Bioinformatics, 35(6), 2019, 995–1001
doi: 10.1093/bioinformatics/bty718
Advance Access Publication Date: 23 August 2018
Original Paper

Gene expression

A Bayesian model for single cell transcript expression analysis on MERFISH data

Johannes Köster[1,2,3,*], Myles Brown[2,4,5] and X. Shirley Liu[4,6,7]

[1]Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, 45147 Essen, Germany, [2]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA, [3]Centrum Wiskunde and Informatica, Amsterdam, The Netherlands, [4]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [5]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02215, USA, [6]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA and [7]School of Life Science and Technology, Tongji University, Shanghai 200092, China

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on October 18, 2017; revised on June 28, 2018; editorial decision on August 16, 2018; accepted on August 21, 2018

Abstract

Motivation: Multiplexed error-robust fluorescence in-situ hybridization (MERFISH) is a recent technology to obtain spatially resolved gene or transcript expression profiles in single cells for hundreds to thousands of genes in parallel. So far, no statistical framework to analyze MERFISH data is available.
Results: We present a Bayesian model for single cell transcript expression analysis on MERFISH data. We show that the model successfully captures uncertainty in MERFISH data and eliminates systematic biases that can occur in raw RNA molecule counts obtained with MERFISH. Our model accurately estimates transcript expression and additionally provides the full probability distribution and credible intervals for each transcript. We further show how this enables MERFISH to scale towards the whole genome while being able to control the uncertainty in obtained results.
Availability and implementation: The presented model is implemented on top of Rust-Bio (Köster, 2016) and available open-source as MERFISHtools (https://merfishtools.github.io). It can be easily installed via Bioconda (Grüning *et al.*, 2018). The entire analysis performed in this paper is provided as a fully reproducible Snakemake (Köster and Rahmann, 2012) workflow via Zenodo (https://doi.org/10.5281/zenodo.752340).
Contact: johannes.koester@uni-due.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The investigation of gene or transcript expression at single cell resolution is of increasing importance in various areas of biological and medical research. In particular, it is used to study heterogeneous tissues like brain or tumors (Darmanis *et al.*, 2015; Patel *et al.*, 2014) and it enables the determination of cell types and states (Trapnell, 2015). Based on RNA sequencing (Eberwine *et al.*, 2014; Nawy, 2014), fluorescence microscopy (Femino, 1998; Lyubimova *et al.*, 2013) and mass spectrometry (Angelo *et al.*, 2014; Giesen *et al.*, 2014), various technologies and protocols to quantify gene, transcript or protein expression in single cells have emerged. Among these, *in situ* methods offer the possibility to preserve spatial information at cellular or even subcellular level. Currently, only fluorescence microscopy based approaches (Femino, 1998; Lubeck *et al.*, 2014; Nilsson *et al.*, 1994) offer information about the position of

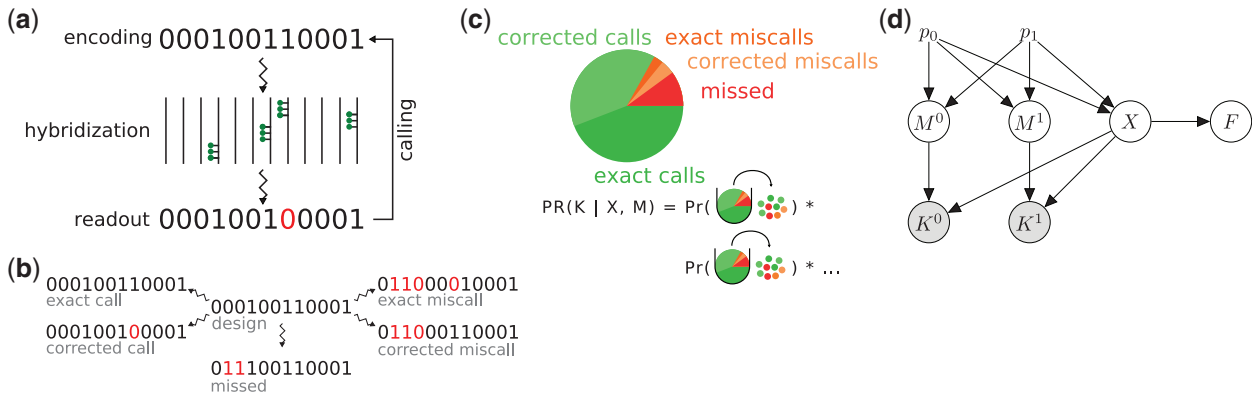

Fig. 1. The MERFISH protocol and Bayesian model for expression analysis on MERFISH data. (**a**) Outline of the MERFISH protocol. A binary encoding is designed for each transcript that shall be measured. Multiple FISH iterations are performed such that the binary word is replicated as an on/off pattern of fluorescent signals. The readout can contain errors, and has to be matched to the encoding to call the correct transcript. For details regarding hybridization and imaging see Chen *et al.*, 2015, Figs 1 and 2. (**b**) Events that can occur during the calling process. (**c**) Urn model to calculate the transcript expression likelihood. From the known error rates, we derive the probabilities for the different events that can happen in the calling process. These constitute an urn model for each transcript. By combining these urns into a joint model over all transcripts, the likelihood of transcript expression given the observed counts can be calculated. (**d**) Graphical representation of the model as Bayesian network. Latent variables are circles, observed variables are shaded circles. The error rate vectors p_0 and p_1 are constant parameters. The moderated fold change F, as well as matrices for expression (X) and miscalls M^0 and M^1 are latent variables. The counts K^0 and K^1 are observed variables

each observed RNA molecule (i.e. transcript) within the cell (Crosetto *et al.*, 2015).

With multiplexed error-robust fluorescence in-situ hybridization (MERFISH), Chen *et al.* (2015) presented a new approach that allows the measurement of hundreds to thousands of different transcripts at the same time, while retaining their coordinates within the cell. In contrast to previous approaches that use multiple colors to measure multiple different transcripts at the same time (Lubeck *et al.*, 2014), MERFISH uses a binary code to represent RNA species. First, a unique binary word (the encoding) of length n is assigned to each gene or transcript of interest (see Fig. 1a). Then, specific readout probes and corresponding fluorescent labels are designed for n rounds of hybridization and bleaching. Thereby probes are designed such that a 1-bit at position i in the binary word corresponds to a fluorescence signal in the ith hybridization round. When imaging each of the n hybridization rounds, RNA molecules can be localized by fluorescent spots, and the on/off pattern of a spot over the n hybridization rounds yields a binary word (the readout) again. By matching the latter against the designed encoding, individual RNA molecules can be assigned to transcripts or genes, and expression can be quantified. Since both the hybridization and the fluorescence measurement are error-prone, individual bits in the readout can differ from the encoding. To enable the matching to deal with such errors, Chen *et al.* (2015) propose to use a binary code with a given minimum Hamming distance (Hamming, 1950) between any pair of words. With a Hamming distance of 2, 1-bit errors can be detected. With a Hamming distance of 4, 1-bit errors can also be corrected, by assigning to the closest encoding.

More than one error in a particular readout can lead to missed calls and misidentification, which both are the source for systematic biases in MERFISH data. Indeed, Chen *et al.* (2015) already report that about 20% of the molecules are missed when using the first version of the MERFISH protocol. A remarkable property of MERFISH is that the average probability for accidentally reading a 1-bit instead of a 0-bit (0-1 error) and vice versa (1-0 error) can be easily determined (Chen *et al.*, 2015; Moffitt *et al.*, 2016). In this work, we use this knowledge to build a Bayesian model for gene or transcript expression on MERFISH data. We show that this model enables the correction of the systematic biases that occur within raw counts. In addition, the model reports full posterior expression distributions for each transcript along with credible intervals. In line with recent reports (Halsey *et al.*, 2015), this enables a thorough exploration of the uncertainties in downstream analyses, which is exemplified in Supplementary Section S6.

2 Materials and methods

We strive to obtain a posterior estimate of transcript (or gene) expression in each single cell that is free of biases (Fig. 1). Here, the unit of expression shall be RNA molecule counts. Using the known rates for 0-1 and 1-0 errors, we derive probabilities for exact calls (i.e. the event that an RNA molecule is identified correctly and no error occured), corrected calls (correct identification, but 1-bit error corrected), exact miscalls (RNA molecule is assigned to the wrong transcript), corrected miscalls (RNA molecule is wrongly assigned although a 1-bit error has been corrected) and dropouts (RNA molecule is missed because the binary readout is not recognized at all). These probabilities define an urn model that allows the calculation of the probability to observe a certain distribution of calls, miscalls and dropouts for a given transcript expression using a multinomial distribution. By combining such urns into a joint model over all transcripts, the joint likelihood of transcript expressions can be calculated. To ensure computational feasibility, we first estimate the maximum likelihood solution using the EM algorithm, and then calculate individual likelihoods per transcript while fixing miscalls and expressions of other transcripts to the maximum likelihood estimates. Bayes' theorem yields corresponding posterior probabilities from which the maximum a posteriori expression along with credible intervals can be calculated.

2.1 Readout probabilities

With MERFISH, each transcript is labeled with a different set of encoding probes. In each of N performed hybridization rounds, different encoding probes are targeted with fluorescent labels. The encoding probes are designed such that for each transcript, there are

m unique fluorescence signals in N hybridization rounds. This *readout* can be seen as a binary word. Chen *et al.* (2015) present two encoding schemes. With MHD4, the binary words (say, *encoding*) assigned to each transcript have length $N=16$ (i.e. number of hybridization rounds), contain $m=4$ 1-bits and the minimum Hamming distance between two words is 4. With MHD2, words have length $N=14$ with $m=4$ 1-bits and a minimum Hamming distance of 2. Technically, the readout for each RNA molecule in a cell is determined using fluorescence microscopy and subsequent computational recognition and assignment of the fluorescence signals to binary words (the readouts). In the following we will use the terms encoding and transcript interchangeably, since every transcript of interest is represented by exactly one encoding (i.e. binary word) in a MERFISH experiment design. Since all stages are error-prone, the measurement of the readouts can lead to flipped bits. At position k in the binary word, we denote with $p_{0,k}$ the probability to accidentally obtain a 1-bit instead of a 0-bit (0-1 error rate) and with $p_{1,k}$ the probability to obtain a 0-bit instead of a 1-bit (1-0 error rate). It is worth noting that our model is agnostic of the underlying process as long as the measured outcome is a binary word and error rates are known. Hence, it will be applicable to any future versions of MERFISH as long as these properties stay the same. The error rates p_0 and p_1 can be estimated from MERFISH data, see Supplementary Section S4. In the following, we derive probabilities for different readout events.

After readouts have been obtained, they can be compared to the designed encodings in order to determine from which transcript they come. With MHD4, it is possible to correct for 1-bit errors during this calling process, with MHD2, it is only possible to detect, but not correct 1-bit errors (Chen *et al.*, 2015). Obviously, these calls can be incorrect and incomplete, which motivates the comprehensive description of the underlying uncertainties as it will be done in the following.

Let N be the number of bits in the used encoding and $m \leq N$ be the number of 1-bits. First, we strive to calculate the probability to obtain a readout that is called as a certain target transcript. This can be an exact call with hamming distance $d=0$ to the target transcript or a corrected call with hamming distance $d=1$. We denote with t the binary encoding of the target transcript and with s the binary encoding of the real origin of the readout (i.e. the source). The probability is denoted as

$$\xi_{d,s,t} = \xi'_{d,s,t,N}$$

with ξ' being the recurrence

$$\xi'_{d,s,t,k} = \begin{cases} p_{1,k}\xi'_{d-1,s,t,k-1} + (1-p_{1,k})\xi'_{d,s,t,k-1} \\ \quad \text{if } s_k = t_k = 1, k > 0, d \geq 0 \\ p_{0,k}\xi'_{d-1,s,t,k-1} + (1-p_{0,k})\xi'_{d,s,t,k-1} \\ \quad \text{if } s_k = t_k = 0, k > 0, d \geq 0 \\ p_{1,k}\xi'_{d,s,t,k-1} + (1-p_{1,k})\xi'_{d-1,s,t,k-1} \\ \quad \text{if } s_k = 1 \neq t_k, k > 0, d \geq 0 \\ p_{0,k}\xi'_{d,s,t,k-1} + (1-p_{0,k})\xi'_{d-1,s,t,k-1} \\ \quad \text{if } s_k = 0 \neq t_k, k > 0, d \geq 0 \\ 1 \text{ if } k = 0, d = 0 \\ 0 \text{ if } d < 0. \end{cases}$$

In the recurrence ξ' we calculate the probability for the prefix of length k. If $s_k = t_k$, any bit flip (the first term of the respective sums) increases the Hamming distance to the target transcript, whereas no bit flip keeps the previous distance (the second term of the sum). In contrast, if $s_k \neq t_k$, a bit flip keeps the previous Hamming distance (because it will flip the source bit to the correct bit of the target encoding), while no bit flip increases the Hamming distance to the target transcript.

The function ξ can be used to calculate probabilities for the different events that can occur to the readout of a certain transcript. Let e be the encoding for an arbitrary transcript and r_e be the obtained readout. Further, let $\delta(r_e, e) = d$ denote the event that the Hamming distance between any readout r and any encoding e is $d \in \mathbb{N}$. The probability that the transcript is called correctly with Hamming distance 0 (an exact call) is

$$\Pr(\delta(r_e, e) = 0) = \xi_{0,e,e}.$$

The probability for a corrected call (Hamming distance 1) of the transcript of origin is

$$\Pr(\delta(r_e, e) = 1) = \xi_{1,e,e}.$$

Note that in both cases the source and target transcript are the same, because we calculate the probability for a correct call. The next type of event that can happen to a transcript is that it is miscalled as another transcript. Let e' be the encoding of the latter. Then, we can calculate the probability for an exact miscall of e' as

$$\Pr(\delta(r_e, e') = 0) = \xi_{0,e,e'}$$

and the probability for a corrected miscall of e' as

$$\Pr(\delta(r_e, e') = 1) = \xi_{1,e,e'}.$$

We consider above probabilities for all e' that have a reasonable Hamming distance to e and denote this set of encodings as neighbors $\mathcal{N}_e$. With the error probabilities known so far, a maximum Hamming distance of 4 is a reasonable threshold because the probability would essentially be zero beyond this [as previously also shown by Chen *et al.* (2015)]. Finally, the probability for a dropout event (i.e. the transcript is neither called correctly nor miscalled because the readout contains too many errors) is the sum of all remaining combinations of bit flips that haven't been covered by above events. Alternatively, the result can be obtained by summing call and miscall probabilities:

$$\Pr(\forall e' \neq e : \delta(r_e, e') > 1) = 1 - \sum_{e' : \{e\} \cup \mathcal{N}_e} \Pr(\delta(r_e, e') = 0) + \Pr(\delta(r_e, e') = 1).$$

2.2 Estimating transcript expression

In the following, we consider the transcript (or gene) expression as the real, but unknown, number of RNA molecules present in a single cell. We first define parameters, latent and observed variables of the model.

Let E be the set of encodings given by a MERFISH codebook. The MERFISH protocol suggests to keep a small set $\bar{E} \subset E$ of encodings as misidentification probes, i.e. they are not assigned to any transcript. Any call of a misidentification probe can be considered as an artifact. There are two reasons for such an artifact. First, it can be generated by a miscall from an expressed transcript. Second, it can be generated out of nothing, by interpreting microscopy or hybridization noise as a real signal. To model the latter, we consider 'noise' to be a special transcript with encoding $e_n = 0^N$, which will be later treated in a special way.

We denote with $n = |E \setminus \bar{E}|$ the number of real transcripts that are measured in the experiment and with $n' = |E|$ the number of all measured encodings (including misidentification probes). Let $X \in \mathbb{N}^{n+1}$ be the vector of expressions of all transcripts and the noise

source. Let $M^d \in \mathbb{N}^{(n'+1)\times(n'+1)}$ with $d \in \{0,1\}$ be the matrices of miscalls with Hamming distance 0 and 1, with $M^d_{e,e'}$ denoting the number of miscalls from encoding e to encoding e'. Finally, let $K^d \in \mathbb{N}^{n'}$ with $d \in \{0,1\}$ be the vector of observed counts with Hamming distance 0 and 1 to the called encoding (e.g. K^0_e are the observed exact counts that have been assigned to the transcript represented by the binary word e). While K is an observed variable, M is a latent variable that we will have to estimate along with the expressions X. See Figure 1d for a graphical representation of the model.

For each transcript with encoding $e \in E \setminus \bar{E}$, we now define an urn model that represents the different events that can occur to obtained readouts. Imagine that the urn contains differently colored balls for exact calls, corrected calls, miscalls to every neighbor $e' \in \mathcal{N}_e$ and dropouts. Their relative quantities are defined by the readout probabilities from Section 2.1. The number of draws from the urn is then equivalent to the expression of the transcript and the distribution of drawn balls is equivalent to the observed and latent events that happen in the experiment. More formally we can write

$$\Pr(K^0_e, K^1_e \mid X, M^0, M^1) = h(\kappa^0_e, \kappa^1_e, M^0_{e,e_1}, M^0_{e,e_2}, \ldots, M^0_{e,e_k}, M^1_{e,e_1}, M^1_{e,e_2}, \ldots, M^1_{e,e_k}, \lambda_e; X_e)$$

with $h(\ldots; X_e)$ being the multinomial probability mass function with event probabilities

$$\Pr(\delta(r_e, e) = 0), \Pr(\delta(r_e, e) = 1),$$
$$\Pr(\delta(r_e, e_1) = 0), \Pr(\delta(r_e, e_2) = 0), \ldots, \Pr(\delta(r_e, e_k) = 0),$$
$$\Pr(\delta(r_e, e_1) = 1), \Pr(\delta(r_e, e_2) = 1), \ldots, \Pr(\delta(r_e, e_k) = 1),$$
$$\Pr(\forall e' \neq e : \delta(r_e, e') > 1)$$

and X_e trials. The support (i.e. the drawn events) consists of the correct exact and inexact calls κ^0_e and κ^1_e, the miscalls to neighboring encodings, and the dropout events λ_e. Thereby it is $\kappa^0_e = K^0_e - \sum_{e' \in \mathcal{N}'_e \cup \{e_n\}} M^0_{e,e'}$ and $\kappa^1_e = K^1_e - \sum_{e' \in \mathcal{N}'_e \cup \{e_n\}} M^1_{e,e'}$, i.e. the correct calls are the observed counts minus miscalls from other transcripts or the artificial noise source (see above). For the dropout events λ_e we take the number of trials X_e minus the sum of the other events. Note that miscalls computed by the urn for one transcript reoccur in the κ terms of neighboring transcripts. While this ensures that the real processes in a MERFISH experiment are properly modeled, it also gives rise to additional computational challenges outlined in Supplementary Section S3.

For the artificial noise source, we define a slightly different urn. While we do not have directly assigned counts for the noise, we can use the misidentification probes as evidence. Instead of having two events for exact and inexact calls (κ^0 and κ^1), we define the two events for each misidentification probe, i.e. $\kappa^0_{\bar{e}}, \kappa^1_{\bar{e}}$ for each $\bar{e} \in \bar{E}$. We use the corresponding probabilities $\Pr(\delta(r_{e_n}, \bar{e}) = 0)$, $\Pr(\delta(r_{e_n}, \bar{e}) = 1)$ for miscalling $\bar{e}$ from noise as event probabilities and keep the rest of the urn model as defined before.

Now we can define a joint model for the likelihood of transcript expression as

$$\Pr(K^0, K^1 \mid X) = \sum_{M^0, M^1} \prod_{e \in E \setminus \bar{E}} \Pr(K^0_e, K^1_e \mid X, M^0, M^1).$$

Naturally, this is infeasible to calculate because of the combinatorial number of summands. However, we can use the EM algorithm (Dempster *et al.*, 1977) to obtain an approximation of the maximum likelihood solution $\hat{X}, \hat{M}^0, \hat{M}^1$ (see Supplementary Section S3).

By keeping miscalls $\hat{M}^0, \hat{M}^1$ fixed and keeping $\hat{X}$ fixed for every transcript except a particular one e, we can approximate the posterior probability distribution for expression X_e as

$$\Pr(X_e = x \mid K^0, K^1) \approx \frac{\Pr(X_e = x) h(x)}{\sum_{x'} \Pr(X_e = x') h(x')}$$

with

$$h(x) := \Pr(K^0, K^1 \mid X_e = x, X_{e'} = \hat{X}'_e, \forall e' \neq e, \hat{M}^0, \hat{M}^1).$$

Depending on the experiment, reasonable prior probabilities could be negative binomial or Poisson distributed. However, MERFISH so far is applied to subsets of all available genes or transcripts. Their expression distribution can be heterogeneous in single cells and strongly affected by the investigated conditions. Finding appropriate priors is, among other tasks, subject to future research (see Supplementary Section S7). For now, we consider flat prior probabilities, which is conservative in the sense that (a) the resulting probability mass functions will be less sharp and (b) no further assumptions about the experiment are made. From the posterior probabilities, we can report the maximum a posteriori probability (MAP) and the 95% credible interval.

3 Results

We strive to evaluate the benefits of using the presented model on both simulated and real MERFISH data. For this, we will first show that the obtained expression estimates are, in contrast to raw counts, unbiased. Since there exists no real MERFISH dataset where the true molecule counts are known on a large scale, we show this on simulated data. On real data, we will show that the presented model enables improved correlation among the same genes across datasets.

3.1 Available data

Currently, two versions of the MERFISH protocol are available: Chen *et al.* (2015) published version 1 with a reported average 0-1 error rate of $p_0 = 4\%$ and a 1-0 error rate of $p_1 = 10\%$. Later, Moffitt *et al.* (2016) published version 2 with significantly improved average error rates of $p_0 = 0.5\%$ and a $p_1 = 1\%$. To assess the accuracy of our method, we simulate MERFISH data for both protocols. Note that we estimated higher error rates as reported for protocol 1 (see Supplementary Section S4). However, we will use the reported lower error rates for our simulations to avoid an exaggeration of the benefits. Further, we use real data from protocol version 1 published by Chen *et al.* (2015): The *MHD4 dataset* measures 130 primary transcripts (plus 10 controls) on human fibroblast cells (IMR90), using two codebooks (i.e. the binary words assigned to the transcripts) with a pairwise Hamming distance of at least 4. The *MHD2 dataset* measures 985 primary transcripts (plus 16 controls) on the same cell line, using a codebook with a pairwise Hamming distance of at least 2.

3.2 Example output

Figure 2 exemplifies the output of our model for two genes of the MHD4 dataset. Since the model reports credible intervals and the entire probability distribution for each estimate, the certainty of results can be properly assessed even in single cells. In most cases, the raw counts obtained by MERFISH are lower than our posterior estimates. This is consistent with the tendency of underestimation reported by Chen *et al.* (2015).

3.3 Evaluation on simulated data

To validate the estimates of the Bayesian model, we simulated MERFISH counts for different codebooks and the two protocol versions. Transcript expressions were drawn from Poisson distributions with different means. For each drawn count, the corresponding binary word from the codebook was mutated according to the reported average rates of 0-1 and 1-0 errors of the used protocol version. In addition we simulated noise by adding 50% of the total counts as 0-only binary words and mutating them in the same way. This is equivalent to expecting $\frac{2}{3}$ of the investigated cell to be covered by transcripts. Finally, the mutated binary words were matched against the codebook, yielding artificial raw counts that are subject to the same miscall and dropout effects as real MERFISH data. Figure 3a–c and Supplementary Figure S1d–f (see Supplementary Section S1) provide results for protocol version 1 with the codebooks from the MHD4 and MHD2 datasets described above. The simulation is able to reproduce the reported underestimation bias of raw counts. In contrast, the posterior estimates reported by our Bayesian model do not show this bias and provide a more accurate estimate of the true expression. This is also robust with respect to uncertainty in the 0-1 and 1-0 error rates (see Supplementary Fig. S3). With protocol version 2, as error rates are much smaller, biases are expected to be less severe. In fact, when using the MHD4 codebook with protocol version 2 (MHD4v2), they disappear even when only considering the raw counts (Supplementary Fig. S1a–c). The corresponding credible intervals predicted by our model are narrow, which reflects the increased certainty in the data. While bias is not an issue with MHD4v2, there is still uncertainty in the data and our model properly captures this in the credible intervals (Supplementary Fig. S2c). The improved error rates of protocol version 2 offer the opportunity to use a more aggressive encoding in order to scale towards the whole genome. To illustrate this, we

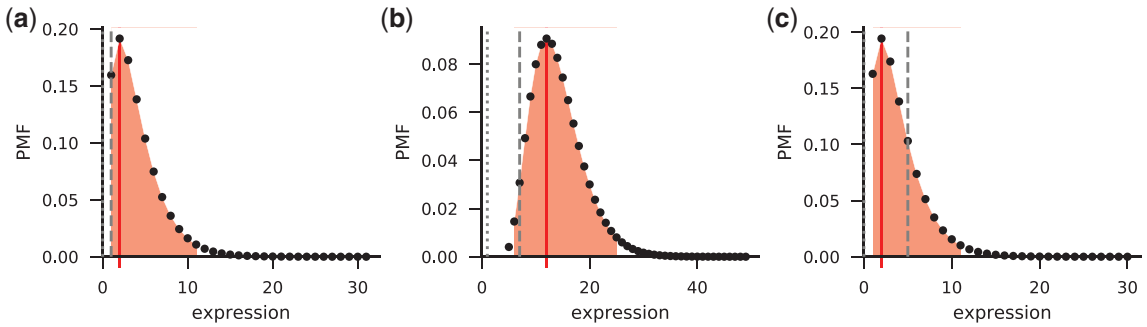

Fig. 2. The presented Bayesian model can estimate probability distributions, maximum a-posteriori estimates (MAP) and credible intervals of transcript expression. The figure shows example probability mass functions (PMF, black dots) of the genes FLNC (**a, b**) and PRKCA (**c**) in three single cells from the MHD4 dataset. The raw total count (number of readouts assigned to the gene) is shown as dashed, and the raw exact count (number of readouts where no bit-correction was necessary) is shown as dotted gray lines. See Supplementary Section S2 for the general distribution of exact versus corrected counts in the MHD4 dataset

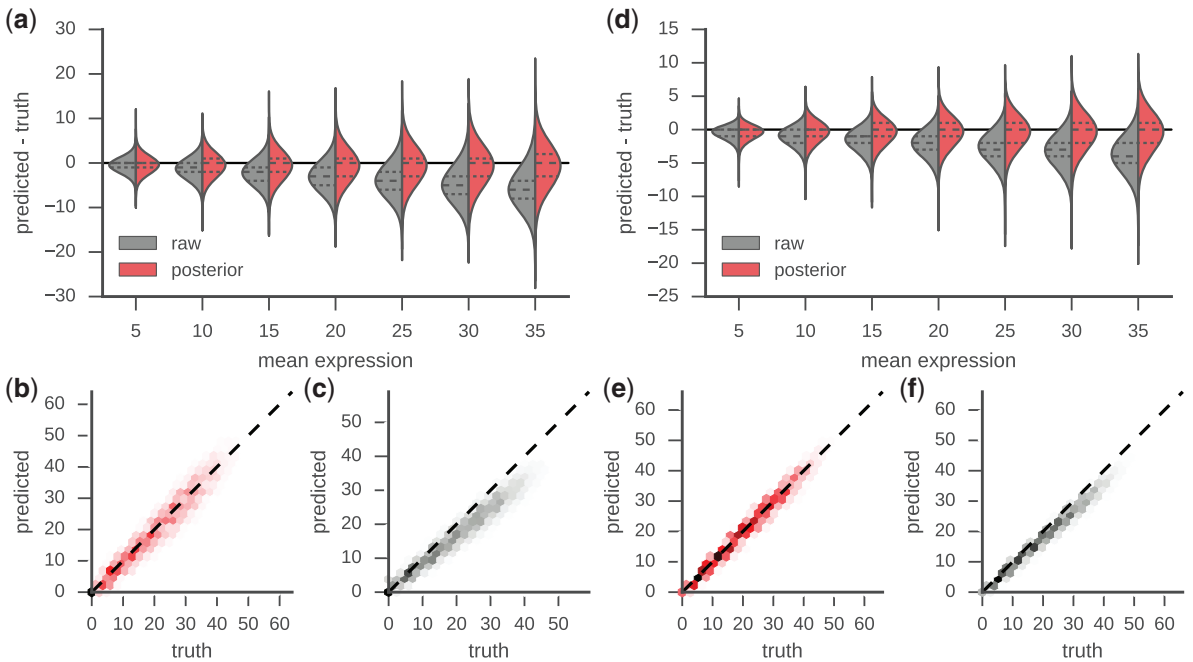

Fig. 3. The presented Bayesian model is an accurate estimator of gene expression. We simulated MERFISH data under protocol version 1 (MHD4 codebook) and protocol version 2 (MHD2v2 codebook). (**a**) Prediction error as violin plots for different mean expressions using posterior estimates (maximum a posterior probability, right violin half) and raw counts (left violin half) with MHD4 encoding. (**b**) Predicted versus true gene expression using posterior estimates (**c**) and raw counts with MHD4 encoding. Data is shown in hexagonal bins with intensities corresponding to bin counts (i.e. a 2D histogram). (**d, e, f**) Corresponding results for MHD2v2 encoding

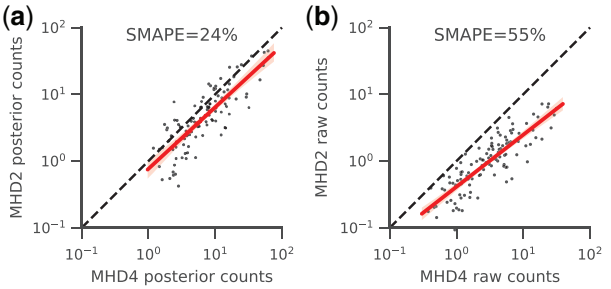

Fig. 4. Correlation of per-gene means between MHD4 and MHD2 datasets. Each dot represents the mean estimate across all cells in the two datasets. SMAPE denotes the symmetric mean absolute percentage error for a prediction between the two datasets. Linear regression is depicted by the line. (**a**) Posterior estimates, (**b**) raw counts

generated a codebook with a pairwise hamming distance of at least 2, binary words of length $N = 16$, and $m = 8$ 1-bits (MHD2v2, see Section 2). Such a codebook permits to measure 12, 870 transcripts at the same time, which is about $\frac{2}{3}$ of the expected number of protein coding genes in the human genome. Figure 3d–f shows that even the improved error rates of MERFISH protocol version 2 cannot avoid biased raw counts for such a codebook (this is also the case for the regular MHD2 codebook in combination with protocol version 2). Again, our Bayesian model removes the bias, thereby enabling an almost whole-genome scale analysis of MERFISH data.

3.4 Evaluation on real data

The MHD4 and MHD2 datasets published by Chen *et al.* (2015) share 107 genes. If both experiments would have perfectly measured all transcripts from these genes on exactly the same cells, we would expect both experiments to yield exactly the same counts. In reality, the experiments have of course been conducted on different cells. Nevertheless, under the assumption that biological conditions have been the same, the mean expression of each gene within both datasets should be the same. We used this property as a proxy for missing gold standard data, and calculated for both raw counts and the posterior estimates yielded by our model how accurately the per-gene means agree between the two datasets. First, we predicted gene expression on both datasets as (a) the posterior estimates yielded by our Bayesian model and (b) the raw counts. Let G be the set of common genes and $m_{g,d}$ be the mean prediction of gene $g \in G$ in dataset $d \in \{\text{MHD2}, \text{MHD4}\}$. Figure 4 depicts the agreement between the two datasets for both prediction methods in terms of (a) the predicted means for each gene and (b) the symmetric mean absolute percentage error

$$SMAPE = \frac{\sum_{g \in G} |m_{g,\text{MHD2}} - m_{g,\text{MHD4}}|}{\sum_{g \in G} m_{g,\text{MHD2}} + m_{g,\text{MHD4}}}$$

between the two datasets for both posterior and raw prediction. A perfect agreement cannot be expected due to the fact that (a) different cells are measured and (b) batch effects exist between the different MERFISH datasets conducted by Chen *et al.* (2015) (see Supplementary Section S6). However, it can be seen that the Bayesian model provides a significantly improved agreement between the datasets compared to using raw counts.

4 Discussion

MERFISH has been identified as a major advance in spatial transcriptomics (Shalek and Satija, 2015). In this work, we presented a Bayesian model for single-cell gene or transcript expression analysis on MERFISH data. Using simulated and real data, we showed that the model provides expression estimates that are free of systematic biases seen with raw MERFISH counts. Reported estimates from our model are complemented by credible intervals and the entire probability distribution, which allows to obtain a complete picture of the uncertainties as it was, e.g. demanded by Halsey *et al.* (2015). In the Supplementary Section S6 we exemplify how this information can be used to control the false discovery rate of differential expression analysis. The presented model provides, for the first time, a framework to assess gene or transcript expression with MERFISH, while properly handling and summarizing uncertainty in the data, even when investigating a single cell. In combination with the improved accuracy of version 2 of the MERFISH protocol (Moffitt *et al.*, 2016) our model even enables to increase the number of measured transcripts by an order of magnitude, thereby scaling MERFISH towards the whole genome.

Uncertainties in MERFISH data stem from errors made during hybridization, fluorescence microscopy and imaging. All these accumulate in the 0-1 and 1-0 error rates that are the main parameter of our Bayesian model (they can be estimated from the data). Given that MERFISH is a new technology, it is not yet possible to quantify how the error rates would differ between labs and microscopes. Since our model takes these rates into account, the provided unbiased estimates of transcript expressions together with information about the uncertainty will help to provide reproducible results across labs and machines, even if error rates turn out to differ significantly.

We have shown that the current version of our model is an accurate predictor for Poisson distributed transcript expressions, although it uses a flat prior. Future work will entail the evaluation of suitable prior probability distributions (e.g. negative binomial), that enable to (a) introduce prior knowledge about expected mean and overdispersion or (b) infer these parameters from the data. Given that MERFISH is a targeted approach, we also plan to support setting priors per gene or group of genes. By this, knowledge about perturbations that are applied to the cells (e.g. the knockout of a gene) can be incorporated into the Bayesian inference.

The MHD4 dataset published by Chen *et al.* (2015) exhibits an additional, protocol-specific type of bias: a caveat in the MERFISH image analysis causes increased dropout rates for encodings with a large Hamming distance to the most abundant encoding. This problem has been eliminated in version 2 of the MERFISH protocol (Moffitt *et al.*, 2016) (personal communication). Due to certain choices in our implementation of the presented Bayesian model we are currently only able to partially eliminate this type of bias. In Supplementary Section S5, we analyze the bias in detail and outline our plans to improve our implementation accordingly.

With MERFISH protocol version 2, the throughput has been improved considerably, enabling to measure hundreds of thousands of cells (Moffitt *et al.*, 2016). This necessitates the efficient data exchange between the different steps of MERFISH analysis. Instead of introducing a custom binary file format, we plan to use an existing generic approach for binary and compressed data encoding, like Apache Arrow (https://arrow.apache.org). This will ensure interoperability with other tools as well as scripting languages typically used for post processing (e.g. Python and R).

With single-cell expression analysis, it is reasonable to consider cells to be in different states and transcript expression to be dynamic over time (e.g. when investigating cell differentiation). Here, an important tool is to label individual cells with a pseudotime, that represents at which point in time of a dynamic process each cell appears

to be. Campbell and Yau (2016) have previously shown how uncertainties in gene expression can be incorporated into a Bayesian model for pseudotime inference. Future work might entail the integration of the uncertainty information provided by our model into such frameworks, such that reliable pseudotime estimates can be calculated for MERFISH data.

Acknowledgements

We thank Jeffrey Moffit for extensive, insightful discussions about technological details of MERFISH. We thank our exceptionally dedicated reviewers, who have contributed a lot to the quality of this work. Further, we thank Peng Jiang, Bo Li and Eric Severson for fruitful discussions.

Author contributions

J.K. developed the model, designed and conducted the analysis, implemented MERFISHtools and wrote the manuscript. M.B. and X.S.L. supervised the research and contributed to the manuscript.

Funding

This work was supported by the National Institutes of Health [R01HG008728 to M.B. and X.S.L., R01GM099409 to X.S.L.].

Conflict of Interest: none declared.

References

Angelo,M. *et al*. (2014) Multiplexed ion beam imaging of human breast tumors. *Nat. Med*., **20**, 436–442.

Campbell,K.R. and Yau,C. (2016) Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Comput. Biol*., **12**, e1005212.

Chen,K.H. *et al*. (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.

Crosetto,N. *et al*. (2015) Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet*., **16**, 57–66.

Darmanis,S. *et al*. (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA*, **112**, 7285–7290.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodological)*, **39**, 1–38.

Eberwine,J. *et al*. (2014) The promise of single-cell sequencing. *Nat. Methods*, **11**, 25–27.

Femino,A.M. *et al*. (1998) Visualization of single RNA transcripts in situ. *Science*, **280**, 585–590.

Giesen,C. *et al*. (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, **11**, 417–422.

Grüning,B. *et al*. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.

Halsey,L.G. *et al*. (2015) The fickle P value generates irreproducible results. *Nat. Methods*, **12**, 179–185.

Hamming,R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech.J*., **29**, 147–160.

Köster,J. (2016) Rust-Bio: a fast and safe bioinformatics library. *Bioinformatics*, **32**, 444–446.

Köster,J. and Rahmann,S. (2012) Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Lubeck,E. *et al*. (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, **11**, 360–361.

Lyubimova,A. *et al*. (2013) Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc*., **8**, 1743–1758.

Moffitt,J.R. *et al*. (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. USA*, **113**, 11046–11051.

Nawy,T. (2014) Single-cell sequencing. *Nat. Methods*, **11**, 18.

Nilsson,M. *et al*. (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, **265**, 2085–2088.

Patel,A.P. *et al*. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, **344**, 1396–1401.

Shalek,A.K. and Satija,R. (2015) MERFISHing for spatial context. *Trends Immunol*., **36**, 390–391.

Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res*., **25**, 1491–1498.