Westoby *et al. Genome Biology* (2020) 21:74
https://doi.org/10.1186/s13059-020-01981-w

Genome Biology

RESEARCH Open Access

Obstacles to detecting isoforms using full-length scRNA-seq data

Jennifer Westoby[1,2*], Pavel Artemov[1], Martin Hemberg[2†] and Anne Ferguson-Smith[1†]

Abstract

Background: Early single-cell RNA-seq (scRNA-seq) studies suggested that it was unusual to see more than one isoform being produced from a gene in a single cell, even when multiple isoforms were detected in matched bulk RNA-seq samples. However, these studies generally did not consider the impact of dropouts or isoform quantification errors, potentially confounding the results of these analyses.

Results: In this study, we take a simulation based approach in which we explicitly account for dropouts and isoform quantification errors. We use our simulations to ask to what extent it is possible to study alternative splicing using scRNA-seq. Additionally, we ask what limitations must be overcome to make splicing analysis feasible. We find that the high rate of dropouts associated with scRNA-seq is a major obstacle to studying alternative splicing. In mice and other well-established model organisms, the relatively low rate of isoform quantification errors poses a lesser obstacle to splicing analysis. We find that different models of isoform choice meaningfully change our simulation results.

Conclusions: To accurately study alternative splicing with single-cell RNA-seq, a better understanding of isoform choice and the errors associated with scRNA-seq is required. An increase in the capture efficiency of scRNA-seq would also be beneficial. Until some or all of the above are achieved, we do not recommend attempting to resolve isoforms in individual cells using scRNA-seq.

Keywords: scRNA-seq, Single cell, Alternative splicing, Isoform, Gene, Isoform choice, Dropouts

Background

Single-cell RNA-seq (scRNA-seq) theoretically enables transcriptomic analysis at single-cell resolution. If measurements are accurate, the data would allow fundamental molecular biology questions regarding how alternative splicing is regulated at the cellular level to be addressed. However, to date, the majority of scRNA-seq studies have been analysed at the gene rather than the transcript level. Isoform quantification remains a challenging problem for bulk RNA-seq [1, 2], and we suspect that many researchers are concerned that the high degree of technical noise associated with scRNA-seq could overwhelm any biological signal from alternative splicing events. Effectively distinguishing between technical and biological noise is made all the more challenging by a lack of orthogonal methods for validating scRNA-seq. Single molecule FISH (smFISH) can be used to validate some of the predictions of scRNA-seq, but resolving between highly similar isoforms remains challenging [3–6]. Although the throughput of smFISH is improving [7], to the best of our knowledge, no smFISH technology currently exists which could accurately resolve a high proportion of the transcriptome at an isoform level.

A recent benchmark of isoform quantification for scRNA-seq found that many isoform quantification softwares perform almost as well for full-length scRNA-seq datasets as for bulk RNA-seq [8]. Whilst this is encouraging, it is important to note that the benchmark only evaluated the ability of quantification softwares to cor-

*Correspondence: jenniwestoby@gmail.com
†Martin Hemberg and Anne Ferguson-Smith contributed equally to this work.
[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH UK
[2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA UK

rectly assign simulated scRNA-seq reads to the transcripts that generated them. It is well known that a substantial amount of scRNA-seq technical noise occurs prior to the bioinformatic analysis of reads, most notably dropouts due to a low capture efficiency and PCR amplification bias due to a low amount of starting material [9–11]. To the best of our knowledge, the impact of these and other sources of technical noise on splicing analysis accuracy in scRNA-seq experiments has not been systematically studied. Consequently, the extent to which it is possible to accurately perform splicing analysis with scRNA-seq is not well understood.

It has been known for some time that not all isoforms are equally likely to be expressed from a given gene. Bulk RNA-seq studies have shown that most genes have a 'major', highly abundant isoform and will sometimes additionally have 'minor', more lowly expressed isoforms [12, 13]. It is not currently understood how isoform choice is regulated at the cellular level for most genes. In particular, it is not clear whether all cells express all isoforms but at different levels, or whether each cell exclusively expresses one or a subset of the total number of possible isoforms for a given gene. Such knowledge has the potential to contribute to a greater understanding of splicing mechanisms. For example, it is not known to what extent a common mechanism might be used to regulate isoform number at a cellular level, or whether every gene is substantially different. Several scRNA-seq studies have found that for genes which expressed multiple isoforms in bulk RNA-seq, only one or a small number of isoforms were detected in matched scRNA-seq [14–17]. However, many of these studies did not consider the impact of dropouts and quantification software errors, potentially confounding their conclusions. The deceptively simple question: 'How many isoforms are produced from a gene in a single cell?' has a central place in our understanding of molecular biology, yet its answer remains unclear.

In this study, we return to this basic biological question using a fundamentally different approach. We take real scRNA-seq datasets and select genes for which four isoforms are detected. We then use these genes to simulate the following four scenarios: (1) all cells express one isoform per gene per cell, (2) all cells express two isoforms per gene per cell, (3) all cells express three isoforms per gene per cell and (4) all cells express four isoforms per gene per cell. Importantly, in each scenario, we explicitly simulate dropout events and quantification errors. We then use the simulated output of each scenario to ask two questions. Firstly, to what extent are we able to distinguish between these global differences in alternative splicing using scRNA-seq? And secondly, what should be done to enable more accurate splicing analysis with scRNA-seq?

Results

A detailed description of our simulation approach can be found in the "Methods" section, where a brief description is given here for convenience. Our approach for the first scenario, in which we simulate one isoform being expressed per gene per cell, is to first identify genes for which the expression of exactly four isoforms is detected in a real scRNA-seq dataset. In the second step, we randomly select one isoform based on a plausible model of isoform choice for the first of our genes in the first cell in our simulated dataset. For our default model of isoform choice, we choose the isoform based on a model of alternative splicing described by Hu et al. [18]. Third, we simulate dropouts based on a Michaelis-Menten model described by Andrews and Hemberg [9]. Fourth, we simulate quantification errors based on isoform detection error estimates based on work by Westoby et al. [8]. We repeat these four steps for every four isoform gene and cell in our simulated dataset, then calculate the mean number of isoforms detected for that gene per cell. The entire process described above is one complete simulation. We run 100 simulations for each of our four scenarios, where each scenario corresponds to one, two, three or four isoforms being expressed per gene per cell. We can then plot the distributions of the mean number of isoforms detected per gene per cell for each scenario. A schematic of our simulation approach is displayed in Fig. 1. Negative control models, in which our simulations are repeated but with no dropouts and/or quantification errors are simulated, can be found in Additional file 1: Figs S1–3.

In Fig. 2, we apply our simulation approach to a dataset of H1 and H9 human embryonic stem cells (hESCs) [19, 20]. In this dataset, each cell's cDNA was split into two groups and sequenced at two different sequencing depths, enabling us to directly compare our simulation results at different sequencing depths without biological confounders. One group was sequenced at approximately 1 million reads per cell and the other group at approximately 4 million reads per cell on average. Our simulation results for the two H1 groups are compared side by side in Fig. 2a. scRNA-seq experiments have been found to saturate in terms of the number of genes detected per cell at approximately 1 million reads per cell [21, 22]. However, we observe differences in the number of isoforms detected per gene per cell at 1 and 4 million reads per cell, indicating that the saturation depth may differ for gene- and isoform-level analyses. Next, we calculate the fraction of overlap between the isoforms expressed in the ground truth and the isoforms detected as expressed in our simulations. In Fig. 2b, we show the distributions of the mean fraction of overlap for each gene. We will refer to each gene's mean fraction of overlap between isoforms expressed in the ground truth and isoforms detected as

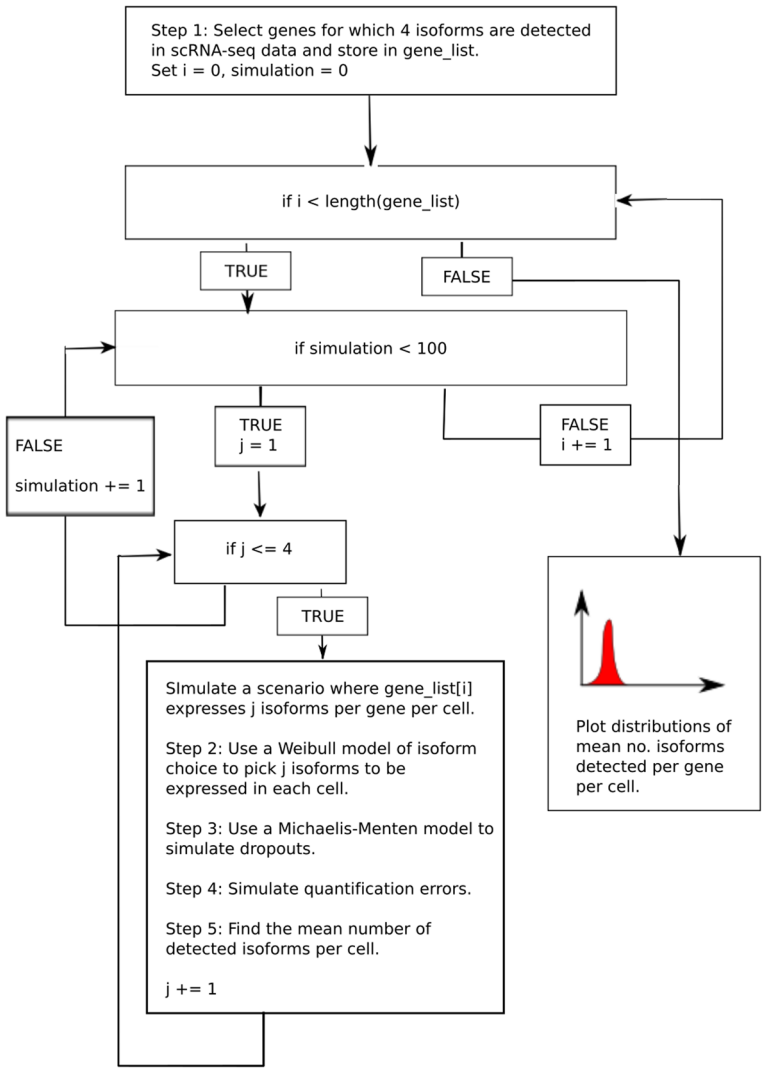

Fig. 1 Schematic of our simulation approach

expressed as the 'overlap fraction' hereafter in the text. The mean overlap fraction is consistently higher at 4 million reads per cell compared to at 1 million reads per cell, indicating that our ability to accurately detect isoforms is improved at higher sequencing depths. Similar results were observed for the H9 hESC dataset in Additional file 1: Fig S4.

Figure 2 a and b illustrate some of the difficulties associated with splicing analysis in scRNA-seq. At both sequencing depths, the distributions of the observed mean number of isoforms per gene per cell are shifted to the left of their true value. In addition, the highest mean overlap fraction observed is less than 0.8, indicating that even in a best case scenario, we fail to detect over 20% of the isoforms expressed in the ground truth. These effects are less extreme, but still present, for the group sequenced at approximately 4 million reads per cell compared to the group sequenced at 1 million reads per cell. This is consistent with the hypothesis that sequencing at higher depth reduces the extent to which isoform number is underestimated. However, even at approximately 4 million reads per cell, our simulations suggest that scRNA-seq substantially underestimates the mean number of isoforms per gene per cell for almost all genes. A naive analysis of these two datasets would most likely underestimate the number of isoforms expressed per gene per cell. This casts doubt on the biological relevance of previous observations suggesting only one isoform was typically produced per gene per cell, although admittedly the sequencing depth per cell was generally much greater than 4 million reads per cell in those studies (for example, Shalek et al. sequenced approximately 27 million reads per cell [14]).

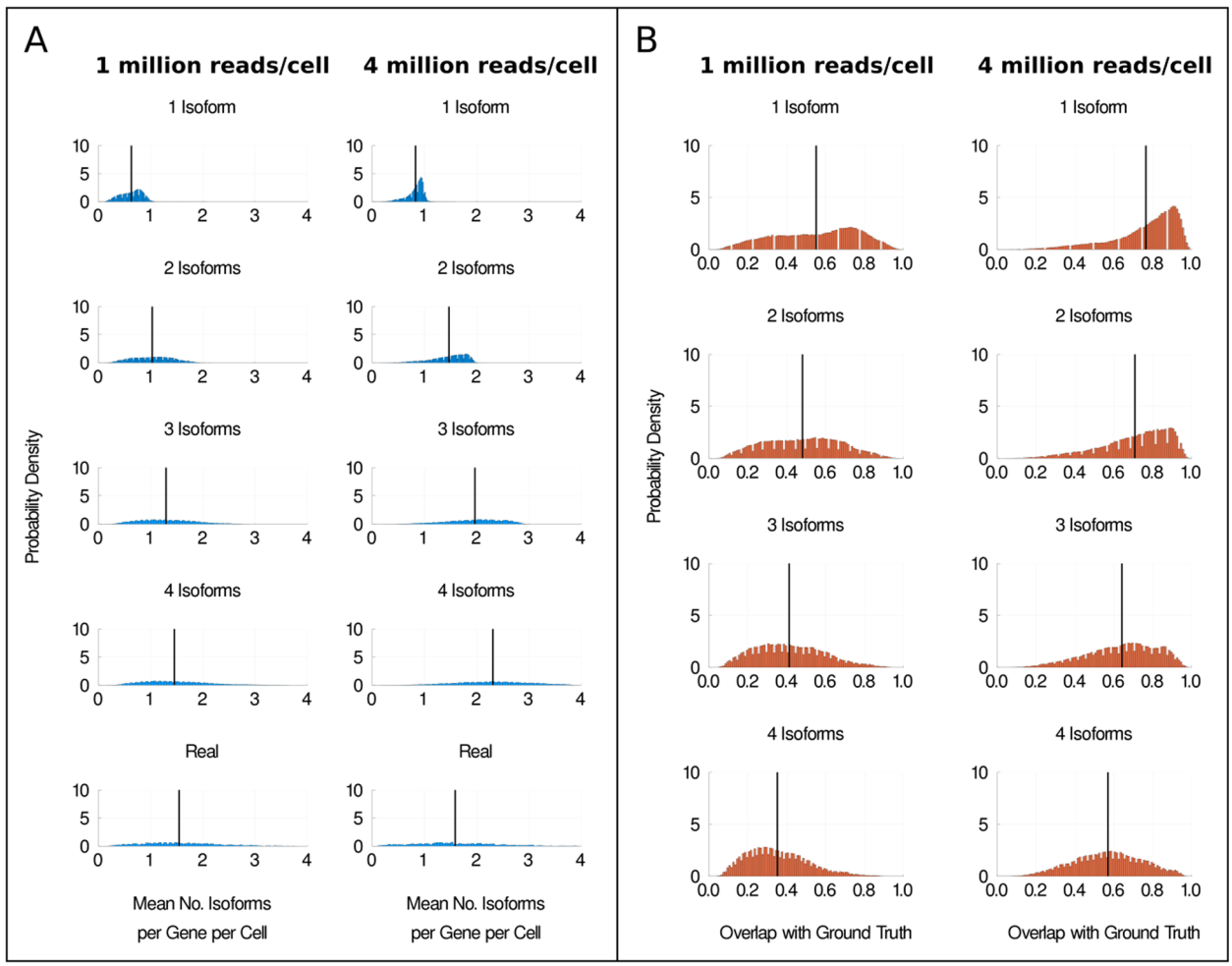

Fig. 2 The effect of sequencing depth on isoform detection. **a** Distributions of the mean number of isoforms detected per gene per cell for H1 hESCs whose cDNA was split and sequenced at approximately 1 million reads per cell or 4 million reads per cell on average. **b** Distributions of the overlap fraction. Black vertical lines represent the mean value of the distributions

One hypothesis for why our ability to detect isoforms increases with increased sequencing depth is that the rate of dropouts is reduced. In Fig. 3a, we investigate this hypothesis by plotting the distribution of the probabilities of dropout for each isoform (p(dropout)), as estimated using the Michaelis-Menten equation [9] (see the "Methods" section). We find that the distribution is skewed towards high probabilities of dropout for the group sequenced at around 1 million reads per cell. In contrast, the distribution for the group sequenced at around 4 million reads per cell is more skewed towards low probabilities of dropouts. This demonstrates that our estimated dropout probabilities are different at the two sequencing depths, as expected.

Overall, the data in Figs. 2 and 3a support the hypothesis that when the rate of technical dropouts decreases, the accuracy of isoform number estimation increases. However, as our dataset was only sequenced at two depths, we only have two data points available to investigate our hypothesis. To extend our investigation, we assume that the distributions of dropout probabilities observed in Fig. 3a can be modelled as beta distributions. The beta distribution is parameterised by two values, α and β, and we find that it approximates our probability distributions well (see bottom panels of Fig. 3a). Therefore, we select five values of α and β that generate differently shaped dropout distributions, as shown in Fig. 3b. We then perform five further simulation experiments. In each simulation experiment, we sample our dropout probabilities from one of our beta distributions. The results of these experiments are shown in Fig. 3c and d.

In Fig. 3c, we show the mean detected number of isoforms per gene per cell for the scenario where each gene produces one isoform per gene per cell. As we move from the top to the bottom of Fig. 3c, the value of α decreases, corresponding to scenarios where the

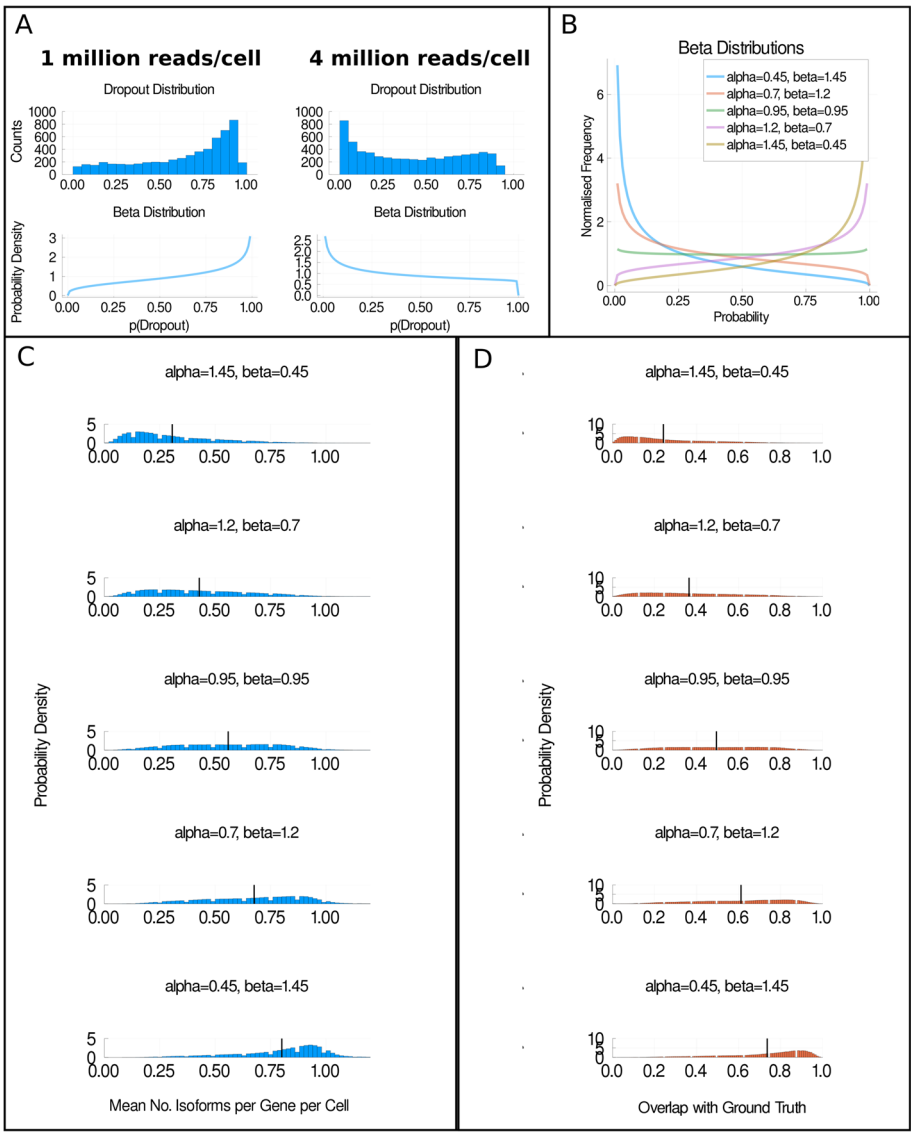

Fig. 3 The impact of dropouts on isoform detection. **a** The distribution of the probabilities of dropouts (*p*(dropout)) in each group of H1 hESCs and an approximation of these distributions using a beta distribution. At 1 million reads per cell, $\alpha = 1.31$ and $\beta = 0.74$ in the approximated beta distribution. At 4 million reads per cell, $\alpha = 0.72$ and $\beta = 1.03$ in the approximated beta distribution. **b** Five beta distributions from which dropout probabilities were sampled from the simulations used to generate **c** and **d**. In **c**, the distribution of the mean number of isoforms detected per gene per cell is shown for simulations in which one isoform was produced per gene per cell. Each plot corresponds to a simulation in which dropout probabilities were sampled from one of the distributions shown in **b**. **d** The overlap fraction for each simulation. Plots shown in **c** and **d** are for H1 hESCs sequenced at 4 million reads per cell. Black vertical lines represent the mean value of the distributions

probability of dropout is more frequently close to zero. As α decreases, the distributions of mean detected isoforms per gene per cell shift further to the right and closer to the true number of isoforms produced per cell. In Fig. 3d, we find that the mean overlap fraction increases as α decreases, corresponding to the mean probability of dropout decreasing. We conclude from Fig. 3c and d that reducing the dropout rate would likely improve the accuracy of splicing analyses performed using scRNA-seq. Similar results were observed for the H9 hESCs in Additional file 1: Fig. S5, lending further support to this conclusion.

Quantification errors are a relatively minor obstacle to studying alternative splicing

A benchmark of isoform quantification softwares in full-length coverage mouse scRNA-seq datasets found that the error rate of many software tools was low and

comparable to bulk RNA-seq [8]. This is encouraging; however, it should be noted that the error rate is likely to be substantially higher for non-model organisms with less well-annotated genomes than the mouse genome. As isoform quantification is a key step of many scRNA-seq alternative splicing analysis pipelines, it would be beneficial to understand how quantification errors impact our ability to study alternative splicing, both when the error rate is high and when the error rate is low.

As our interest in this study is the detected number of isoforms per gene per cell, we are only interested in quantification errors which lead to changes in the number of isoforms detected. We simulate two types of quantification errors, false positives and false negatives. In this context, a false positive occurs when an isoform is called as expressed by the quantification software when there are no reads from that isoform. Note that this means that if an isoform is expressed in a cell but no reads are captured from it (i.e. a dropout), but the quantification software calls it as expressed, we would define this as a false-postive event. A false negative occurs when an isoform is not called as expressed by the isoform quantification software when reads from that isoform are present. Based on our previous benchmark [8], we estimate that the probability of false-positive events (*pFP*) is around 1% and that the probability of false negative (*pFN*) events is around 4% (see the "Methods" section). In our simulations in Fig. 4, we vary both of these probabilities in the range of 0 to 50% . Figure 4a shows how the mean number of isoforms detected per gene per cell distributions changes as the probability of false positives and false negatives alters when every gene expresses one isoform per cell. Importantly, even when the probability of false positives and false negatives is zero, there are many genes for which the mean number of detected isoforms per gene per cell is not equal to one, the true number of expressed isoforms. This indicates that even if a perfect, 100% accurate isoform quantification tool existed, there would still be substantial barriers to studying alternative splicing using scRNA-seq. We suspect that the reason a 100% accurate isoform quantification tool would underestimate the number of isoforms per gene per cell is that isoform quantification tools usually only quantify the reads that are present. Due to the high number of dropouts in scRNA-seq, many expressed isoforms do not generate reads and thus would be called as unexpressed by a 100% accurate isoform quantification tool, leading to an underestimate of the number of isoforms present.

Unsurprisingly, increasing the probability of false positives causes an increase in the mean number of detected isoforms, whilst increasing the probability of false negatives causes the mean number of detected isoforms to decrease, as shown in Fig. 4b. Somewhat counterintuitively, increasing the probability of false positives from 0.0 to 0.1 could be considered to 'improve' the accuracy of the mean number of isoforms detected by shifting the distribution to slightly higher values and away from zero. This is probably because slightly increasing the probability of false positives allows some dropout events to be detected. In Additional file 1: Fig. S6, we investigate how the overlap fraction is affected by changes in the probability of false positives and negatives. We find that the overlap fraction increases as the probability of false positives increases, supporting the hypothesis that some dropout events are 'rescued' by false positive events. However, we note that in addition to 'rescuing' some dropouts, many unexpressed isoforms are also called as expressed, as indicated by the mean numbers of detected isoforms per gene per cell that are greater than one. Interestingly, when the probability of false positives and false negatives are equally increased (the diagonal of Fig. 4a), the mean number of detected isoforms increases, suggesting that the increased rate of false positives dominates over the increased rate of false negatives. This is likely because more isoforms are unexpressed than are expressed, and thus, there are more opportunities for false positive events than for false negative events. Overall, we find that high probabilities of false positives and false negatives decrease our ability to accurately detect expressed isoforms in scRNA-seq.

In Fig. 4a, we showed that even when isoform quantification is 100% accurate, we underestimate the number of expressed isoforms for many genes. One hypothesis for why we are less able to detect isoforms in scRNA-seq data compared in bulk RNA-seq data is that the sequencing depth is typically lower. A lower sequencing depth could mean that for many expressed isoforms, there are too few or no reads that would allow the expressed isoform to be uniquely identified.

To investigate whether sequencing depth could explain the difference in our ability to detect isoforms in bulk and scRNA-seq, we first identified a matched bulk and scRNA-seq dataset. The dataset we selected was a mouse embryonic stem cell (mESC) dataset in which mESCs were cultured in 2i + LIF media [23, 24]. In the mESC dataset, each cell was sequenced to approximately 7 million reads on average, whilst the matched bulk data was sequenced to approximately 44 million reads.

To determine whether sequencing depth was responsible for the difference in our ability to detect isoforms in bulk and scRNA-seq, we randomly downsampled the bulk mESC RNA-seq dataset to 7 million reads 50 times. Using the original, un-downsampled bulk RNA-seq dataset as the ground truth, in Additional file 1: Fig. S7, we plotted the mean overlap fractions for each gene in the downsampled bulk RNA-seq dataset and the matched scRNA-seq dataset. We found that the mean overlap fraction was significantly higher ($p < 2.2e{-}16$, Welch two sample t test) for the downsampled bulk RNA-seq than for the matched

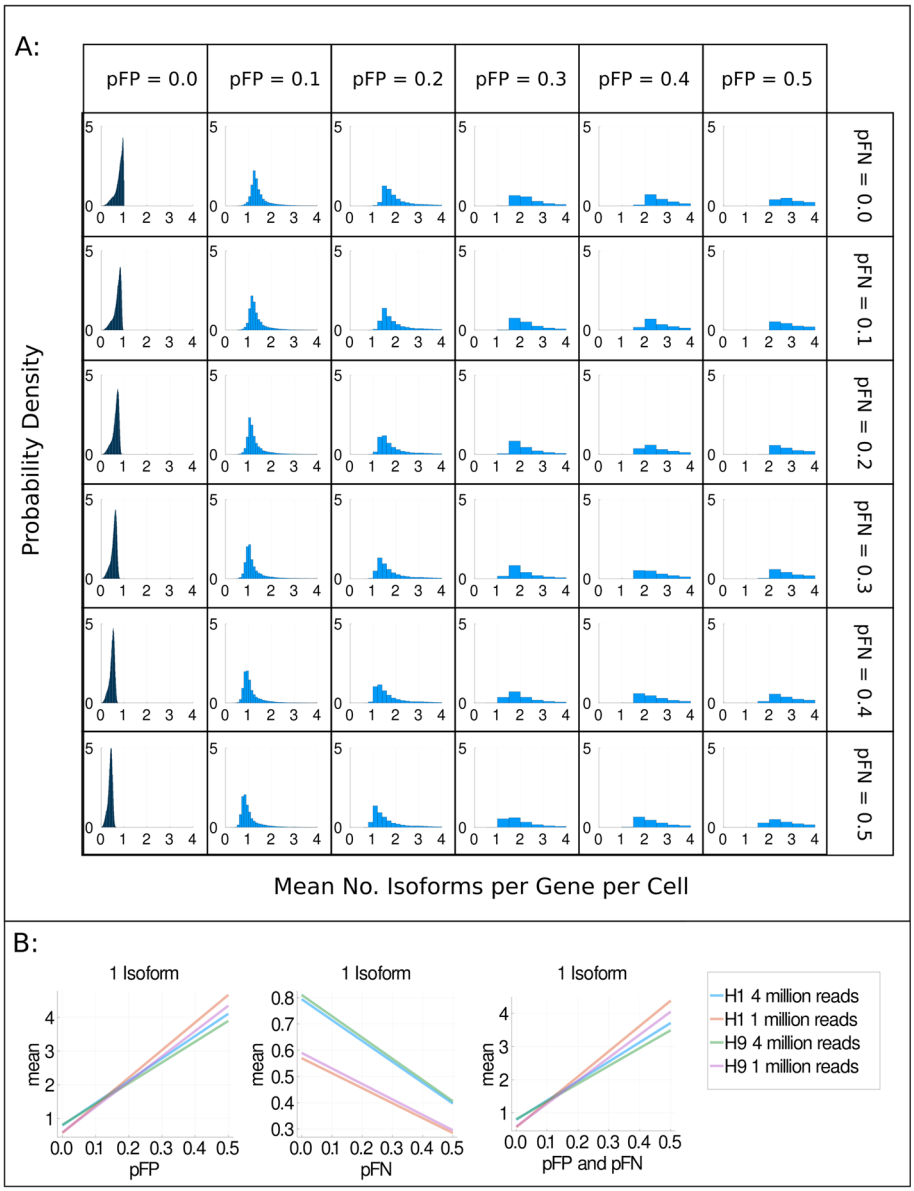

Fig. 4 The impact of quantification errors on isoform detection. **a** Distributions of the mean number of isoforms detected per gene per cell when one isoform is expressed per gene per cell. The probability of false positives (*pFP*) increases from left to right, and the probability of false negatives (*pFN*) increases from top to bottom. The dataset shown is H1 hESCs whose cDNA was split and sequenced at approximately 4 million reads per cell on average. **b** Summary plots of the average of the mean number of isoforms detected per gene per cell when *pFP*, *pFN*, or *pFP* and *pFN* are increased

scRNA-seq. This indicates that a lower sequencing depth does reduce our ability to detect isoforms, but that this does not fully explain the reduction in ability to detect isoforms between bulk and scRNA-seq. One explanation for the reduction in ability to detect isoforms in scRNA-seq, over and above the reduction expected due to reduced sequencing depth, is that there could be heterogeneous isoform expression between individual cells. If this were the case, using the isoforms detected in bulk RNA-seq as the ground truth would not be appropriate. There are also potential technical explanations for the reduced ability to detect isoforms using scRNA-seq. For example, the enzymatic reactions associated with library preparation may have reduced efficiency when there is a lower amount of starting material, as is the case for scRNA-seq. Determining to what extent heterogeneous isoform expression and technical factors are responsible for our reduced ability to detect isoforms in scRNA-seq will require further study of cellular isoform heterogeneity and the technical noise associated with scRNA-seq.

Different models of isoform choice meaningfully change our simulation results

It is possible that different mechanisms of isoform choice at the cellular level could alter our ability to correctly detect which isoforms are present in scRNA-seq. Because there is uncertainty over the mechanism of isoform choice within single cells, we implement four different models of isoform choice in our simulations. We then ask whether different models of isoform choice alter the mean number of detected isoforms per gene per cell in our simulations.

We give a detailed description of how each of these models was implemented in the "Methods" section; here, we provide a brief description of each model and the rationale behind it. We first model the alternative splicing process as a type III Weibull distribution, using a model described by Hu et al. [18]. Based on observations about the molecular process of alternative splicing, Hu et al. suggested that the process could be well modelled by an extreme value distribution, and they found that a Weibull distribution best fit the expression levels of isoforms in bulk RNA-seq. In our second implemented model, we attempt to infer the probability of each isoform being 'chosen' to be expressed in a cell. We calculate the probability of an isoform being chosen based on the observed probability of the isoform being detected. Our third model is identical to the second except that we allow the probability of an isoform being 'chosen' to vary between cells. We achieve this by sampling the probability of an isoform being chosen from a beta distribution, using a similar approach as Velten et al. [4]. In our final model, we choose a random number between 0 and 1 for each isoform. The random number is assigned to be that isoform's probability of being chosen, weighted against the probabilities of the gene's other isoforms being chosen. For brevity, we will refer to these four models as the Weibull model, the inferred probabilities model, the cell variability model and the random model below.

Figure 5 shows the distributions of the mean number of detected isoforms when one, two, three or four isoforms are expressed per gene per cell for each model. Figure 5 shows our simulation results for the H1 hESC dataset sequenced at 4 million reads; results for the other hESC datasets including distributions of overlap fractions can be found in Additional file 1: Figs. S8–14. Importantly, the distributions in Fig. 5 visibly differ between models. To quantitatively confirm this, we perform a K-sample Anderson-Darling test on each row of graphs in Fig. 5. We find that the distributions for 1, 2 and 3 isoforms significantly differ between the isoform choice models ($p<0.001$, see Additional file 1: Supplementary Tables for details). In contrast, the distributions for 4 isoforms have a p value of 0.999999, consistent with these distributions originating from the same population. This is as expected, as in the 4 isoform simulations all of the isoforms are picked, and thus, we would not expect isoform choice to matter. Our qualitative and quantitative analyses indicate that different mechanisms of isoform choice alter our ability to detect splice isoforms in scRNA-seq. Therefore, a better understanding of the mechanism of isoform choice across the transcriptome could be key to enabling splicing analysis using scRNA-seq data. Without knowing how best to model isoform choice, our results suggest the presence of a substantial confounder.

Interestingly, our simulation results when using the inferred probability model compared with the cell variability model are almost identical. Given that the only difference between these models is whether or not isoform preference is allowed to vary between cells, this indicates that cellular heterogeneity in isoform preference does not change our ability to detect isoforms under the inferred probability model. We perform a K-sample Anderson-Darling test between the inferred probabilities and cell variability models for each row of Fig. 5, and we find that these distributions do not significantly differ (see Additional file 1: Supplementary Tables). Interestingly, the results of the random model of isoform choice look more like the inferred probability and cell variability models than the Weibull model. This could be because the Weibull model determines the probability of an isoform being chosen based on the rank of that isoform, whereas all of the other models do not use a rank-based approach. These observations and the difficulty we have interpreting them illustrate the need for a better understanding of how best to model isoform choice.

We hypothesise that the reason that different models of isoform choice differ in ability to detect isoforms could be because some models of isoform choice prefentially pick isoforms with a low probability of dropout, whereas other models do not exhibit this preference. To investigate whether different models of isoform choice differ in their preference for picking isoforms with a low probability of dropout, in Additional file 1: Figs. S15–18, we plot the distributions of the probabilities of dropout for the isoforms chosen when one, two, three or four isoforms are picked using each of our four models. We would expect models with a preference for picking isoforms with a low probability of dropout to have distributions of dropout probabilities more skewed towards zero when small numbers of isoforms are chosen. When larger numbers of isoforms are chosen, we would expect to observe less skewed distributions, because the model is effectively forced to choose isoforms with higher probabilities of dropout due to a lack of alternatives. In contrast, if a model had no preference for picking isoforms with a low probability of dropouts, we would expect the distributions of the probabilities of dropout to be identical regardless of whether one, two, three or four isoforms are chosen.

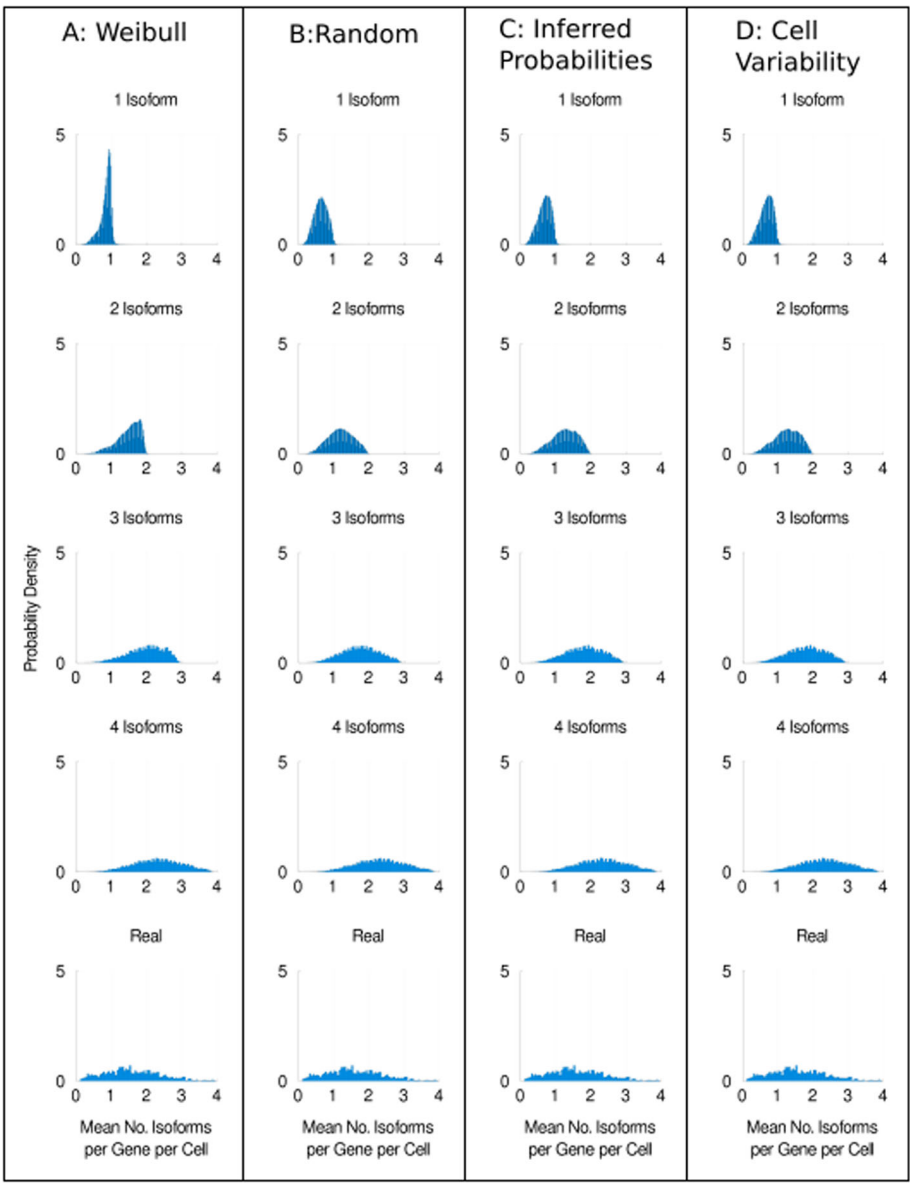

Fig. 5 Different models of isoform choice alter our ability to detect isoforms. **a** Distributions of the mean number of isoforms detected per gene per cell for H1 hESCs sequenced at approximately 4 million reads per cell using the Weibull model of isoform choice. **b** The same distributions when the random model is used. **c** The distributions when the inferred probabilities model is used. **d** The distributions when the cell variability model is used. See the main text for a detailed description of each model

In Additional file 1: Figs. S15–18, we find that only the random model does not exhibit any preference for choosing isoforms with a low probability of dropout. Of the Weibull, inferred probability and cell variability models, the Weibull model has the dropout probability distribution most skewed towards zero when one isoform is picked, indicating that the Weibull model has the strongest preference for picking isoforms with a low probability of dropout. The Weibull model also detects the highest mean number of isoforms per gene per cell when one isoform is expressed in the ground truth, consistent with the hypothesis that the difference in the performance of the isoform choice models may be related to their preference for picking isoforms with a low probability of dropout.

If isoform detection ability of the isoform choice models is mainly determined by their preference for picking isoforms with a low probability of dropout, we would expect that if the probability of dropout was globally changed, it would alter the isoform choice models' abilities to detect

isoforms. We investigate this in Additional file 1: Fig. S19 by sampling dropout probabilities from the beta distributions shown in Fig. 3b. We find that more isoforms are detected by all isoform choice models when dropouts are sampled from distributions that are more skewed towards zero. This supports the hypothesis that choosing isoforms with a low probability of dropout improves the ability of isoform choice models to accurately detect isoforms.

Some models of isoform choice are more plausible than others

In the previous section, we observed that our simulation results for the inferred probability and cell variability models were extremely similar. To investigate how general our observation that allowing isoform preference to vary between cells does not alter our simulation results is, we developed three additional models of isoform choice. In the first model, the probability of selecting each isoform was sampled from a truncated normal distribution with a mean of 0.25 and a standard deviation of 0.06 in each cell. In the second model, we sample the probability of selecting each isoform from a Bernoulli distribution, in which the value 1 is chosen 25% of the time and the value 0 is chosen 75% of the time in each cell. In the final model, the probability of selecting each isoform is always 0.25 (the '$p = 0.25$' model). The three models are illustrated in Fig. 6a, and additional details are given in the "Methods" section. Under the normal and the Bernoulli models, the probability of picking each isoform varies between cells, whereas the probability of picking each isoform is constant between cells under the $p = 0.25$ model. Importantly, although the distributions we are sampling isoforms from have very different shapes, the mean probability of picking each isoform is 0.25 for all three distributions.

In the second row in Fig. 6a, we show the distribution of the mean number of isoforms detected per gene per cell when we simulate one isoform being expressed per gene per cell. There is no visible difference between our simulation results regardless of which model of isoform choice is used. This is supported by a non-significant result in a K-sample Anderson-Darling test ($p = 0.998$). These findings are consistent with the hypothesis that our simulation results are unchanged whether or not the model of isoform choice used allows cell variability in isoform choice. We suggest that this is because we are reporting the mean number of isoforms detected per gene per cell in our simulations. Across many cells and rounds of simulation, the mean probability of selecting isoforms seems to determine the shape of our simulation result distributions, whereas the higher moments of the isoform choice probability distribution are apparently unimportant. Thus, including cell variability in our isoform choice model appears to not matter. For future scRNA-seq studies in which the mean number of isoforms detected per gene per cell is an important metric, we conjecture that there is no need to model cellular variability in isoform choice, regardless of whether or not such variability exists in reality. Of course, if future studies are interested in precisely what isoforms are present in individual cells rather than a population mean, understanding whether or not cell variability in isoform choice exists is likely to be important.

We have established that our ability to detect isoforms using scRNA-seq is severely affected by the high rate of dropouts in scRNA-seq. Therefore, attempts to infer a biologically meaningful model of isoform choice from scRNA-seq data are likely to fail. However, we can make some general observations to help rule out certain models of isoform choice. In Fig. 6b, we have ranked isoforms by their mean expression relative to other isoforms from the same gene (so for example, an isoform with rank 1 has the highest mean expression, an isoform with rank 2 has the second highest mean expression and so on). Unsurprisingly, we find that the most highly ranked isoforms are substantially more highly expressed than lowly ranked isoforms. This is consistent with the finding that many genes appear to have a 'major', more highly expressed isoform, and one or more 'minor', less highly expressed isoform [12, 13]. We suggest that this behaviour needs to be represented in some way in future models of isoform choice, and models that do not represent it (for example, our random, normal, Bernoulli and $p = 0.25$ models) are probably overly simplistic. In Fig. 6c, we rank isoforms by their probability of dropout, where the isoform with the lowest probability of dropout compared to other isoforms from the same gene has rank 1. We observe a very similar pattern in which highly ranked isoforms have a substantially lower probability of dropout relative to lowly ranked isoforms, further supporting the finding that 'major' and 'minor' isoforms exist for many genes. The results shown in Fig. 6 are for the H1 hESCs sequenced at 1 million reads per cell; equivalent plots and overlap fraction distributions for all of the hESC datasets can be found in Additional file 1: Figs. S20–24.

A mixture modelling approach suggests genes for which four isoforms are detected typically express around three isoforms per cell

We ask whether our simulation-based approach could shed any light on the biological question of how many isoforms are expressed per gene per cell. To do this, we simulate one, two, three and four isoforms being expressed per gene per cell and compare the mean isoforms detected distributions to the distribution of isoforms detected per gene per cell for genes for which four isoforms were detected in the real dataset (see Fig. 7a and b). We then approximate each distribution as a log normal distribution and take a mixture modelling approach to estimate

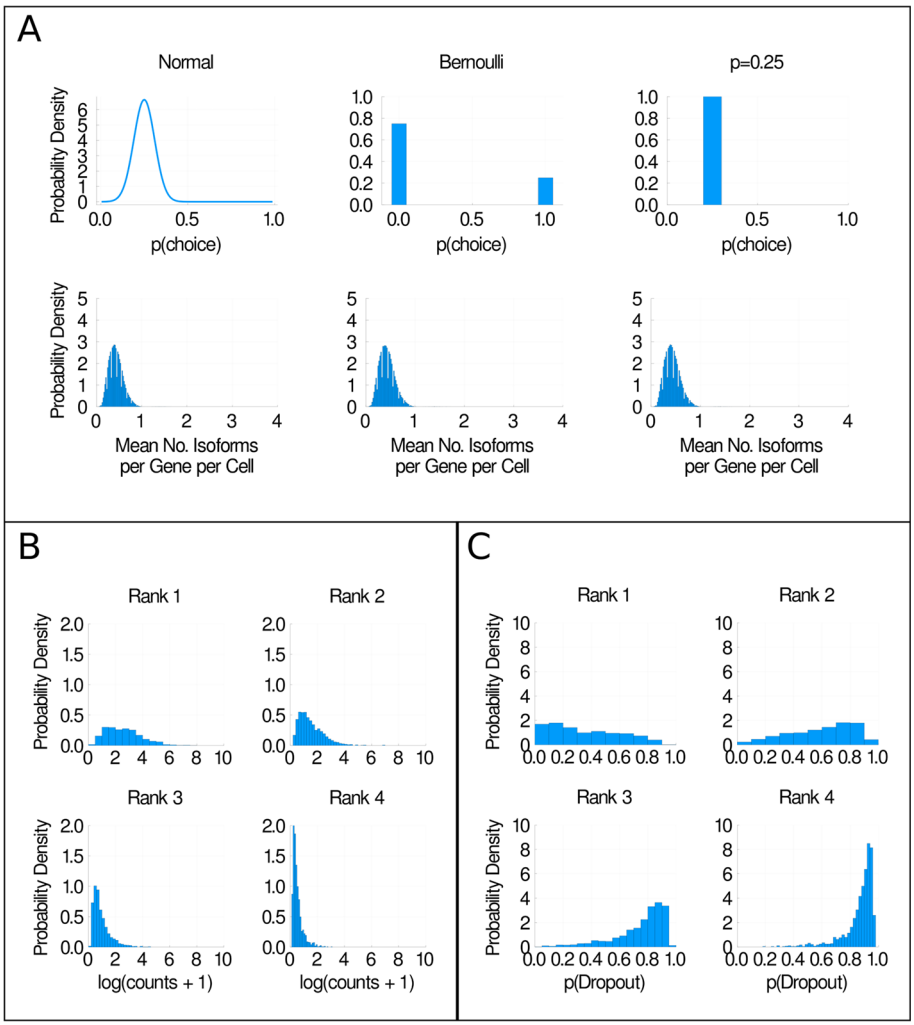

Fig. 6 Some models of isoform choice are more plausible than others. **a** We model the probability of picking any given isoform as a normal distribution, a Bernoulli distribution and a constant probability, all with the same mean (0.25) (top row of graphs). In the bottom row, we show the distributions of the mean number of isoforms per gene per cell detected when each model of isoform choice is used. **b** Histograms of mean isoform expression, ordered by isoform rank. **c** Histograms of dropout probability, ordered by isoform rank. All plots shown are for H1 hESCs sequenced at 1 million reads per cell

the mixing fraction for each of our simulated distributions in the real distribution.

Figure 7c shows the mixing fractions found over 100 iterations of expectation maximisation for H1 hESCs sequenced at approximately 1 million reads per cell. In Fig. 7c, the mixing fraction for the distribution corresponding to four isoforms being expressed per gene per cell is over 90%. This suggests that genes detected to express four isoforms in this dataset typically express four isoforms per gene per cell. However, in Fig. 6d, after 100 iterations of expectation maximisation for H1 hESCs sequenced at 4 million reads per cell, the distribution with the largest mixing fraction is that corresponding to three isoforms per gene per cell. This suggests that genes detected to express four isoforms in this dataset most often express three isoforms per gene per cell. As the cDNA sequenced at 1 and 4 million reads per cell came from the same population of cells, it is unlikely that both of these statements are true. We propose several possible explanations for why we might observe this result.

First, we might be over-estimating the dropout rate at 1 million reads per cell. As there is less information with which to infer the dropout rate at 1 million reads per cell compared to at 4 million reads per cell, it is plausible that our estimates of the dropout rate are less accurate at 1 million reads per cell. Whether or not there is a systematic bias towards over-estimating the dropout rate at low sequencing depths is unknown and goes beyond the scope of this paper.

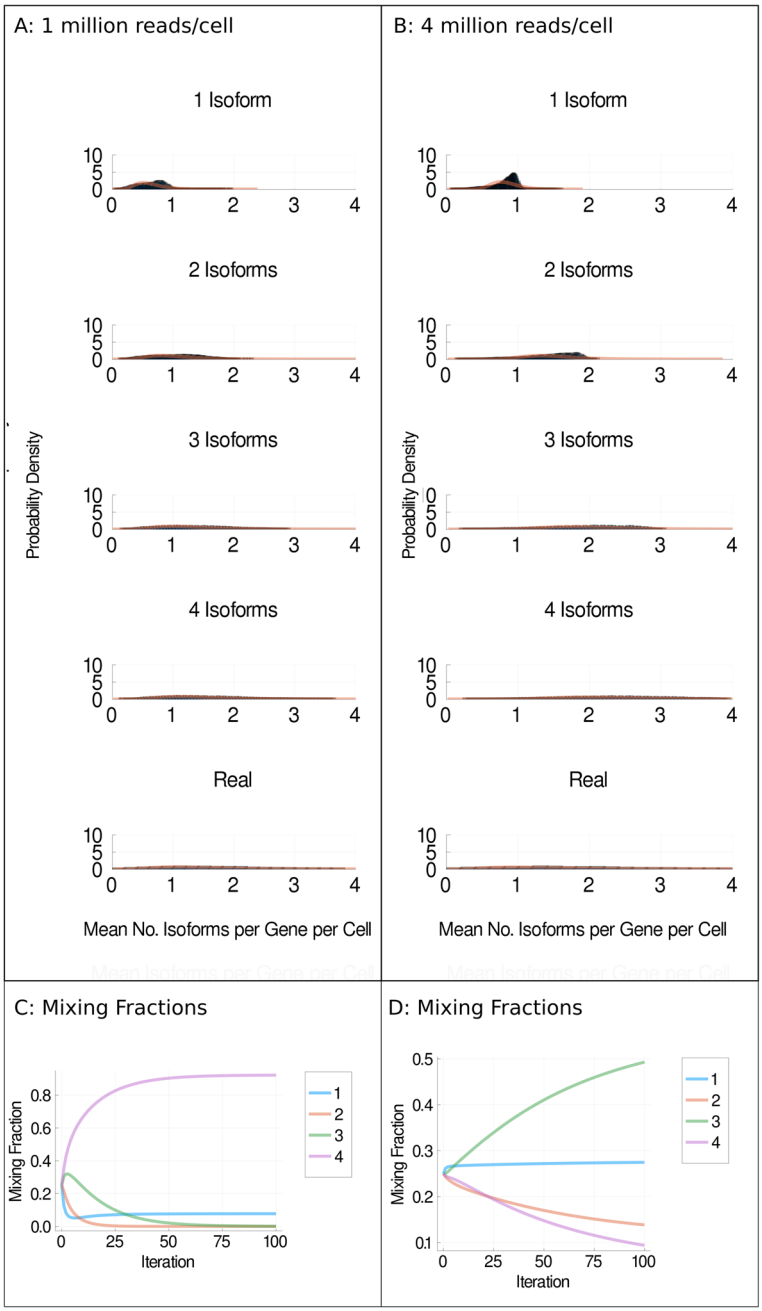

Fig. 7 Mixture models. **a**, **b** Distributions of detected isoforms per gene per cell (blue) and log normal fitted distributions (orange) for H1 cells sequenced at 1 million reads per cell (**a**) or 4 million reads per cell (**b**) under the Weibull model. **c**, **d** Mixing fractions vs iterations of expectation maximisation for 1 million reads per cell (**c**) and 4 million reads per cell (**d**). Each coloured line represents the distributions for one, two, three or four isoforms being simulated as expressed per gene per cell. Equivalent plots for other isoform choice models and H9 cells can be found in Additional file 1: Figs. S25–31

Second, we have established that the model of isoform choice influences the outcome of our simulations but we do not know which model of isoform choice is correct. Therefore, we are (almost certainly) attempting to fit distributions that do not represent reality. Figure 5 shows our mixture modelling approach using the Weibull model of isoform choice. We note however that fitting our alternative models of isoform choice achieves a similar result, in that the largest mixing fraction goes to four isoforms at 1 million reads per cell and to three or fewer isoforms at 4 million reads per cell (see Additional file 1: Figs. S25–31).

Third, the genes detected to express four isoforms differ between the sequencing depths of 1 and 4 million reads. More genes are detected to express four isoforms at 4 million reads (1443 versus 1543 for the H1 cells, 1453 versus 1524 for the H9 cells). Whilst this is not a dramatic difference, it does mean that the mixing fractions between these two depths could genuinely differ, although this is unlikely to fully explain the observed difference.

Fourth, we assume all genes for which four isoforms are detected in the real data actually express four isoforms. Due to dropouts and quantification errors, this may not be accurate, and some genes for which four isoforms are detected may express a different number of isoforms in reality.

Fifth, our parameter estimation for quantification errors and isoform choice modelling is not 100% accurate. We can not rule out that this could be confounding the results of our mixture modelling approach.

Our mixture modelling experiments broadly support the hypothesis that it might be common for a cell to produce more than one isoform per gene. However, there are clearly a lot of potential confounders in our approach, many of which relate to uncertainty about dropouts, quantification errors and isoform choice. We note that without having either a ground truth knowledge of how many isoforms are produced from given genes in given cells, or good estimates of dropout probabilities, quantification errors and isoform choice mechanism, it is hard to imagine how an accurate and reliable estimate of the number of isoforms produced per gene per cell could be obtained.

Discussion

In this study, we use a novel simulation-based approach to ask whether it is possible to study alternative splicing at the level of individual cells using scRNA-seq. In our simulations, we simulate four scenarios in which every gene produces one, two, three or four isoforms per gene per cell. That we struggle to clearly distinguish between these four situations emphasises the challenges associated with distinguishing the much subtler and more complex patterns of alternative splicing that likely exist in reality. Whilst scRNA-seq is capable of detecting some splicing events, confounding due to dropouts means we are likely to underestimate the number of splicing events occurring in individual cells.

We next ask what limitations must be overcome to make alternative splicing analysis possible using scRNA-seq. We find that reducing the probability of dropouts improves our ability to accurately detect isoform number. Therefore, reducing the frequency of dropouts could be one method to improve the accuracy of splicing analyses in scRNA-seq. To some extent, this could be achieved by sequencing cells more deeply, although we note that at 4 million reads per cell we still substantially underestimate isoform number in the H1 hESCs. Unfortunately, extremely deeply sequenced datasets (e.g. >10 million reads per cell) are likely to suffer more with PCR artefacts and potentially a higher false positive rate of isoform detection [11, 25]. Fundamentally, the low capture efficiency of scRNA-seq is likely a consequence of a small amount of starting material. This can probably be rescued to some extent by more PCR cycles and sequencing at higher depths; however, we would not expect this to fully solve the problem.

A more radical way to overcome confounders due to dropouts would be if scRNA-seq technologies changed in some fundamental way that increased capture efficiency. Whether this is feasible is unclear. Alternatively, we note that if we could estimate the probability of dropout for each isoform more accurately, in theory, it should be possible to correct for confounding due to dropouts in splicing analyses. Therefore, to enable splicing analysis using scRNA-seq, either the capture efficiency of the technology needs to improve or more work characterising the probability of dropouts at an isoform level is required.

We note that in our study, we exclusively consider the impact of technical dropouts on isoform detection. However, it is known that many genes are heterogeneously expressed, whether due to 'bursty' transcription or cell type-specific expression [26]. Ideally, the impact of biological dropouts on isoform detection would be evaluated alongside the impact of technical dropouts. Unfortunately, to the best of our knowledge, there is currently no reliable methodology to distinguish between biological and technical dropouts. The goal of imputation approaches is to identify and correct for technical dropouts, but a recent benchmark found that imputation approaches often introduce a high rate of false positive results [27]. This indicates that the problem of distinguishing between biological and technical dropouts is not yet solved. As it is not currently possible to resolve between biological and technical dropouts, it is also challenging to accurately model biological dropouts, as little is known about their prevalence and how the frequency of biological dropouts might vary with genomic features. We hope that future work in this space will enable more accurate identification of biological and technical dropouts, thus enabling studies such as ours to be extended to account for biological as well as technical dropouts.

Long read technologies could in theory enable 100% accurate isoform quantification, if issues due to a high base calling error rate could be overcome [28]. However, we find that even when no isoform detection errors occur, our ability to accurately detect isoforms is very limited. Therefore, long read technologies or isoform quantification software improvements alone are not sufficient to

enable accurate splicing analysis in scRNA-seq. In addition, we note that at present, the read throughput of long read platforms is too low to enable meaningful isoform detection and quantification across a large number of cells [29]. A more immediate way in which long read technologies could improve isoform quantification accuracy is by using long read technologies to improve transcriptome annotations. In many non-model organisms, a high proportion of isoforms are missing from reference transcriptomes, making the problem of isoform detection and quantification substantially harder. Long read approaches combined with tissue-specific transcriptome curation could dramatically improve isoform quantification accuracy in poorly annotated organisms. More accurate isoform detection and quantification would in turn improve our ability to gain biological insight from sequencing data collected from these organisms.

A limitation of this study is that our approach for simulating quantification errors is very simplistic. In particular, we assume that the probability of a false positive or a false negative event is constant and does not depend on the GC content, length, magnitude of expression or any other relevant features of the isoform being simulated. In reality, the probability of isoform detection errors probably does depend on factors such as GC content and how highly expressed the isoform is. However, relatively little research has been done into the relationship between features of isoforms, such as GC content and magnitude of expression, and the probability of isoform detection errors. Further research into how genomic and other features of isoforms affect the likelihood of isoform detection and quantification errors would enable more accurate error models to be built in future. This would be valuable both in studies such as this one and more generally, as it would enable more sophisticated error correction models to be developed.

Little is known about the biological process of isoform choice in individual cells for most genes. Thus, accurately modelling this process is challenging. We find that different models of isoform choice alter our simulation results. This indicates that without better understanding of the process of isoform choice, alternative splicing analyses are potentially confounded by this unknown factor. Research into the process of isoform choice within individual cells across the transcriptome would enable more accurate models of isoform choice to be built, reducing or removing this confounder from future alternative splicing analyses. An important finding from our study is that the ability of isoform choice models to accurately detect isoforms is correlated with the preference of isoform choice models for choosing isoforms with a low probability of dropout. It would therefore be highly relevant to establish whether cells have a preference for expressing isoforms with a low probability of dropout. Isoforms with a low probability of dropout are in practice usually isoforms which are highly expressed. Therefore, if cells have a preference for expressing highly expressed isoforms with a low probability of dropout, we would expect it to be relatively easy to accurately detect how many isoforms are expressed in individual cells. In contrast, if it is common for cells to express lowly expressed isoforms with a high probability of dropout, we would expect it to be much harder to accurately detect the number of expressed isoforms using scRNA-seq. Establishing which scenario is more biologically relevant would therefore be highly valuable to the single-cell community.

It is important to note that the probabilistic models of isoform choice used in our study are unlikely to be realistic models of isoform choice for two reasons. Firstly, we know little about the underlying biological process of isoform choice for most genes. Therefore, at best, the models we have devised in this study are educated guesses as to what the true underlying process might be. Secondly, it is likely that the isoforms chosen by our isoform choice models will have an impact on the probability of a quantification error occurring. Different isoforms have different read generation biases and will generate reads with different mapping properties. In our simulations, we have not modelled the impact of, for example, different splice junction abundances on our ability to detect isoforms, although factors such as this are likely to have an impact on our ability to detect isoforms. We would welcome future studies addressing the more nuanced issues associated with the interplay between isoform choice and quantification errors, although we believe that a better understanding of how to accurately model isoform choice and quantification errors would be a prerequisite to such studies. If isoform expression is found to be heterogeneous between cells, interplay between isoform choice and isoform quantification errors could partly explain why we were less able to detect isoforms present in mESC scRNA-seq data than in downsampled bulk RNA-seq.

We observe that when studying the mean number of isoforms detected per gene per cell, it appears to be unimportant whether or not there is cell variability in isoform choice from a modelling perspective. Of course, if the goal is to accurately detect which isoforms are present in each cell, establishing whether cell variability exists and modelling any variability will be essential. However, we note that imputation remains challenging and often inaccurate at the gene level for scRNA-seq [27]. We therefore anticipate it will be some time before accurate imputation is feasible at the isoform level.

We are able to detect evidence in support of 'major' and 'minor' isoforms, and propose that future models of isoform choice should attempt to capture this behaviour. However, we note that whilst our observations help discard models of isoform choice, we believe that scRNA-seq

is currently too confounded by dropouts to accurately infer a model of isoform choice at the single-cell level. We suggest that smFISH would be a more appropriate technology to investigate how isoform choice is regulated in individual cells. Indeed, smFISH has previously been used to study alternative splicing and isoform choice in individuals cells for a small number of genes [3, 4, 6]

The results of our mixture modelling experiments are consistent with multiple isoforms being produced per gene per cell; however, we note that our mixture modelling experiments are heavily confounded by a lack of understanding about dropouts, isoform choice and perhaps quantification errors to a lesser extent. Therefore, we argue that at this time, scRNA-seq will not be able to provide the answer to basic biological questions about how many isoforms are produced per gene per cell.

We note that in our study, we exclusively focused on full-length scRNA-seq datasets collected using the SMARTer or SMART-seq2 protocols [19, 23, 30, 31]. Our rationale for using full-length scRNA-seq data is that full-length library preparation protocols have reduced 3′ bias relative to UMI-based protocols [30, 31]. Consequently, we would expect to be more able to accurately resolve between isoforms using a full-length protocol compared to a UMI-based protocol. However, a disadvantage of full-length protocols is that they do not usually contain UMIs, meaning that we have limited ability to correct for PCR amplification bias. A new library preparation protocol called SMART-seq3 was recently developed by Hagemann-Jensen et al. which generates both full-length reads and UMI-containing reads [32]. It is currently unclear whether information from the full-length reads and the 3′ biased UMI-containing reads could be utilised in some way to combine the advantages of reduced coverage bias and PCR bias correction in the context of isoform detection. Clearly, if this is feasible, it would be highly relevant to solving some of the problems associated with isoform quantification using scRNA-seq data.

We have exclusively considered the problem of isoform detection using isoform quantification tools in this study. We have chosen to use isoform quantification software in preference of exon centric approaches, such as the approach used by MISO [33], because an independent benchmark of the performance of isoform quantification tools run on scRNA-seq data has been performed [8]. To the best of our knowledge, there is no independent benchmark of the performance of exon centric approaches run on scRNA-seq data. As most exon centric approaches were designed for bulk RNA-seq, this is potentially problematic as it is unclear whether existing exon centric software gives accurate results when run on scRNA-seq data. Consequently, we have focused exclusively on isoform quantification software in this study. However, we hypothesise that dropouts are also likely to be a confounder when studying scRNA-seq using exon centric approaches.

In addition to detecting isoforms, isoform quantification tools attempt to determine how highly expressed isoforms are. Isoform quantification is a substantially harder problem than isoform detection. Due to uncertainties over how highly expressed isoforms are in individual cells, how best to model PCR amplification bias and differences in library sizes between individual cells and how best to incorporate relative expression into a model of isoform quantification errors, we suspect isoform quantification is also likely to be substantially harder to model than isoform detection. For these reasons, we have focused on isoform detection in this study, but suggest that future work investigating our ability to detect the relative expression of isoforms would be highly valuable to the field. We note that although we have not directly evaluated our ability to resolve the relative expression magnitude of isoforms in this study, that we often struggle to accurately detect isoforms implies that we would often struggle to determine how highly expressed they are.

Based on our findings in this study, at this time, we do not recommend attempting alternative splicing analysis using scRNA-seq. However, we make actionable suggestions for how splicing analysis could be enabled in the future. An improved understanding of the prevalence of technical dropouts at the isoform level could enable us to reduce confounding due to dropouts. Improvements to the capture efficiency of scRNA-seq would similarly reduce confounding. Increased study of isoform choice at the single-cell level using technologies such as smFISH would enable better models of isoform choice to be generated, eliminating confounders. Although we find quantification errors to be a relatively small confounder, further reducing quantification errors using long read technologies and more accurate quantification tools would be welcome. Although we have concluded that accurate alternative splicing analysis with scRNA-seq is not possible today, we are optimistic that it could become possible in the near future.

Conclusions

At present, alternative splicing analyses using scRNA-seq are substantially confounded. Better characterisation of dropouts or improvements in capture efficiency would reduce confounding due to dropouts. Further research into the process of isoform choice at a single-cell level would reduce confounding due to a lack of knowledge about isoform choice. Quantification errors are a relatively minor confounder, although improvements in this area are still welcome. At present, to the best of our knowledge, a large-scale unconfounded analysis of the number of isoforms produced per gene per cell has not been performed.

Therefore, we still do not know how many isoforms are typically produced per gene per cell.

Methods

Data preprocessing

Our simulation approach requires an isoform-cell count matrix as input. To generate isoform-cell count matrices, we used Kallisto to quantify reads from each cell against the Gencode mouse vM20 transcriptome for our mESC datasets and against the Gencode human v20 transcriptome for our hESC dataset [34, 35].

Simulation approach

Our simulation approach is summarised as an algorithm below.

Algorithm 1: Our simulation approach

```
Step 1: Select genes for which four isoforms are
detected in scRNA-seq data
for simulation in 1:100 do
    for gene in DetectedGenes do
        for i in 1:4 do
            for j in 1:NumCells do
                Step 2: Choose i isoforms to be
                expressed the jth cell based on isoform
                choice model
                Step 3: Introduce dropouts based on
                Andrews and Hemberg's
                Michaelis-Menten model
                Step 4: Introduce isoform
                quantification errors
            end
            Step 5: Find mean number of isoforms per
            gene per cell.
        end
    end
end
Step 6: Plot distributions of mean number of isoforms
per gene per cell (eg. as in Fig. 1)
```

We expand upon each step below.

Step 1: Select genes for which four isoforms are detected in scRNA-seq data

Our simulation approach takes an isoform-cell count table as input. We define an isoform as detected if it has more than five counts in at least two cells. We select genes for which exactly four isoforms pass this threshold.

Step 2: Choose i isoforms to be expressed the jth cell-based on the isoform choice model

In this step, we probabilistically choose i isoforms to be expressed in each cell, where i is one, two, three or four. The default model used in this study was the Weibull model, which was used to produce all of our main figures unless otherwise stated.

The Weibull model In [18], Hu et al. found that the median frequency, $mf(k, M)$, of the kth dominant isoform of a gene with M detected isoforms can be described as:

$$mf(k, M) = \frac{1}{k \times H_M} \exp\left[-\left(1 + \frac{k}{M}\right)^2\right]$$

where H_M is the Mth generalised harmonic number:

$$H_M = \sum_{m=1}^{M} \frac{1}{m} \exp\left[-\left(1 + \frac{m}{M}\right)^2\right]$$

In our implementation of this model of isoform choice, we first rank the isoforms in order of magnitude expression for each gene, with the most highly expressed isoform having rank 1, the second most highly expressed isoform having rank 2 and so on. We calculate the magnitude of expression by summing the total number of counts across all cells for that isoform. We then use the median frequency formula above to find the predicted median frequency for each isoform. We define the probability of picking an isoform with rank k for a gene with M detected isoforms as:

$$p(\text{isoform}_k) = \frac{mf(k, M)}{\sum_{m=1}^{M} mf(m, M)}$$

With $M = 4$, the probabilities become [0.55, 0.28, 0.12, 0.05].

The inferred probability model In this model, we attempt to infer the probability of an isoform being chosen from its probability of being detected. The formula below relates the probability of choosing an isoform, $P(\text{Choice})$, to its probability of being detected, $P(\text{Detected})$:

$$P(\text{Detection}) = P(\text{Choice})P(\text{Detection}|\text{Choice}) + P(\neg\text{Choice})P(\text{Detection}|\neg\text{Choice}) \quad (1)$$

where $P(\neg\text{Choice})$ is the probability of not choosing an isoform. In practice:

$$P(\text{Detection}) = P(\text{Choice})P(\neg\text{Dropout})(1 - pFN) + P(\neg\text{Choice})pFP \quad (2)$$

where $P(\neg\text{Dropout})$ is the probability that there is not a dropout, pFN is the probability that there is a false negative event due to a quantification error and pFP is the probability that there is a false positive event due to a quantification error. This rearranges to:

$$P(\text{Choice}) = \frac{|P(\text{Detection}) - pFP|}{|P(\neg\text{Dropout})(1 - pFN) - pFP|}$$

In practice, we sometimes find $P(\text{Choice})$ is greater than 1, probably because our estimation of $P(\neg\text{Dropout})$,

pFN and/or *pFP* is inaccurate for that isoform. When this occurs, we set *P*(Choice) equal to one. We take absolute values of the numerator and denominator to avoid negative or complex numbers, which probably also occur due to inaccurate estimation of *P*(¬Dropout), *pFN* and/or *pFP*.

In our simulations, we calculate *P*(Choice) for each isoform from a given gene. The probability of picking a particular isoform to be expressed in our simulation is that isoform's *P*(Choice) divided by the sum of *P*(Choice)s for that gene's isoforms.

The cell variability model The cell variability model is identical to the inferred probabilities model except that the probability of picking a given isoform i is allowed to vary between cells. This is achieved by sampling the probability of picking isoform i in a given cell c, p_ic, from a beta distribution, taking a similar approach to that described in [4] :

$$p_ic \sim beta(\alpha, \beta)$$

where

$$\alpha = (\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}) \times \mu^2$$

$$\beta = \alpha \times (\frac{1}{\mu} - 1),$$

where μ is the mean probability of choosing i across all cells, i.e. $\mu = P$(Choice). Based on attempts to characterise the mean-variance relationship for the probability of choosing a particular gene by Velten et al. [4], we estimate that the sample standard deviation, σ, is approximately 0.002. We find p_ic for each isoform for a given gene. In our simulations, the probability of picking isoform i in cell c is that isoform's p_ic divided by the sum of p_ics for that gene's isoforms.

The random model For this model, each isoform is associated a weight randomly sampled between zero and one. The probability of picking a particular isoform to be expressed in our simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms.

The normal model The weights for each isoform were sampled from a truncated normal distribution with a mean of 0.25 and a standard deviation of 0.06. This sampling was performed for each isoform in each cell. Within each cell, the probability of picking a particular isoform to be expressed in our simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms.

The Bernoulli model The weights for each isoform were sampled from a Bernoulli distribution with a mean of 0.25. This sampling was performed for each isoform in each cell. Within each cell, the probability of picking a particular isoform to be expressed in our simulation is that isoform's weight divided by the sum of all the weights for that gene's isoforms. If all four isoforms for a given gene had a zero weight, we set the probability of picking each isoform to 0.25.

The p = 0.25 model The probability of choosing each isoform was always 0.25.

Step 3: Introduce dropouts based on Andrews et al.'s Michaelis-Menten model

We calculate the probability of dropouts for each isoform using a Michaelis-Menten model proposed by Andrews and Hemberg [9]. We calculate the probability of dropouts for each isoform as:

$$P(\text{Dropout}) = 1 - \frac{S}{K_M + S}$$

where S is the mean expression of that isoform across cells and K_M is the Michaelis-Menten constant. To find S and K_M, we normalise the isoform expression values by converting counts to counts per million (CPM), as suggested in the M3Drop vignette [9]. We estimate the value of K_M for each dataset by applying maximum-likelihood estimation using the equation above and the rate of dropouts and the mean expression of isoforms across the entire transcriptome.

Step 4: Introduce quantification errors

Based on our previous benchmarking study [8], we estimate that the probability of a false positive given an isoform has no reads mapping to it, *pFP*, is about 0.01 and the probability of a false negative given an isoform has reads mapping to it, *pFN*, is about 0.04 for Kallisto when run on full-length coverage scRNA-seq data. Unless otherwise stated in the text, these were the error rates applied in our simulations.

Step 5: Find mean number of isoforms per gene per cell

After iterating over every cell in our simulation, we sum the number of isoforms detected in each cell and divide by the number of cells to find the mean number of detected isoforms per gene per cell.

Step 6: Plot distributions of mean number of isoforms per gene per cell

Step 5 is carried out in each simulation, for every gene in which four isoforms were detected in the real scRNA-seq data. Consequently, a large list of mean number of detected isoforms per gene per cell is generated which we plot as distributions (e.g. see Fig. 1).

Mixture modelling

In our mixture modelling experiments, we begin by fitting log normal distributions to each of our simulation distributions and to the distribution of mean isoforms detected for genes with four detected isoforms in the real data. We then use expectation maximisation to estimate the mixing fraction of each of the simulated distributions in the real distribution. In our expectation step, we calculate the probability that each data point belongs to a given distribution, which we refer to as the responsibility. The responsibility for the ith mean number of isoforms per gene per cell and the cth simulation distribution is:

$$r_i c = \frac{k_c \times LN(x_i|\mu_c, \sigma_c)}{\sum_{j=1}^{j=4} k_j \times LN(x_i|\mu_j, \sigma_j)}$$

where k is the mixing fraction, x_i is the ith mean number of isoforms per gene per cell and $LN(x_i|\mu_c, \sigma_c)$ is the probability density function for the log normal with mean μ_c and variance σ_c^2. The maximisation function for the mixing fraction is:

$$k_c = \frac{\sum_i r_i c}{n}$$

where n is the number of datapoints in $r_i c$. Note that we only perform expectation maximisation for the mixing fractions of the distributions and not for the means or standard deviations.

Overlap fraction

The overlap fraction is the proportion of isoforms detected in our simulations that were expressed in the ground truth. The formula for the overlap fraction is:

$$OverlapFraction = \frac{|GroundTruth \cap Detected|}{|GroundTruth|}$$

where *GroundTruth* is the set of isoforms that are expressed in the ground truth and *Detected* is the set of isoforms that are detected in our simulations. The overlap fractions reported in all figures and Supplementary Figures are the mean overlap fractions for each gene, averaged across all of the simulated cells in that simulation round.

Downsampling

Random downsampling of reads was performed using seqtk [36].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-01981-w.

Additional file 1: Supplementary figures and tables. This PDF file contains all of the supplementary figures and tables for this paper.

Additional file 2: Review history.

Acknowledgments

We would like to thank Stephanie Telerman for helpful discussions.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The peer review history is available as Additional file 2.

Authors' contributions

JW, AFS and MH conceived the study and designed the experiments. JW and PA carried out the experiments. JW wrote the manuscript. JW, AFS and MH supervised the experiments. The authors read and approved the final manuscript.

Funding

JW was supported by a BBSRC DTP studentship BB/M011194/1. MH was funded by a core grant from the Wellcome Trust. AFS was funded by the Medical Research Council, MR/R009791/1.

Availability of data and materials

The Kolodziejczyk et al. mESC data was accessed from the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) using the accession number E-MTAB-2600, as described in the Kolodziejczyk et al. paper [23, 24]. The hESC datasets were accessed under GEO accession number GSE85917 [19, 20]. Our quantification pipelines, which download scRNA-seq data, perform transcript-level quantification and generate an isoform-cell matrix, and were used to carry out the downsampling study, can be found at https://github.com/jenni-westoby/Isoform_Cell_Matrix_Generation and at Zenodo [37–39]. Our quantification pipelines are licensed by the GNU General Public License v3.0. Our simulation pipeline is license by the MIT license and can be found at https://github.com/jenni-westoby/Obstacles and at Zenodo [40].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 10 October 2019 Accepted: 3 March 2020
Published online: 23 March 2020

References

1. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics. 2015;14(2):130–42. https://doi.org/10.1093/bfgp/elu035. Accessed 24 Nov 2017.
2. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics. 2017;18(1):583. https://doi.org/10.1186/s12864-017-4002-1. Accessed 21 Aug 2017.
3. Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, Woehler A, Akalin A, Ulitsky I, Chekulaeva M. Alternative 3′ UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Res. 2019;47(5):2560–73. https://doi.org/10.1093/nar/gky1270. Accessed 9 May 2019.
4. Velten L, Anders S, Pekowska A, Järvelin AI, Huber W, Pelechano V, Steinmetz LM. Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. Mol Syst Biol. 2015;11(6):812. https://doi.org/10.15252/msb.20156198. Accessed 28 Apr 2019.
5. Chen J, McSwiggen D, Ünal E. Single molecule fluorescence in situ hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis. J Visualized Exp. 2018;135:. https://doi.org/10.3791/57774. Accessed 15 Aug 2019.

6. Waks Z, Klein AM, Silver PA. Cell-to-cell variability of alternative RNA splicing. Mol Syst Biol. 2011;7:506. https://doi.org/10.1038/msb.2011.32. Accessed 8 May 2019.
7. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. Proc Natl Acad Sci U S A. 2016;113(39):11046–51. https://doi.org/10.1073/pnas.1612826113. Accessed 29 Apr 2019.
8. Westoby J, Herrera MS, Ferguson-Smith AC, Hemberg M. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. Genome Biol. 2018;19(1):191. https://doi.org/10.1186/s13059-018-1571-5. Accessed 28 Apr 2019.
9. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq,. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty1044. Accessed 24 June 2019.
10. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. https://doi.org/10.1038/nmeth.2967. Accessed 28 Apr 2019.
11. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11(2):163–6. https://doi.org/10.1038/nmeth.2772. Accessed 28 Apr 2019.
12. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6. https://doi.org/10.1038/nature07509. Accessed 6 Aug 2019.
13. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14(7):70. https://doi.org/10.1186/gb-2013-14-7-r70. Accessed 6 Aug 2019.
14. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498(7453):236–40. https://doi.org/10.1038/nature12172. Accessed 28 Apr 2019.
15. Zhao Z, Tu J, Lu Z, Liu S. Dominant isoform in alternative splicing in HeLa s3 cell line revealed by single-cell RNA-seq. In: Proceedings of the 7th International Conference on Computational Systems-Biology and Bioinformatics - CSBio '16. New York: ACM Press; 2016. p. 1–7. https://doi.org/10.1145/3029375.3029376. http://dl.acm.org/citation.cfm?doid=3029375.3029376. Accessed 21 Aug 2017.
16. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24(3):496–510. https://doi.org/10.1101/gr.161034.113. Accessed 28 Apr 2019.
17. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, Yeo GW. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol Cell. 2017;67(1):148–1615. https://doi.org/10.1016/j.molcel.2017.06.003. Accessed 5 Dec 2017.
18. Hu J, Boritz E, Wylie W, Douek DC. Stochastic principles governing alternative splicing of RNA,. PLoS Comput Biol. 2017;13(9):1005761. https://doi.org/10.1371/journal.pcbi.1005761. Accessed 23 Nov 2018.
19. Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods. 2017;14(6):584–6. https://doi.org/10.1038/nmeth.4263. Accessed 17 Apr 2017.
20. Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. SCnorm: robust normalization of single-cell RNA-seq data. RNA-seq and scRNA-seq datasets. Gene Expr Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85917.
21. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–7. https://doi.org/10.1038/nmeth.4220. Accessed 6 Mar 2017.
22. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–6434. https://doi.org/10.1016/j.molcel.2017.01.023. Accessed 28 Apr 2019.
23. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015;17(4):471–85. https://doi.org/10.1016/j.stem.2015.09.011. Accessed 28 Apr 2019.
24. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. RNA-seq and scRNA-seq datasets. Array Express. 2020. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2600/. Accessed 3 Jan 2020.
25. Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng. 2003;96(4):317–23. https://doi.org/10.1016/S1389-1723(03)90130-7. Accessed 2 Oct 2019.
26. Urban EA, Johnston RJ. Buffering and amplifying transcriptional noise during cell fate specification. Front Genet. 2018;9:591. https://doi.org/10.3389/fgene.2018.00591. Accessed 6 Dec 2019.
27. Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Research. 2018;7:1740. https://doi.org/10.12688/f1000research.16613.2. Accessed 28 Apr 2019.
28. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol. 2019;20(1):26. https://doi.org/10.1186/s13059-018-1605-z. Accessed 5 Feb 2019.
29. Arzalluz-Luque Á, Conesa A. Single-cell RNAseq for the study of isoforms-how is that possible? Genome Biol. 2018;19(1):110. https://doi.org/10.1186/s13059-018-1496-z. Accessed 28 Apr 2019.
30. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protocol. 2014;9(1):171–81. https://doi.org/10.1038/nprot.2014.006. Accessed 28 Apr 2019.
31. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777–82. https://doi.org/10.1038/nbt.2282. Accessed 28 Apr 2019.
32. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramskold D, Hendriks G-J, Larsson AJM, Faridani OR, Sandberg R. Single-cell RNA counting at allele- and isoform-resolution using Smart-seq3. BioRxiv. 2019. https://doi.org/10.1101/817924. Accessed 30 Oct 2019.
33. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7(12):1009–15. https://doi.org/10.1038/nmeth.1528. Accessed 21 Aug 2017.
34. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519. Accessed 4 Apr 2016.
35. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):766–73. https://doi.org/10.1093/nar/gky955. Accessed 8 Aug 2019.
36. lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk. Accessed 10 Dec 2019.
37. Westoby J. jenni-westoby/Isoform_Cell_Matrix_Generation: downsampling. Zenodo. 2020. https://doi.org/10.5281/zenodo.3659545. https://doi.org/10.5281/zenodo.3659545.
38. Westoby J. jenni-westoby/Isoform_Cell_Matrix_Generation: scnorm. Zenodo. 2020. https://doi.org/10.5281/zenodo.3659546. https://doi.org/10.5281/zenodo.3659546.
39. Westoby J. jenni-westoby/Isoform_Cell_Matrix_Generation: E-MTAB-2600. Zenodo. 2020. https://doi.org/10.5281/zenodo.3659542. https://doi.org/10.5281/zenodo.3659542.
40. Westoby J. jenni-westoby/Obstacles: v1.0.0. Zenodo. 2020. https://doi.org/10.5281/zenodo.3659553. https://doi.org/10.5281/zenodo.3659553.

Publisher's Note