponding use of a relatively small number of degrees of freedom, leads to an efficient inference. Finally, the Bayesian nature of the method gives rise to a set of candidate target probability distributions. This reflects the natural indeterminacy present in any process corrupted by noise. The ability to express indeterminacy in the solutions is crucial for any real application of a deconvolution algorithm. We further show that the method is robust even under strong noise in real datasets.

The method we propose is applicable so far only to one-dimensional distributions. In the case of multidimensional datasets, the algorithm could be immediately applied to one-dimensional projections of the data along the different dimensions, provided the channels are independent of each other. Furthermore, it is straightforward to apply the theory underlying Bayesian deconvolution to higher dimensional models, although in this case additional computational challenges might have to be faced.

Our results reveal nested sampling as a very robust method that enables exploring the posterior distribution extensively, as required by the deconvolution problem. The main drawback of this approach is the computational cost associated with the exploration, which increases exponentially with the dimensionality of the parameter space to be explored. Two sources of improvement over the current implementation could be explored in future work. The first is to reduce the evaluation time of the likelihood function. Usual flow cytometry datasets have on the order of thousands to tens of thousands of cells. For K_n mixture components of the noise and K_t components of the target, a single evaluation of the likelihood function scales as $\mathcal{O}(K_n \times K_t \times N)$, where N is the number of measures in the Bayesian model. Handling large datasets is a well established problem in Bayesian statistics, and some lines of work have explored solutions to reduce this computational cost, which might be worth exploring in the context of the single-cell Bayesian deconvolution method proposed here[36–38].

As for the second target for improvement, we note that mixture models have degeneracy under label exchange. The nested sampler is not able to detect such degeneracy, and spends much effort in probability peaks that virtually represent the same target distribution. This leads to a costly exponential slowdown during the search, as more particles are required to explore all the peaks of the distribution, instead of focusing on exploring only the different distributions consistent with the data. In order to solve this problem, two promising directions come to the mind of the authors. First, an asymmetric choice of priors that break the symmetry between components may solve the degeneracy of probability peaks and hence reduce the number of peaks that need to be explored by the algorithm. Second, mode search methods like the ones used for ground state search in protein folding, although being point based, can lead fast explorations of parameter space and give rise to lists of candidate peaks in high dimensional models, giving the possibility to scale the present models to very high number of components[39].

ACKNOWLEDGEMENTS

We thank the CRG Tissue Engineering Unit and the UPF/CRG Flow Cytometry Unit for continuous support. This work was supported by the Spanish Ministry of Science and Innovation and FEDER, under projects FIS2017-92551-EXP and PGC2018-101251-B-I00, by the "Maria de Maeztu" Programme for Units of Excellence in R&D (grant CEX2018-000792-M), and by the Generalitat de Catalunya (ICREA Academia programme). GT is supported by a FPU doctoral fellowship from the Spanish Ministry of Education and Universities (reference FPU18/05091). D.O. acknowledges funding from Juan de la Cierva Incorporación with Project no. IJC2018-035298-I, from the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI).

[1] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," Science **297**, 1183 (2002).

[2] G. M. Süel, R. P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, and M. B. Elowitz, "Tunability and noise dependence in differentiation dynamics," Science **315**, 1716–1719 (2007).

[3] H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang, "Transcriptome-wide noise controls lineage choice in mammalian progenitor cells," Nature **453**, 544–547 (2008).

[4] D. Osumi-Sutherland, C. Xu, M. Keays, A. P. Levine, P. V. Kharchenko, A. Regev, E. Lein, and S. A. Teichmann, "Cell type ontologies of the human cell atlas," Nature Cell Biology **23**, 1129–1135 (2021).

[5] S. Nowotschin, M. Setty, Y.-Y. Kuo, V. Liu, V. Garg, R. Sharma, C. S. Simon, N. Saiz, R. Gardner, S. C. Boutet, D. M. Church, P. A. Hoodless, A.-K. Hadjantonakis, and D. Pe'er, "The emergent landscape of the mouse gut endoderm at single-cell resolution," Nature **569**, 361–367 (2019).

[6] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein, "Lineage tracing on transcriptional landscapes links state to fate during differentiation," Science **367**, eaaw3381 (2020).

[7] M. Mittnenzweig, Y. Mayshar, S. Cheng, R. Ben-Yair, R. Hadas, Y. Rais, E. Chomsky, N. Reines, A. Uzonyi, L. Lumerman, A. Lifshitz, Z. Mukamel, A.-H. Orenbuch, A. Tanay, and Y. Stelzer, "A single-embryo, single-cell time-resolved model for mouse gastrulation," Cell **184**, 2825–2842.e22 (2021).

[8] J. Garcia-Ojalvo and A. Martinez Arias, "Towards a statistical mechanics of cell fate decisions," Current Opinion in Genetics & Development **22**, 619–626 (2012).

[9] P. Rué and J. Garcia-Ojalvo, "Modeling gene expression in time and space," Annual Review of Biophysics **42**, 605–627 (2013).

[10] J. P. Corsetti, S. V. Sotirchos, C. Cox, J. W. Cowles, J. F. Leary, and N. Blumburg, "Correction of cellular autofluorescence in flow cytometry by mathematical modeling of cellular fluorescence," *Cytometry*, Cytometry **9**, 539–547 (1988).

[11] M. Roederer and R. F. Murphy, "Cell-by-cell autofluorescence correction for low signal-to-noise systems: Application to epidermal growth factor endocytosis by 3t3 fibroblasts," Cytometry **7**, 558–565 (1986).

[12] S. Alberti, D. R. Parks, and L. A. Herzenberg, "A single laser method for subtraction of cell autofluorescence in flow cytometry," *Cytometry*, Cytometry **8**, 114–119 (1987).

[13] J. A. Steinkamp and C. C. Stewart, "Dual-laser, differential fluorescence correction method for reducing cellular background autofluorescence," *Cytometry*, Cytometry **7**, 566–574 (1986).

[14] We do not discuss in what follows other deconvolution problems that are not relevant to our situation, such as those related with composition of distributions, which usually require complete knowledge of the noise distribution[40,41].

[15] L. A. Stefanski and R. J. Carroll, "Deconvolving kernel density estimators," Statistics **21**, 169–184 (1990).

[16] M. C. Liu and R. Y. Taylor, "Simulations and computations of nonparametric density estimates for the deconvolution problem," Journal of Statistical Computation and Simulation **35**, 145–167 (1990).

[17] J. Fan, "On the optimal rates of convergence for nonparametric deconvolution problems," The Annals of Statistics **19**, 1257–1272 (1991).

[18] P. J. Diggle and P. Hall, "A Fourier approach to nonparametric deconvolution of a density estimate," Journal of the Royal Statistical Society: Series B (Methodological) **55**, 523–531 (1993).

[19] M. H. Neumann and O. Hössjer, "On the effect of estimating the error density in nonparametric deconvolution," *Journal of Nonparametric Statistics*, Journal of Nonparametric Statistics **7**, 307–330 (1997).

[20] F. Comte and C. Lacour, "Data-driven density estimation in the presence of additive noise with unknown distribution," Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**, 601–627 (2011).

[21] A. Delaigle and P. Hall, "Parametrically assisted nonparametric estimation of a density in the deconvolution problem," Journal of the American Statistical Association **109**, 717–729 (2014).

[22] J. Fan and J.-Y. Koo, "Wavelet deconvolution," IEEE Transactions on Information Theory **48**, 734–747 (2002).

[23] M. Lee, P. Hall, H. Shen, J. S. Marron, J. Tolle, and C. Burch, "Deconvolution estimation of mixture distributions with boundaries," Electronic Journal of Statistics **7**, 323–341 (2013).

[24] J. Staudenmayer, D. Ruppert, and J. P. Buonaccorsi, "Density estimation in the presence of heteroscedastic measurement error," Journal of the American Statistical Association **103**, 726–736 (2008).

[25] A. Sarkar, B. K. Mallick, J. Staudenmayer, D. Pati, and R. J. Carroll, "Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors," *Journal of Computational and Graphical Statistics*, Journal of Computational and Graphical Statistics **23**, 1101–1125 (2014).

[26] A. Sarkar, D. Pati, A. Chakraborty, B. K. Mallick, and R. J. Carroll, "Bayesian semiparametric multivariate density deconvolution," *Journal of the American Statistical Association*, Journal of the American Statistical Association **113**, 401–416 (2018).

[27] Throughout the paper we use greek subindices, in particular η and λ, when summing over basis functions, and roman subindices when summing over samples –cells–.

[28] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis, Third Edition* (Taylor & Francis, 2013).

[29] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, "Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells," Science **329**, 533 (2010).

[30] J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Analysis **1**, 833–859 (2006).

[31] F. Feroz and M. P. Hobson, "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses," Monthly Notices of the Royal Astronomical Society **384**, 449–463 (2008).

[32] E. Higson, W. Handley, M. Hobson, and A. Lasenby, "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation," Statistics and Computing **29**, 891–913 (2019).

[33] J. S. Speagle, "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences," Monthly Notices of the Royal Astronomical Society **493**, 3132–3158 (2020).

[34] H. J. Fehling, G. Lacaud, A. Kubo, M. Kennedy, S. Robertson, G. Keller, and V. Kouskoff, "Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation," Development **130**, 4217–4227 (2003).

[35] D. A. Turner, M. Girgin, L. Alonso-Crisostomo, V. Trivedi, P. Baillie-Johnson, C. R. Glodowski, P. C. Hayward, J. Collignon, C. Gustavsen, P. Serup, B. Steventon, M. P. Lutolf, and A. M. Arias, "Anteroposterior polarity and elongation in the absence of extra-embryonic tissues and of spatially localised signalling in gastruloids: mammalian embryonic organoids," Development **144**, 3894–3906 (2017).

[36] Z. Huang and A. Gelman, "Sampling for Bayesian computation with large datasets," (2005), available at SSRN: https://ssrn.com/abstract=1010107.

[37] C. Nemeth and C. Sherlock, "Merging MCMC subposteriors through Gaussian-process approximations," Bayesian Analysis **13**, 507–530 (2018).

[38] A. Vehtari, A. Gelman, T. Sivula, P. Jylänki, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. P. Robert, "Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data," Journal of Machine Learning Research **21**, 1–53 (2020).

[39] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," *The Journal of Chemical Physics*, The Journal of Chemical Physics **120**, 9911–9917 (2004).

[40] B. Efron, "Empirical Bayes deconvolution estimates," Biometrika **103**, 1–20 (2016).

[41] O.-H. Madrid-Padilla, N. G. Polson, and J. Scott, "A deconvolution path for mixtures," Electron. J. Statist. **12**, 1717–1751 (2018).