Modeling degradation bias for accurate transcript quantification from long-read RNA-seq

by

Joseph Lee Jing Xian

Honours Project in Computational Biology

Faculty of Science

National University of Singapore

Supervisor:

Dr. Jonathan Göke

Genome Institute of Singapore

Co-supervisor:

Prof. Greg Tucker-Kellogg

National University of Singapore

2021/2022

Abstract

Long-read RNA-seq technologies have enabled the profiling of full-length reads while mitigating biases in previous generations of RNA-seq technologies, improving the accuracy of isoform abundance estimates. However, biases present in long-read RNA-seq data and their effects on transcript quantification have not yet been extensively explored in the literature.

In this thesis, we examine *degradation bias* present in long-read direct RNA-seq, where reads are truncated and map to multiple isoforms, leading to ambiguity in read-to-isoform assignment and erroneous isoform abundance estimates. We characterise degradation in real datasets, develop a bias-aware model for transcript quantification and derive statistical methods for inferring isoform abundance estimates. By accounting for degradation bias, we demonstrate improvements in transcript quantification on simulated datasets with known degradation rates and real datasets with sequencing spike-ins.

Acknowledgements

I thank Jonathan Göke for conceptualising this work and granting me the opportunity to work on it. I also thank him for his continual guidance, not only for the duration of this thesis, but also from the time I joined his lab in the summer of 2020. His support and patience have been invaluable.

My gratitude extends also to Chen Ying and Andre Sim for their contributions to this work: the weekly discussions on this subject, the resources they have provided, and the suggestions they have given me have all shaped this work crucially. Their own work is brilliant and has given me much insight and piqued my interest in all things long-read.

I also thank Prof. Greg Tucker-Kellogg for his co-supervision and for conducting great courses in bioinformatics at NUS that have laid foundations instrumental to this work.

Last but not least, I thank Prof. Choi Hyungwon for playing a large part in my statistical education which has come to fruition with this work.

Contents

List of Figures

List of Tables

Glossary

cDNA	Complementary DNA.
DNA	Deoxyribonucleic acid.
dRNA	Direct RNA.
EM	Expectation maximization.
mRNA	Messenger RNA.
ONT	Oxford Nanopore Technologies.
PacBio	Pacific Biosciences.
PCR	Polymerase Chain Reaction.
RNA	Ribonucleic acid.
RNA-seq	Ribonucleic acid sequencing.
SG-NEx	The Singapore Nanopore Expression Project.

Chapter 1

Introduction

Third generation sequencing technologies have enabled the production of long reads ranging from tens to hundreds of kilobases in length [1], and have shown promise in resolving many challenges in genomics and transcriptomics [2–7]. In particular, long-read technologies enable greater insight into the transcriptome and its complexity, which is crucial in understanding the functioning of cells and their biological processes. These technologies allow accurate detection of a larger proportion of full-length transcripts and novel splice junctions while mitigating biases associated with short-read technologies, enabling more accurate abundance estimates of reference and novel isoforms. Nevertheless, biases are still present in long-read technologies, albeit to a lesser extent.

This thesis focuses on a particular bias present in long-read direct RNA-seq referred to as *degradation bias*. This bias arises due to the fact that from the time a transcript is generated to when it is sequenced, it is subject to multiple factors that results in its degradation. Consequently, the reads obtained from sequencing are often truncated, resulting in ambiguity in read assignment to transcript isoforms. This, in turn, leads to erroneous isoform abundance estimates.

In this thesis, we attempt to characterise degradation bias and its effects on quantification from long-read direct RNA-seq, and develop a framework to model and correct such bias for accurate transcript quantification and isoform abundance estimation.

1.1 Review

Here, we review various concepts and existing literature relevant to our aim of modeling bias in long-read RNA-seq. We first examine (i) biases in short-read RNA-seq technologies and how they are accounted for by existing methods, which provide useful ideas on how to handle biases in long-read RNA-seq. Next, we review (ii) long-read RNA-seq technologies and how they mitigate biases in (i), and (iii) biases that long-read technologies themselves possess.

1.1.1 Bias modeling in short-read RNA-seq

Short-read RNA-seq technologies enable deep sequencing of highly accurate short reads, and has been the dominant technology for profiling the transcriptome since its popularisation in the early 2010s [8]. Typically, library preparation protocol for an RNA-seq experiment following RNA extraction involves RNA fragmentation, followed by the use of random hexamer primers for priming of the fragments to synthesize one strand of cDNA. After second strand synthesis, the resulting double stranded cDNA are size-selected and amplified via PCR amplification to generate enough cDNA for sequencing [9].

Biases in short-read RNA-seq have been extensively studied [10–17], and can often be traced back to specific steps in the protocol described above. For instance, [10] showed that priming with random hexamer primers induces bias in the nucleotide composition of the reads and results in non-uniform representation of reads across the length of the transcript. In addition, size selection of fragments results in an over-representation of fragments of a certain length, while RNA degradation and mRNA selection can lead to over-representation of fragments that are located towards either the beginning or end of the transcript [14, 16, 17]. PCR amplification before sequencing also introduces bias by preferential amplification of fragments with certain GC content [15, 17]. The combination of these biases affect quantification estimates, and if not corrected for, leads to erroneous estimates [14, 17].

Existing methods to address these biases can be categorised into two broad classes. The first class of methods involve innovations in library preparation methods to reduce bias from their source of origin [18]. For instance, an amplification-free RNA-seq protocol was proposed in [19] to reduce amplification bias, while thermostable polymerases that exhibit relatively lower GC bias were identified in [20]. We focus on the second class of methods that involve modeling and correcting for the bias *in silico*. In particular, we focus on the approaches in the literature that are most relevant to our aim.

In modeling bias, one needs to estimate parameters describing the bias from the data. The first approach adopted by many short-read RNA-seq methods for modeling bias is to use single isoform genes for estimating bias parameters. For instance, Cufflinks corrects for fragment length bias by estimating fragment length distributions based on a set of single isoform gene [14], while alpine estimates bias offsets and coefficients from single isoform genes for fragment length, read start sequence preference and GC content [17]. This approach ensures that the bias terms are computed from a subset of the data with the least ambiguity in assignment to the correct isoform. Furthermore, due to the complexity of real data, non-parametric density estimation is often used to model certain features. Again, both Cufflinks and alpine use kernel density estimates to fit an empirical fragment length distribution to correct for positional bias, i.e., the enrichment of fragments at the start or end of the transcripts. RSEM uses an empirical distribution to model the position of reads within fragments [12]. Other complex parametric approaches have been used, such as a mixture of Gaussian in [21], but appear less popular possibly due to parametric constraints. Last and perhaps most relevant is the use of curves to characterise and model for bias. In [22], coverage over single isoform genes is used to compute global and local bias curves which are then used to describe the non-uniformity of read distributions across all genes and for each gene respectively. In particular, the coverage over a gene is binned and transformed into a step function over exons of a gene [22]. We consider all these approaches when characterising bias and modeling for it in the following chapters.

1.1.2 Long-read technologies

As the name suggests, long-read sequencing technologies allow the sequencing of long reads ranging from one to hundreds of kilobases, improving over short-read technologies that yield fragments ranging from 50 to 600 bp. This has been made possible by the development of biophysical techniques to capture full-length DNA and RNA [23]. While the increased read length provided by these technologies has enabled leaps in genomics and transcriptomics research [2–7], these same technologies suffer from high sequencing error rates and low throughput [24–26].

Two of the most widely used platforms for long-read sequencing include Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). We omit discussion on PacBio data for the time being as we work only with ONT data in this thesis. ONT is a nanopore-based sequencing technology that currently provides three protocols for transcriptomic sequencing: direct RNA, direct cDNA and PCR-cDNA sequencing. In all three protocols, DNA or RNA is processed and ligated with sequencing adaptors comprising a motor protein that ensures unidirectional single-nucleotide translocation through the nanopore on the order of milliseconds [27]. Each nanopore is associated with a sensor that measures changes in ionic current caused by differences in the nucleotides occupying the pore [27], giving rise to raw electric current signal. In addition, the nanopores are embedded in a synthetic electrically-resistant membrane to ensure that all current only passes through the pore, resulting in a cleaner signal [28]. The raw electric current signal is then translated to nucleotide sequence via base-calling algorithms [29].

We briefly describe the similarities and differences between the three ONT protocols, with a focus on direct RNA-seq. In direct RNA-seq (Fig. 1.1), native poly(A) RNA molecules are enriched and directly sequenced from the 3' end. An optional reverse transcription step stabilizes the RNA strand by reducing intramoleulcar secondary structure of the RNA, resulting in better sequencing output [30]. In direct cDNA and PCR-cDNA protocols, poly(A) RNA is reverse transcribed followed by second strand synthesis to produce double stranded cDNA. Strand-switching is used to increase the proportion of full-length cDNAs. In PCR-cDNA, double stranded cDNAs are amplified prior to sequencing, giving higher throughput. Across all three protocols, GC bias was evaluated to be minimal in the data, including data from the PCR-cDNA protocol, provided that the number of PCR cycles was capped [31].

Compared to cDNA protocols, direct RNA-seq mitigates any form of bias or errors due to PCR or reverse transcription [30, 32] which is important for obtaining accurate and reliable isoform abundance estimates. Despite this, a recent evaluation of ONT direct RNA-seq showed that although direct RNA-seq yielded significantly longer reads compared to short-read RNA-seq, the proportion of reads corresponding to full-length transcripts remained low due to read truncation [33]. Assigning direct RNA-seq reads to their isoform of origins remains highly nontrivial, in large part due to the number of secondary alignments for each read [33]. In the following section, we review some possible explanations for the observed read truncation.

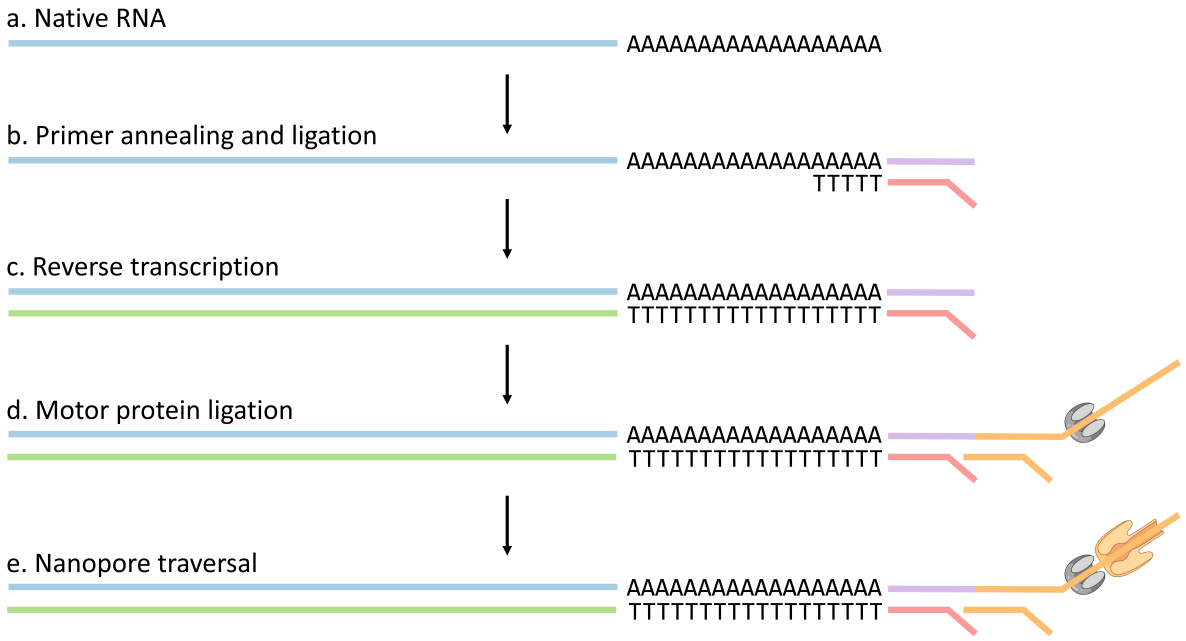

Figure 1.1: ONT direct RNA-sequencing protocol. **a.** Native RNA with 3' poly(A) tails are enriched for sequencing. poly(A) tails can be added to non-poly(A) RNA with a poly(A)-tailing kit. **b.** A poly(T) adaptor is annealed to native RNA to prime first-strand synthesis of cDNA. **c.** Reverse transcription of the RNA strand stabilises it for sequencing and improves sequencing output. **d.** A sequencing adaptor comprising a motor protein is ligated to the RNA strand and is necessary for guiding it through the nanopore. **e.** RNA traverses through the nanopore, producing raw signal output which is then basecalled.

1.1.3 Biases in long-read RNA-seq

We now discuss biases in long-read ONT data with a particular focus on degradation bias observed in the direct RNA protocol.

GC bias According to ONT, datasets sequenced with all three protocols (direct RNA, direct cDNA, PCR cDNA) exhibit virtually no GC bias [34], with low correlations between the GC content and read count for each gene. This was corroborated by studies finding that compared to other high throughput sequencing methods such as short-read RNA-seq or PacBio sequencing, there was no relative coverage biases across regions of different GC content [35, 36].

Degradation bias Here, we consider sources of degradation bias observed in long-read direct RNA-seq, where the reads are degraded or truncated [33, 37]. This results in ambiguity in read-to-isoform assignment [33] (Fig. 1.2).

Broadly, factors that influence transcript degradation can be broadly classified as being intra-cellular (*in vivo*) or extra-cellular (*in vitro*). We first examine the dominant source of degradation *in vivo*, which is RNA decay. RNA decay is a well studied phenomena, and can occur due to the action of both exo- and endoribonucleases [38–41]. In the first instance, RNA decay occurs due

to exoribonuclease cleavage from the 5' or 3' ends of the transcript. In mRNAs, this results in the removal of modifications such as 5'-capping or 3'-polyadenylation that help to stabilise the transcript for translation initiation, initiating RNA degradation. Such degradation can occur in both the 5'-3' or 3'-5' directions, with the former being the more dominant pathway [39].

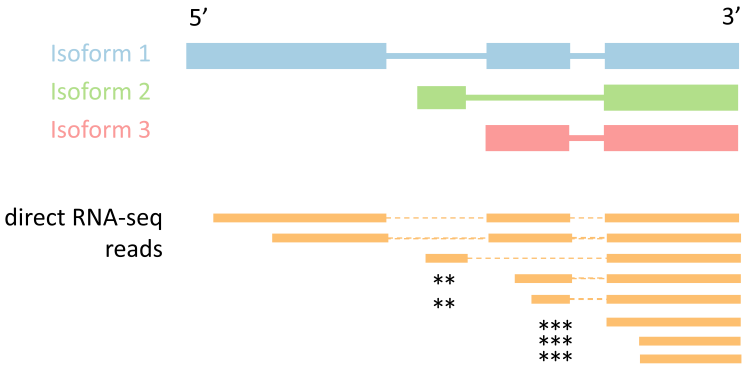

Figure 1.2: Degradation bias results in multi-mapping reads. In this hypothetical scenario, there are three isoforms transcribed (blue, green, red). Their transcripts are sequenced to produce reads (orange). Due to degradation, the observed reads are truncated, predominantly from the 5' end. This results in multi-mapping reads (starred). The number of possible isoforms each read is compatible with is indicated by the number of stars adjacent to the read.

A minor pathway for RNA decay involves endoribonuclease cleavage within the body of the transcript [39]. Endoribonuclease cleavage exposes unprotected 5' and 3' ends of the distal and proximal segments of the cleaved transcript, which then expose these segments to degradation by exoribonucleases [39, 42]. Differential rates of mRNA decay in different functional sets of genes have also been studied. For instance, transcription factor mRNAs tend to exhibit higher rates of mRNA decay compared to mRNA that code for proteins [43]. Other classes of RNA also exhibit variable decay patterns [41]; [44] described large variability in the decay rates of lncRNA compared to mRNAs, finding lncRNAs that were both extremely short-lived and long-lived.

Compared to *in vivo* effects, extra-cellular *in vitro* effects are less well studied, due to the relative novelty of nanopore sequencing. In [37], it was suggested that read truncation in nanopore sequencing could be due to electronic signal noise from current spikes of unknown origin. In addition, pore blocking, which results in the stoppage of RNA translocation through the nanopore, could also result in truncated reads [45]. Besides these considerations, generic factors that are not specific to nanopore sequencing, such as the fragmentation of transcripts during library preparation, can also result in the observed read truncation.

Existing methods for transcript quantification from long-read RNA-seq, such as FLAIR [46] and NanoCount [47], do not attempt to correct for degradation bias. To the best of our knowledge, the only method that currently models degradation bias is Bambu [48], a tool for transcript discovery and quantification. To assign reads to isoforms, Bambu uses an alignment compatibility weight between the read class (reads sharing the same splice junctions) and isoform. In multi-mapping

read classes, these weights are adjusted to account for constant degradation based on the effective length of the read class [48].

Keeping the ideas from this review in mind, we aim to characterise degradation in long-read direct RNA-seq and develop a model for correcting such bias.

1.2 Organisation

The chapters of this thesis are organised as follows:

- In Chapter 2, we formalise the notion of degradation, and characterise degradation in long-read direct RNA-seq data in reference isoforms and sequencing spike-ins across multiple cell lines and samples.
- In Chapter 3, we develop a generative model for degradation-aware transcript quantification and derive an expectation maximization algorithm for parameter inference.
- In Chapter 4, we evaluate our model and inference algorithms on simulated datasets with known degradation and real datasets with sequencing spike-ins. We benchmark our model against existing long-read transcript quantification methods in the literature.
- In Chapter 5, we summarise the ideas of this thesis and discuss potential directions of future work.

Chapter 2

Characterising degradation

In this chapter, we aim to characterise transcript degradation and read truncation from long-read RNA-seq data. Transcript degradation from the 5' end results in truncated reads and a decrease in coverage with increasing distance from the 3' end for a given isoform. Thus, even though the observed degradation occurs from the 5' end, it is helpful to characterise degradation as the resultant decrease in coverage from the 3' end. We formalize the notion of degradation by defining the *degradation rate*.

Definition 2.0.1 (Normalized coverage)**.** Let the maximum coverage over an isoform be $\mathrm{cov}_{\max}$ and the coverage at base b be cov_b. The normalized coverage of the isoform at base b is defined as

$$\mathrm{ncov}_b = \frac{\mathrm{cov}_b}{\mathrm{cov}_{\max}} \tag{2.1}$$

Definition 2.0.2 (Degradation rate)**.** Let ncov be the normalized coverage over an isoform, and x be the distance in kb from the 3' end. The degradation rate of an isoform is defined as the rate of change in normalized coverage with respect to distance in kb from the 3' end of the isoform:

$$d = \lim_{\Delta x \to 0} \frac{\Delta \mathrm{ncov}}{\Delta x} \tag{2.2}$$

The degradation rate of an isoform at a base b with distance x_b away from the 3' end is the value of this limit evaluated at $x = x_b$.

Intuitively, the degradation rate can be interpreted as the gradient of the plot of normalized coverage against distance from the 3' end. We will refer to this plot as the **degradation curve**. To illustrate this, we visualise degradation curves for a hypothetical isoform of length 2 kb with different degradation rates (Fig. 2.1).

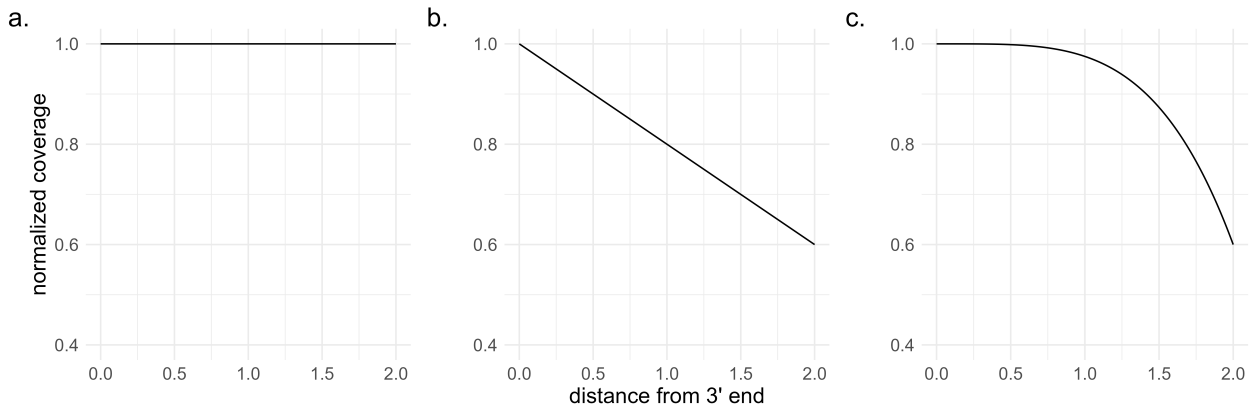

Figure 2.1: Normalized coverage plots for a hypothetical isoform of length 2 kb illustrating different degradation rates. **a.** Degradation rate is 0. **b.** Degradation rate is constant (0.2). **c.** Degradation rate is variable.

In Fig. 2.1a, the degradation rate or gradient is 0, implying that all reads from the isoform are full-length, with no drop in coverage over the isoform body. Conversely, in Fig. 2.1b, the gradient is a constant value of 0.2 over the isoform body, implying that for every 1 kb from the 3' end, normalized coverage drops by 0.2. The last plot in Fig. 2.1c shows variable gradient over the isoform body. In particular, the gradient is low in magnitude towards the 3' end and increases with distance from the 3' end.

In the following sections, we first describe a coverage-based approach for characterising degradation in long-read direct RNA-seq data, and validate this approach with simulated data where the degradation is known. We then examine degradation in real data, characterising degradation by isoform features and in sequencing spike-ins. Finally, we develop a method for efficient read length-based degradation estimation that can be used for obtaining degradation-aware isoform abundance estimates.

2.1 Coverage-based degradation estimation

We first sought to determine if patterns of degradation were consistent across transcript isoforms for a given direct RNA-seq dataset. To that end, we selected single-isoform, multi-exon genes from GRCh38 reference annotations, and further restricted the set of isoforms to those that do not intersect with any other annotated features in reference annotations. We refer to these as **lone isoforms**. Estimating bias from lone isoforms reduces ambiguity in the isoform of origin of the reads we use to estimate bias; such approaches were also adopted in [14] and [17] for bias estimation in short-read data. Filtering yielded approximately 5,000 lone isoforms. For each of these isoforms, we obtained coverage over the isoform body using the `genomecov` module from bedtools (v.2.27.1). Next, we further apply a median coverage filter, filtering out lone isoforms with low median coverage (min median coverage $= 10$). Finally, a degradation curve is fitted to the data with a smoothing spline.

To validate this approach, we simulated two datasets where the expected degradation rates for all isoforms is constant ($\mathbb{E}[d] \in \{0.2, 0.4\}$, see Appendix A for more details on degraded read simulation). Visualising the degradation curves for each isoform, we observed noise in the form of deviation from the expected degradation curve (grey lines, Fig. 2.2), likely due to sampling noise in the simulation data generation process. This noise can be mitigated by computing a global degradation curve across all isoforms (red line, Fig. 2.2). We find that the global degradation curves reflect the known degradation rates from the simulated data qualitatively (Fig. 2.2). In addition, we assessed the global degradation curves quantitatively by computing the average gradients of each curve, and found good agreement between the estimates and the expected degradation ($\mathbb{E}[d] = 0.2$, estimated $= 0.201$; $\mathbb{E}[d] = 0.4$, estimated $= 0.403$).

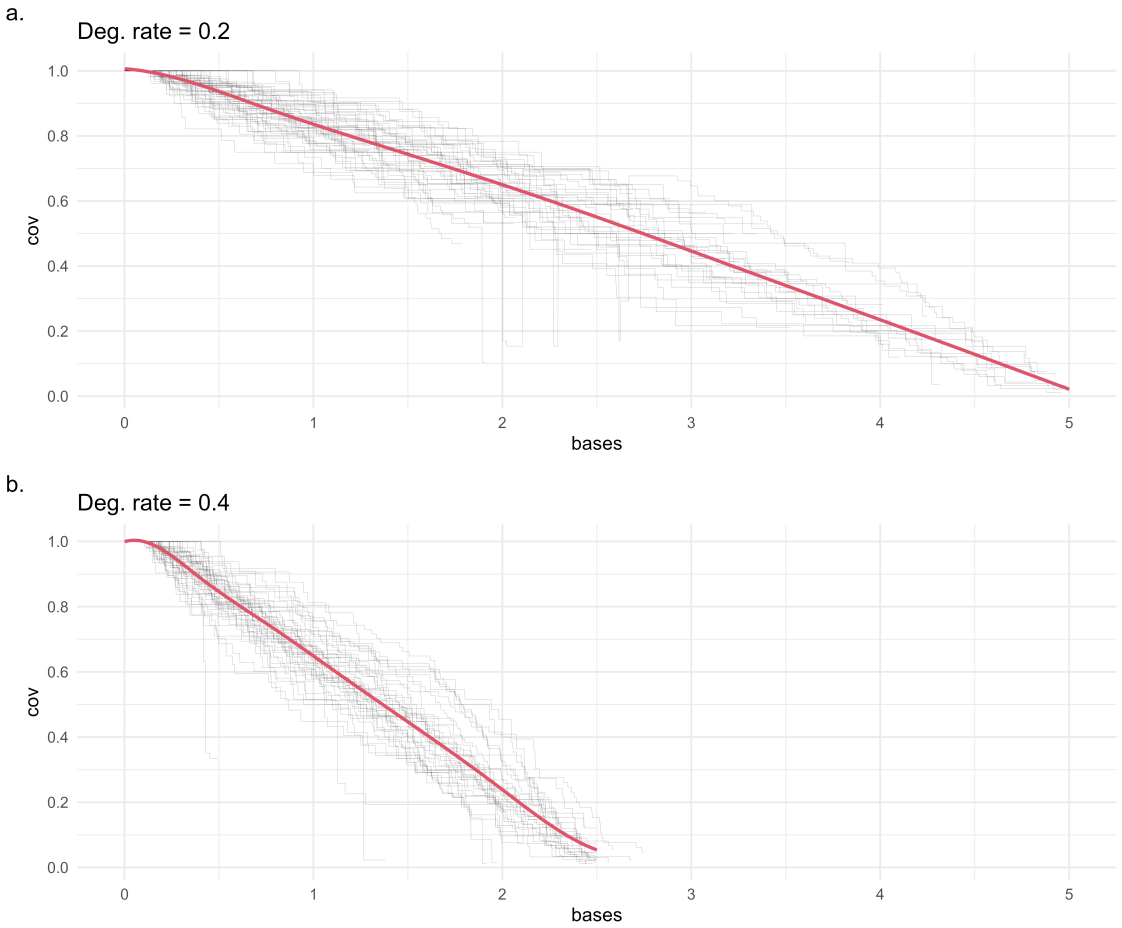

Figure 2.2: Degradation curves on simulation datasets based on coverage. Grey translucent lines represent the coverage curves for individual isoforms. The red line represents the spline fitted across all isoforms. **a.** Degradation curve for dataset with $\mathbb{E}[d] = 0.2$. Coverage drops close to 0 at 5kb with a degradation rate of 0.2. The estimated degradation rate is 0.201. **a.** Degradation curve for dataset with $\mathbb{E}[d] = 0.4$. Coverage drops close to 0 at 2.5 kb. The estimated degradation rate is 0.403.

We applied this approach in real direct RNA-seq data from the SG-NEx project [49]. Within each sample, we found consistent patterns of degradation across isoforms (Fig. 2.3). However, we note that for most real datasets, filtering for lone transcripts and by median coverage yields very few isoforms remaining for estimating degradation rates. For instance, in a HepG2 sample (Fig. 2.3a), we obtained only 57 isoforms, while in a MCF7 sample (Fig. 2.3b), we obtained only 52 isoforms. Across the samples, the median number of lone isoforms post-filtering was 21. While this approach separates signal and noise by considering only lone isoforms, it may be overly restrictive. We consider an improved approach for estimating the degradation rate in Section 2.2 based on observed read length distributions.

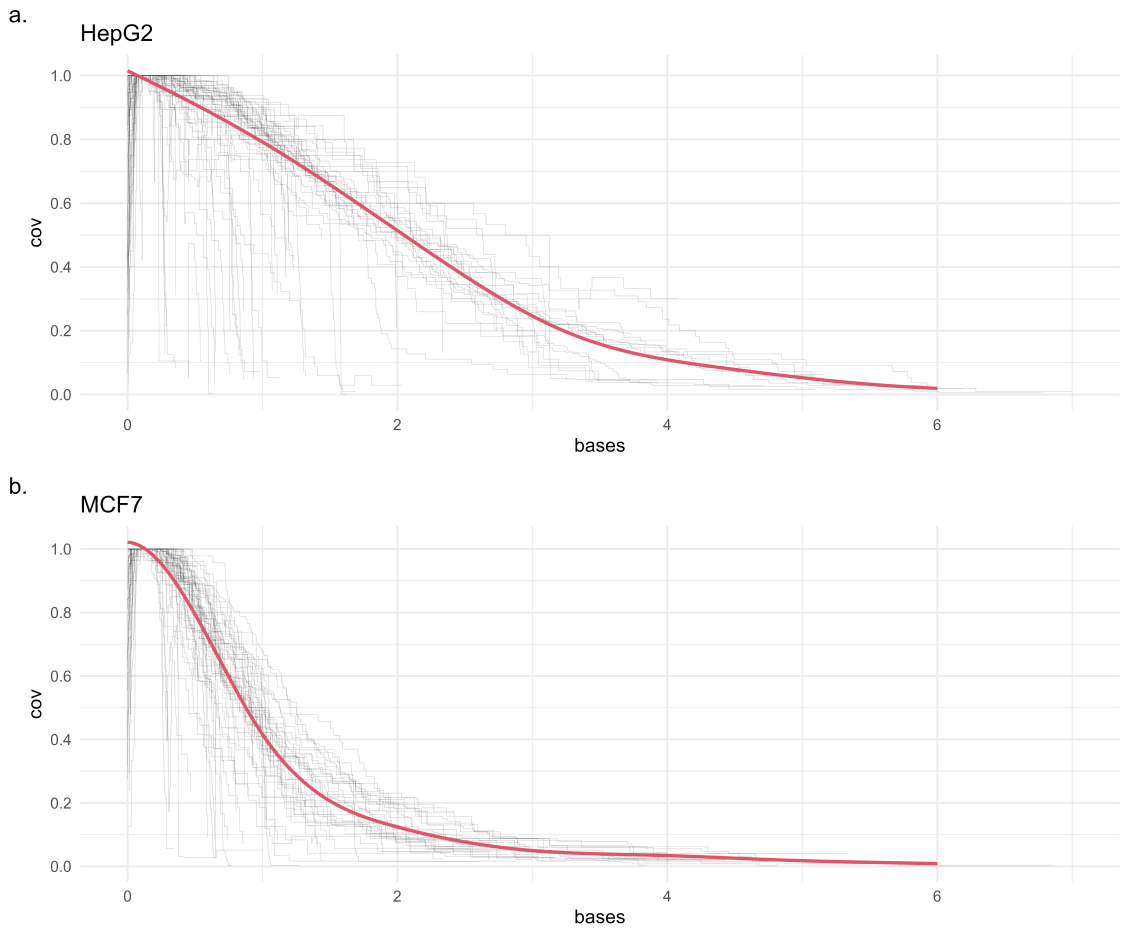

Figure 2.3: Degradation curves on real datasets based on coverage. Grey translucent lines represent the coverage curves for individual isoforms. The red line represents the spline fitted across all isoforms. These two datasets had the most number of isoforms post-filtering for degradation rate estimation (n = 57, 52). **a.** Degradation curve for a HepG2 cell line sample. **b.** Degradation curve for a MCF7 cell line sample.

2.1.1 Degradation by isoform features

Here, we explore whether degradation rates vary between isoforms of different features. In particular, we striate isoforms by their annotated length (Fig. 2.4a), observed median coverage (Fig. 2.4b) and biotype (Fig. 2.4c). On a representative sample from the MCF7 cell line, estimated degradation rates do not appear to vary significantly between isoforms of different annotated length or median coverage (Fig. 2.4a,b). However, the converse is true for different transcript biotypes. In particular, we observe higher degradation rates for processed pseudogenes (n = 7) and long non-coding RNAs (n = 3) as compared to protein coding genes (n = 42, Fig. 2.4c). While the sample size here is small, this result corroborates findings in [50] and [44], which identified larger variance in the stabilities of lncRNAs.

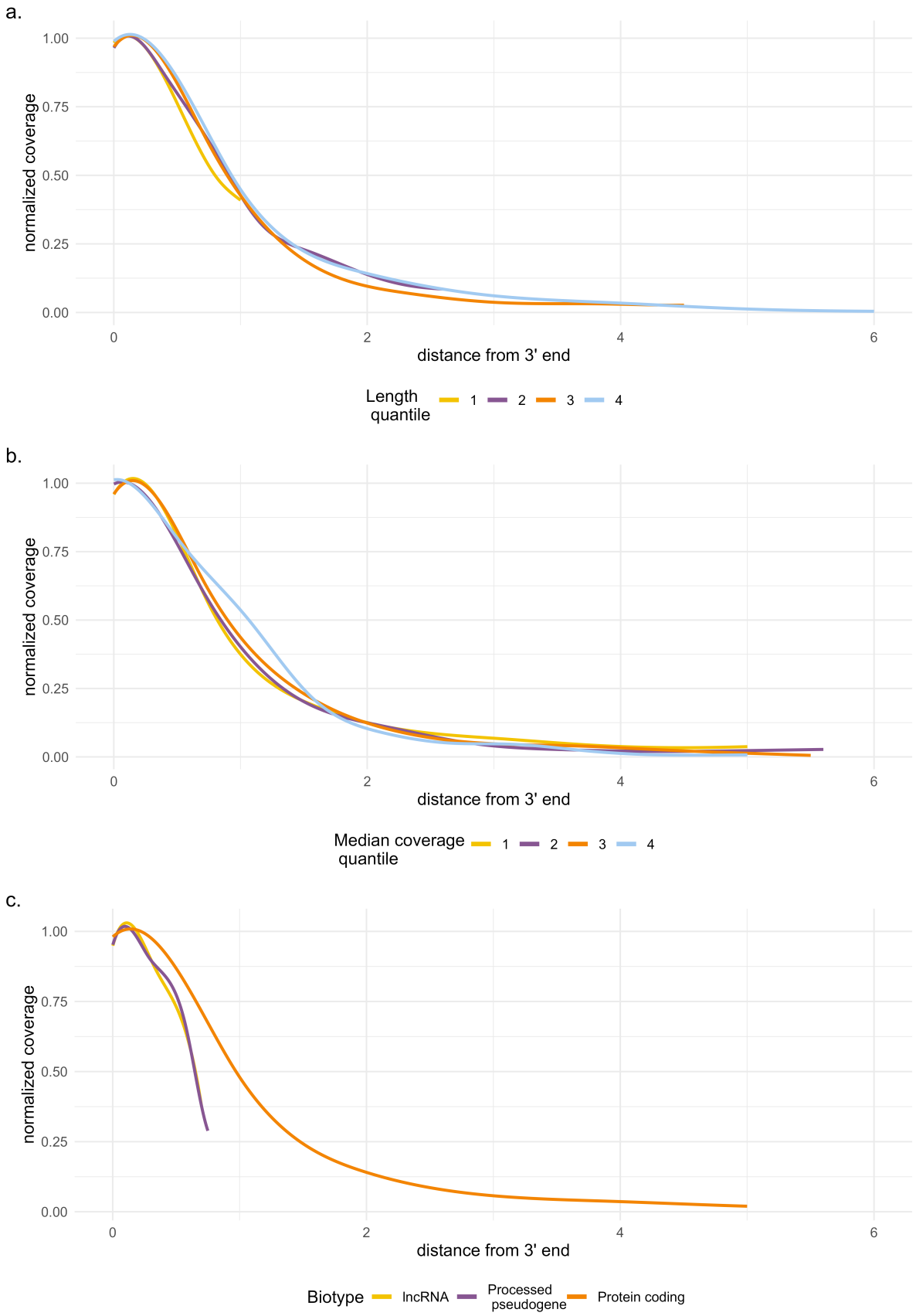

Figure 2.4: Degradation curves in MCF7 striated by features. **a.** Degradation curves fit for isoforms within each length quantile separately, with quantiles at 0%: 320, 25%: 1360, 50%: 2820, 75%: 5590, 100%: 9038. **b.** Degradation curves fit for isoforms within each median coverage quantile separately, with quantiles at 0%: 7, 25%: 12, 50%: 18, 75%: 31, 100%: 381. **c.** Degradation curves fit for isoforms of different biotypes.

2.1.2 Degradation in spike-ins

We now analyse possible degradation in sequencing spike-ins, which are synthetic RNA molecules added to endogenous RNA samples before library preparation [51]. By doing so, we attempt to elucidate the relative contributions of *in vivo* RNA decay and other extra-cellular factors unrelated to decay, such as library preparation or sequencing artifacts.

We examine SIRVs (Spike-in RNA Variants) [51] present in a subset of SG-NEx samples. Fortunately, a subset of SIRV isoforms are non-overlapping, allowing us to easily apply coverage-based approaches for estimating the degradation. The degradation curves estimated on the SIRVs for six H9 samples (2 runs with 3 replicates each) show consistent patterns amongst themselves (dotted lines, Fig. 2.5), suggesting a constant extraneous factor that acts on degrading SIRV reads. In fact, the degradation rate appears to be constant for each sample, and the average degradation rate is consistent across all samples (run 1: 0.121, 0.148, 0.128; run 2: 0.149, 0.124, 0.183).

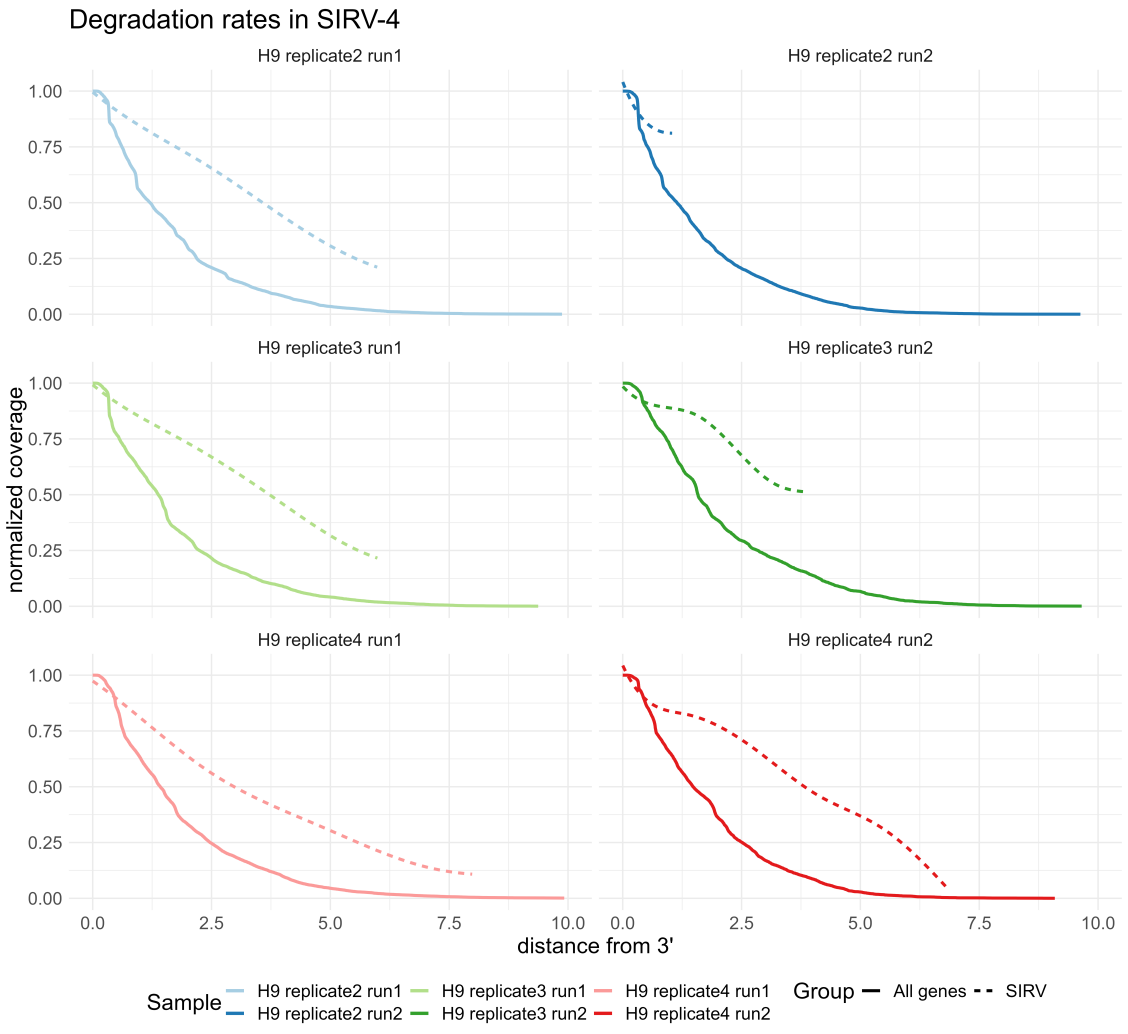

Figure 2.5: Degradation curves for SIRVs based on coverage for six H9 samples. The number of SIRV transcripts used to fit the degradation curve differs between samples, resulting in curves of different lengths.

When contrasted against degradation curves estimated on all isoforms from within the same sample (solid lines, Fig. 2.5), the SIRVs show lower rates of degradation. This might suggest that the degradation in endogenous RNAs is a combination of RNA decay *in vivo* and extraneous *in vitro* factors.

2.2 Read length-based degradation estimation

From the coverage-based estimation of degradation in the preceding sections, we gleaned that patterns of degradation tend to be consistent across transcript isoforms for a given sample across annotated length and median coverage. This holds at least for protein coding isoforms. Based on these observations, we develop a more efficient approach for estimating the degradation rate via the observed read length distributions.

Recall that our objective is to estimate the degradation curve, from which we can easily derive the degradation rate. The key observation for developing a read length-based degradation estimation approach is to note that the degradation curve is essentially the survival function on degraded read lengths. Let X be a random variable denoting the length of a read. Then, its survival function is given by

$$S(x) = P(X > x) = 1 - F(x) \tag{2.3}$$

where F is the cumulative distribution function of X. Intuitively, the survival function gives the proportion of reads that exceed (survive after) a given length x. We make this clear with an illustration on a hypothetical dataset with a degradation rate of 0.2 (Fig. 2.6). Here, the degradation curve provides an easy way to read off the proportion of degraded reads that are at least 1 kb in length, which is $S(1) = 0.8$.

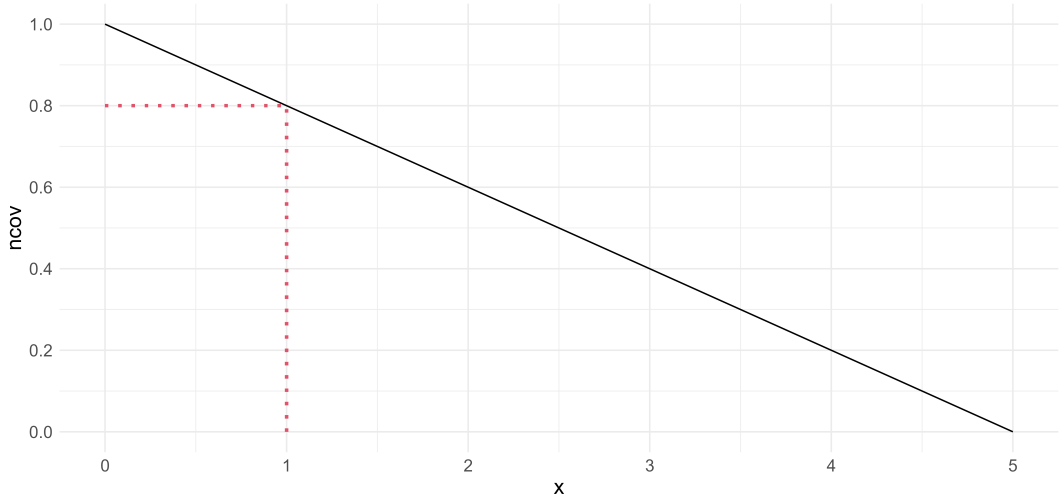

Figure 2.6: Equivalence between degradation curves and survival functions on degraded read lengths.

Based on this observation, we develop an approach for estimating degradation based on read

length distributions:

1. Bin lone isoforms by their length and sample a fixed number of isoforms from each bin. We do so to avoid capturing the isoform length distribution in the degraded read length distribution.

2. Extract degraded reads based on the isoforms sampled in step 1. We define degraded reads as those whose lengths are not within the annotated full length by some threshold (in practice, we use 50 bp).

3. Compute the empirical cumulative distribution function F on the read lengths obtained in step 2. Deriving the degradation curve $1 - F$ is trivial.

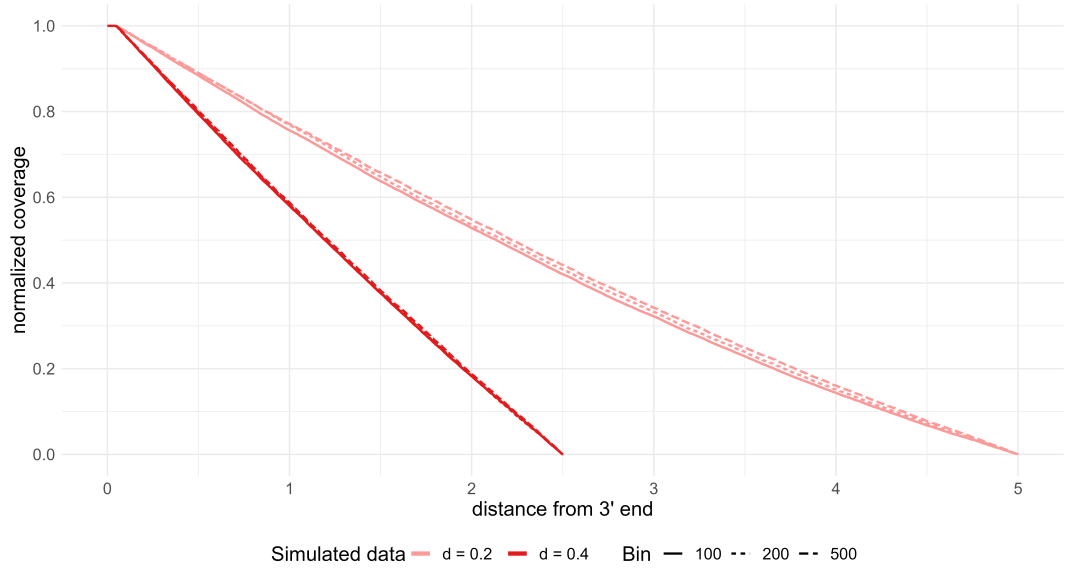

Figure 2.7: Degradation curves on simulated datasets based on read length distribution. For degradation rate of 0.2, coverage drops to 0 at 5 kb (light red), while for degradation rate of 0.4, coverage drops to 0 at 2.5 kb (dark red). The degradation rates are relatively robust to bin size.

	Bin size		
Simulated dataset	100	200	500
d = 0.1	0.118	0.113	**0.107**
d = 0.2	0.207	0.206	**0.204**
d = 0.4	**0.408**	0.409	**0.408**
d = 0.5	0.513	0.512	**0.511**

Table 2.1: Degradation estimates on simulated datasets based on read length distribution. The length bin size is varied for 100, 200 and 500 bp, and the average degradation rate is computed.

We validate this approach on the same simulated datasets with degradation rates of 0.2 and 0.4 and test the robustness of the length bin size used. The degradation curves qualitatively

reflect the expected degradation rate (Fig. 2.7) and appear robust to the bin size chosen (Table 2.1). We note that in the range of bin sizes we tested, a larger bin size yielded estimates closest to the expected degradation rates. Intuitively, a smaller bin size captures more variance in the degraded read length distribution, making the estimates noisier.

With this validation in hand, we estimated the degradation curves for 32 direct RNA-seq samples spanning six cell lines and six sequencing runs (Table 2.2) for multiple bin sizes. We observed similar trends in the degradation across all samples, with low rates of degradation toward the 3' end, followed by a relatively linear segment and tapering of towards the 5' end asymptotically. Within each cell line, the degradation curves were mostly consistent, with the exception of a few outlier samples for H9 and HepG2 (Fig. 2.8).

Next, we computed a regression line based on the degradation curve and calculated its gradient as a proxy for the *average degradation rate*. Here, the average degradation rate allows for relative comparisons between samples and acts as a summary statistic. In practice, we do not use these point estimates to correct for degradation bias; rather, the entire degradation curve is used to do so (Section 3.2.3).

We report the average degradation across the samples for multiple bin sizes, and group them by cell line (Fig. 2.9a) and sequencing run (Fig. 2.9b). We first observe that the estimated degradation rates are relatively robust to bin size. Next, for certain cell lines, there is high variance in estimates, presumably due to the samples being sequenced in multiple different runs (e.g. Hct116). These observations were paralleled on the level of sequencing runs. It is likely that the combination of sequencing runs and cell lines influence the observed degradation, although fitting a linear model for the average degradation rate did not produce significant results.

Lastly, we computed correlations between the estimated degradation rates with raw sequencing metrics (Fig. 2.10). We observed moderate negative correlations between degradation rate and raw sequenced lengths, and weak correlation with sequencing quality. Interestingly, the strongest correlation was achieved with read length standard deviaion (SCC = -0.90). We postulate that higher degradation rates result in a smaller possible range of read lenghts, resulting in lower variance in lengths.

	Run1	Run2	Run3	Run4	Run5	Run6
A549	4	0	0	0	0	0
H9	4	3	0	0	0	0
Hct116	3	1	2	1	1	1
HepG2	3	1	1	0	0	0
MCF7	2	1	0	0	0	0
K562	4	0	0	0	0	0

Table 2.2: Description of SG-NEx samples across cell lines and sequencing runs

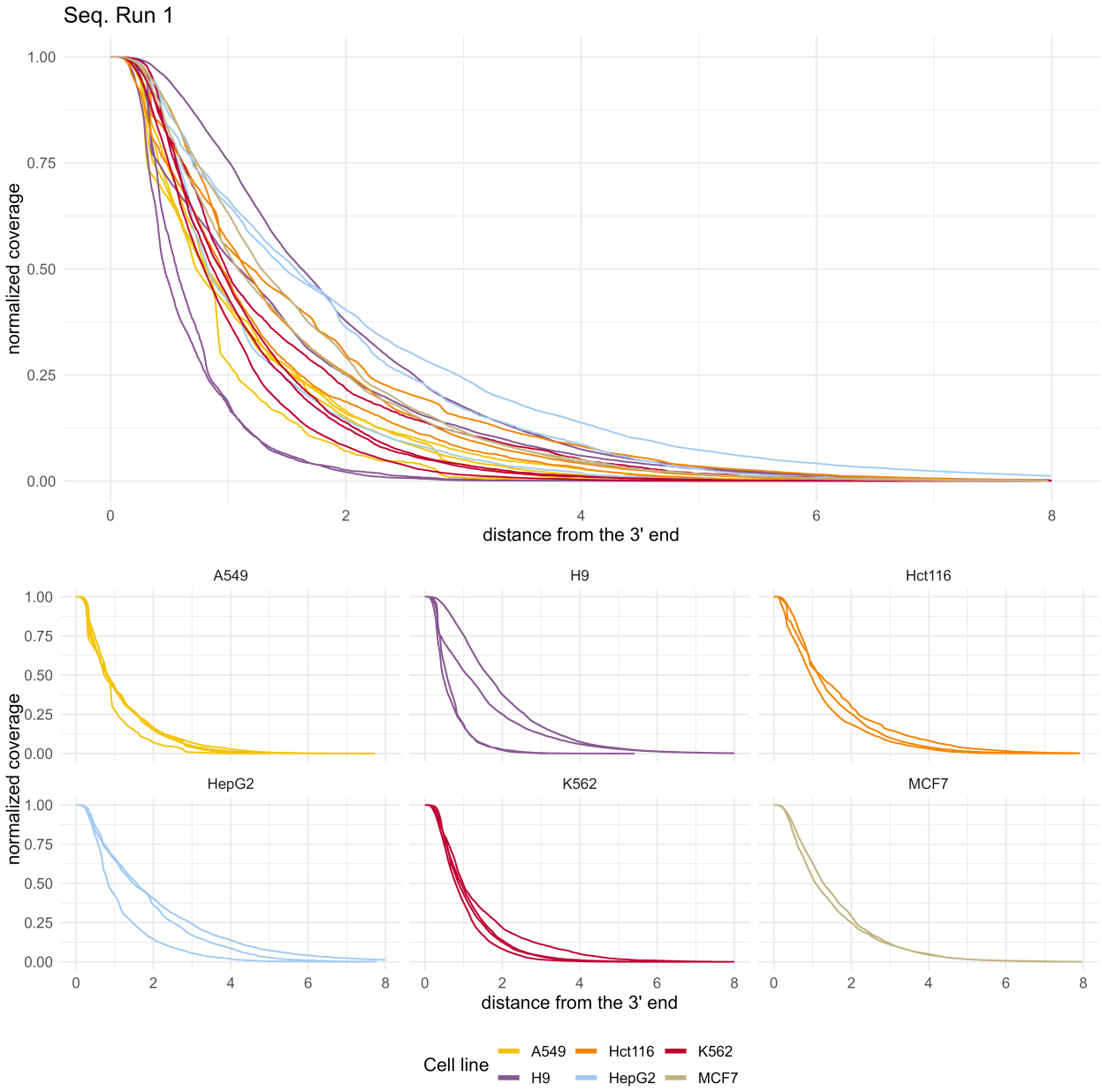

Figure 2.8: Degradation curves for real datasets based on read length distribution for sequencing run 1 with bin size 500. The top panel shows all samples combined. The bottom panel splits the samples by cell lines. In general, the degradation curves show consistent patterns, with low rates of degradation toward the 3' end, followed by a linear segment and a tapering off as distance from the 3' end increases.

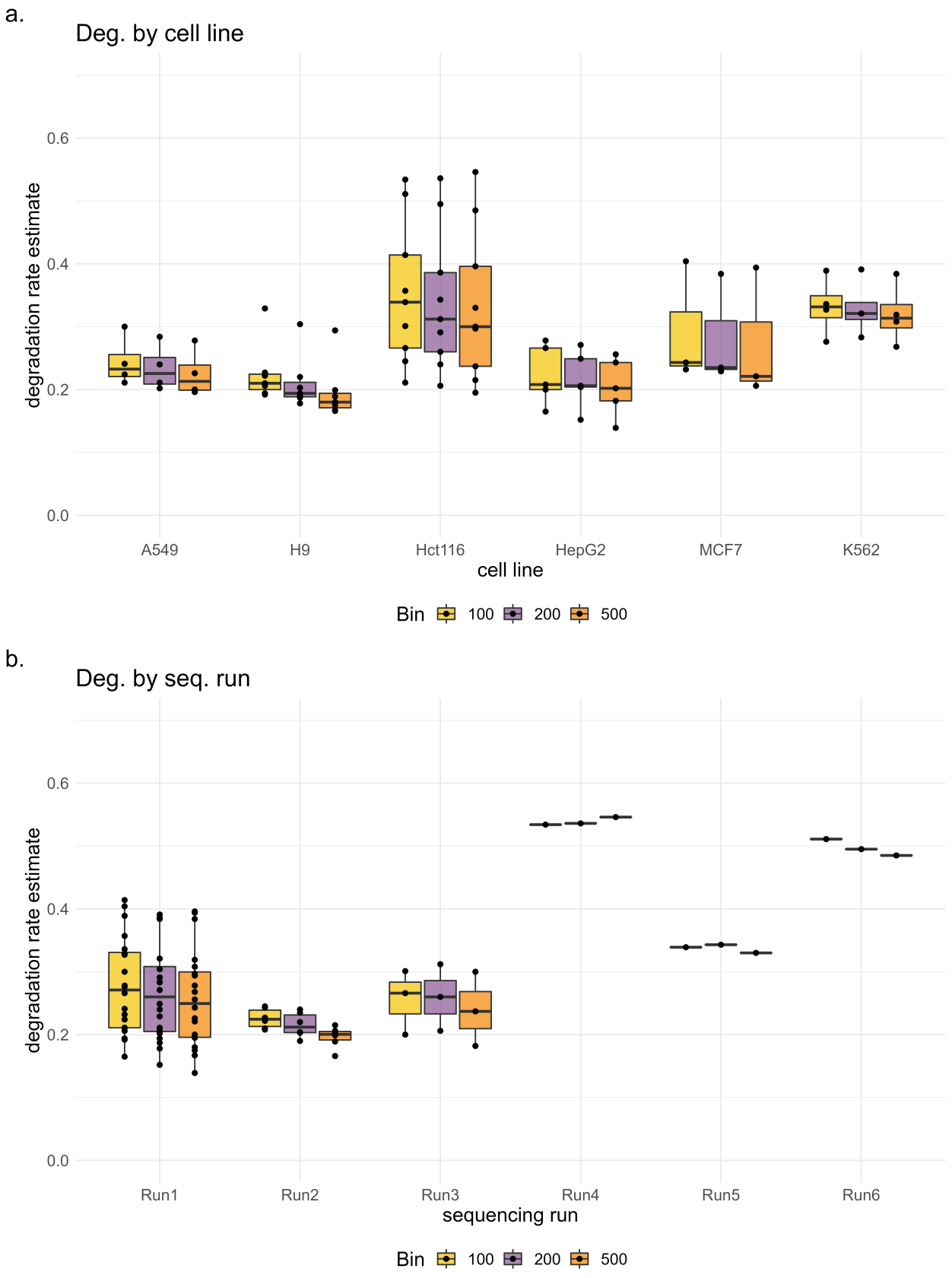

Figure 2.9: Degradation estimates for real datasets based on read length distribution. The average degradation rate is robust to eh bin size chosen. **a.** Samples grouped by cell line. **b.** Samples grouped by sequencing run.

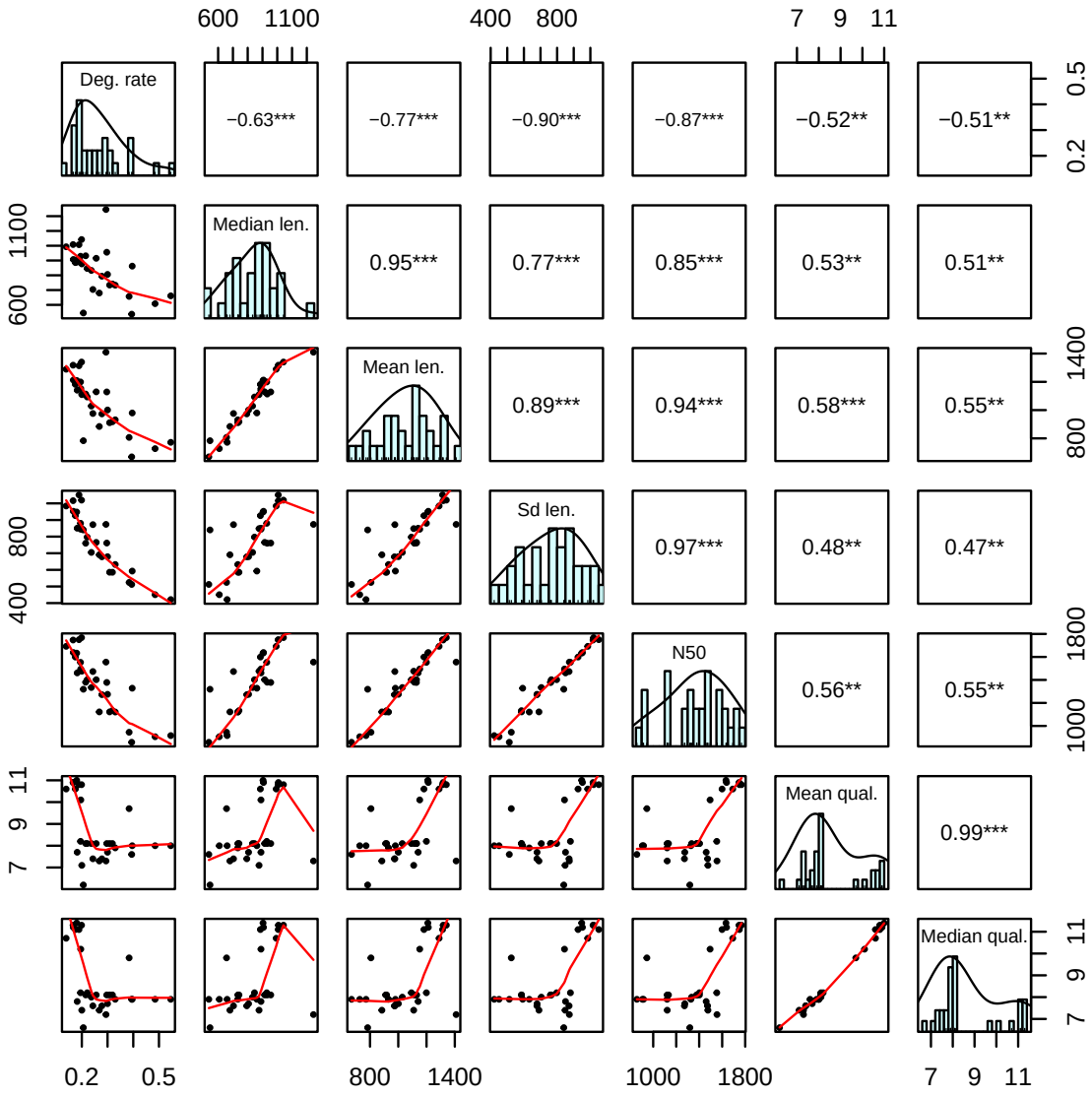

Figure 2.10: Correlation of degradation estimates with other sequencing metrics. From left to right: degradation rate, median length, mean length, read length standard deviation, N50, mean quality and median quality.

2.3 Discussion

In this chapter, we formalized the notion of the degradation rate, and described two approaches for estimating it from long-read direct RNA-seq data. The first approach involves computing the coverage over isoforms and fitting a curve to the coverage track. Based on this approach, we find that degradation curves appear to be mostly consistent across isoforms, and is not dependent on the length of the isoform or the median coverage over the isoform. However, degradation appears to differ by the biotype of the RNA, which corroborates established knowledge.

Next, we examined the degradation rates in SIRVs and endogenous RNAs from the same samples. SIRV reads exhibited mostly constant degradation rates while endogenous RNAs exhibited variable and steeper degradation rates. This suggests that the observed degradation in endogenous RNA reads is a combination of *in vivo* RNA decay, which contributes a variable component to the degradation, and extraneous *in vitro* factors such as artifacts due to library preparation or sequencing, which contribute a constant component to the degradation.

Finally, we developed an efficient way of estimating degradation curves based on the degraded read length distribution which does not rely on computing coverage or fitting curves, removing noise and arbitrary parameters in the estimation process. We apply this approach on 32 samples from the SG-NEx project, finding consistent patterns of degradation across all samples. In particular, the degradation curve is flat towards the 3' end, extends into a linear segment and plateaus as distance from the 3' end increases. The low degradation rate observed at the 3' end could be a consequence of protection due to poly(A) tailing and poly(A) binding proteins *in vivo* [52]. It also suggests that there is a lower bound on the minimum sequence length attainable with ONT sequencing technologies. In the next chapter, we develop a model that uses the estimated degradation curves for bias-aware transcript quantification.

Chapter 3

Bias-aware quantification

In this chapter, we specify a generative model for transcript quantification from long-read direct RNA-seq that accounts for bias due to the observed truncation of reads. We detail the assumptions of our model, formulate the generative process and derive statistical algorithms for inference of the parameters of the model.

3.1 Model assumptions

The input to our model consists of reads obtained from sequencing transcripts with the ONT direct RNA-seq protocol aligned to the reference transcriptome. We assume there is a bijective mapping between reads and transcript, i.e., each read originates from only one transcript, and each transcript generates only one read. Reads are generated independently and identically distributed (iid) from a distribution to be specified, and the 3' end of each read aligns within some region close to the 3' end of the isoform it originated from. Finally, we assume that a global expected degradation rate per base exists, i.e., the expected degradation curves of all isoforms are the same. This claim is supported by findings from the previous chapter.

3.2 Generative model

Our model aims to capture the generative process of the observed degraded reads from their isoforms of origin. Here, we formalize our model and derive the complete data log likelihood for maximum likelihood estimation of the model parameters.

3.2.1 Notation and formulation

Let N be the number of observed reads and M be the number of isoforms the reads were generated from.

- We observe a set of reads $\boldsymbol{R} = \{r_1, ..., r_N\}$, with r_i representing the i^{th} read. The variable r_i can be thought of as a collection of properties about the i^{th} read that are relevant for quantification, such as length, start and end positions or GC content.
- Our goal is to infer the unknown relative abundances of the isoforms $\boldsymbol{\theta} = \{\theta_1, ...\theta_M\}$. These relative abundances sum to one and are non-negative, i.e., $\Sigma_j \theta_j = 1, \theta_j \geq 0 \ \forall j$.
- We introduce N latent variables $\boldsymbol{Z} = \{z_1, ..., z_N\}$, where $z_i = (z_{i1} \ ... \ z_{iM})^T$ is a binary vector in M-dimensions with $\Sigma_j z_{ij} = 1$. z_i describes the assignment of r_i to one of M

isoforms, where

$$z_{ij} = \begin{cases} 1 & \text{if } r_i \text{ originated from isoform } j \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The directed graphical model in Figure 3.1 illustrates the relationship between the variables $\boldsymbol{R}, \boldsymbol{Z}$ and parameters $\boldsymbol{\theta}$, and provides an easy way to read off the factorization of the joint distribution over the observed and latent variables:

$$p(\boldsymbol{R}, \boldsymbol{Z}) = p(\boldsymbol{R} \mid \boldsymbol{Z}) \cdot p(\boldsymbol{Z}) \tag{3.2}$$

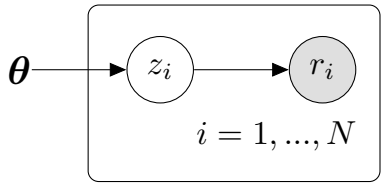

Figure 3.1: Graphical model for long-read RNA-seq. Nodes are random variables, and directed edges represent the dependencies between them. Shaded nodes are observed, while unshaded nodes are latent. Variables within the plate are replicated N times.

The distribution over the latent variables is categorical:

$$p(z_i) = \prod_{j=1}^{M} \theta_j^{z_{ij}} \tag{3.3}$$

The conditional distribution of observing the read r_i given z_i is

$$p(r_i \mid z_i) = \prod_{j=1}^{M} p(r_i \mid z_{ij})^{z_{ij}} \tag{3.4}$$

We now model the probability that read i originated from isoform j, i.e., $p(r_i \mid z_{ij} = 1)$. This probability is proportional to:

- The alignment compatibility $a_{ij} \in (0, 1)$ between read i and isoform j, which measures the alignment score (AS, as defined in SAM specification [53]) of read i against isoform j scaled by the best alignment score of read i against all isoforms:

$$a_{ij} = \frac{\mathrm{AS}_{ij}}{\max_{j'} \mathrm{AS}_{ij'}} \tag{3.5}$$

- The probability of observing a read of length ℓ_i given the degradation rate d_j of isoform j, i.e., $p(\ell_i \mid d_j, z_{ij} = 1)$, where ℓ_i is the length of r_i. We refer to this probability as the *read length-isoform agreement* model.

Combining alignment compatibility and the read length-isoform agreement model, we have

$$p(r_i \mid z_{ij} = 1) = a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \tag{3.6}$$

We make some remarks on equation 3.6. First, a_{ij} measures how well read i aligns to isoform j compared to its best alignment. Note that if we do not observe an alignment between read i and isoform j, then $\mathrm{AS}_{ij} = 0$, $a_{ij} = 0$, and thus $p(r_i \mid z_{ij} = 1) = 0$ as expected. Second, the read length-isoform agreement model captures the expected read length distribution for isoform j given its degradation rate and length. In the absence of degradation, we expect that the lengths of all reads originating from isoform j will be equal to the length $\mathrm{len}(j)$ of isoform j, i.e., $p(\ell_i = \mathrm{len}(j) \mid d_j, z_{ij} = 1) = 1$ and 0 otherwise.

3.2.2 Exact read length-isoform agreement

We specify the read length-isoform agreement model with the assumption that the expected degradation is constant, i.e., $\mathbb{E}[d] = c \in (0, 1)$. For example, if the degradation rate for all transcripts is $d_j = 0.2$, then for every kb from the 3' end, we expect 20% less reads. A constant expected degradation rate implies that the maximum possible sequenced read length $\ell_{\max}$ is bounded, and is given by:

$$-c \cdot \ell_{\max} + 1 = 0 \implies \ell_{\max} = 1/c \tag{3.7}$$

For instance, a $\mathbb{E}[d] = 0.2$ implies that $\ell_{\max} = 5$ kb. The assumption made here is unrealistic, given that ONT devices are capable of sequencing reads upwards of 10 kb in length. Nevertheless, this simplifying assumption enables us to test a simpler read length-isoform agreement model on simulated datasets. We develop a more flexible model allowing for variable degradation in the following section.

Let the length of isoform j in kb be $\mathrm{len}(j)$. For constant expected degradation, we have

$$p(\ell_i \mid d_j, z_{ij} = 1) = \begin{cases} \dfrac{d_j}{10^3} & \text{if } \ell_i < \min(\mathrm{len}(j), \ell_{\max}) \\ 1 - d_j \cdot \mathrm{len}(j) & \text{if } \ell_i = \mathrm{len}(j) \end{cases} \tag{3.8}$$

We explain Eq. 3.8 beginning with the simpler case, where the read length ℓ_i is equal to the length of the isoform $\mathrm{len}(j)$. This implies that the read is a full-length read; the probability of observing this read is thus 1 − the probability of observing a degraded read, which is $d_j \cdot \mathrm{len}(j)$ (see Section 2.2). When expected degradation is constant, the degraded read lengths are uniformly distributed (discretely). For an isoform of length $\mathrm{len}(j)$ in kb, there are $\mathrm{len}(j) \cdot 10^3$ possible degraded read lengths. We thus divide the probability of observing a degraded read by this quantity, which yields $d_j/10^3$.

3.2.3 Empirical read length-isoform agreement

To model real data, we develop an empirical read length-isoform agreement model based on the degradation curves derived from Section 2.2. This model has no constraints on the degradation rate. Let X be a random variable denoting the length of degraded reads, and $F(x) = P(X \leq x)$ be the corresponding empirical cumulative distribution function on X. Then, we have

$$p(\ell_i \mid d_j, z_{ij} = 1) = \begin{cases} \lim_{\delta \to 0}[F(\ell_i + \delta) - F(\ell_i - \delta)] & \text{if } \ell_i < \min(\text{len}(j), \ell_{\max}) \\ 1 - F(\text{len}(j)) & \text{if } \ell_i = \text{len}(j) \end{cases} \tag{3.9}$$

Recall again from section 2.2 that the probability of observing a read of length ℓ_i or greater is $S(\ell_i) = 1 - F(\ell_i)$. Thus, for a transcript of length $\text{len}(j)$, the probability of observing a full length read is simply $1 - F(\text{len}(j))$. The corresponding probability of observing degraded reads is $F(\text{len}(j))$. To compute the probability of observing a point read length given a empirical continuous density, we take limits:

$$\begin{aligned} P(X = \ell_i) &= \lim_{\delta \to 0}[P(X < \ell_i + \delta) - P(X < \ell_i - \delta)] \\ &= \lim_{\delta \to 0}[F(\ell_i + \delta) - F(\ell_i - \delta)] \end{aligned} \tag{3.10}$$

Since the length of isoform j is $\text{len}(j)$, we normalize this quantity by considering probabilities only up to $\text{len}(j)$:

$$\lim_{\delta \to 0} \frac{F(\ell_i + \delta) - F(\ell_i - \delta)}{F(\text{len}(j))} \tag{3.11}$$

Finally, we multiply the above expression by the probability of observing degraded reads $F(\text{len}(j))$, arriving at the quantity in Eq. 3.9.

3.3 Parameter inference

In this section, we derive an expectation maximization algorithm to infer the parameters $\boldsymbol{\theta}$ that maximize the likelihood of the observed data.

3.3.1 Likelihood formulation

The marginal distribution of r_i is obtained by summing the joint distribution (Eq. 3.2) over all M possible states of z_i. From Eqs. 3.3, 3.4 and 3.6, we obtain

$$p(r_i) = \sum_{z_i} p(z_i) \cdot p(r_i \mid z_i) = \sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \tag{3.12}$$

Since we assume the reads are generated iid, the likelihood is a product over N reads:

$$p(\boldsymbol{R} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \tag{3.13}$$

The corresponding log-likelihood is

$$\log p(\boldsymbol{R} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \right] \tag{3.14}$$

It is difficult to maximize Eq. 3.14 with respect to $\boldsymbol{\theta}$ because the log acts on the sum. If we had access to the latent variables $\boldsymbol{Z}$, the likelihood will decompose. To demonstrate this, we write the complete data likelihood as

$$p(\boldsymbol{R}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \theta_j^{z_{ij}} \cdot [a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1)]^{z_{ij}} \tag{3.15}$$

The corresponding complete data log-likelihood is

$$\log p(\boldsymbol{R}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \cdot \log \left[\theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \right] \tag{3.16}$$

3.3.2 Expectation maximization

The EM algorithm is a powerful method for finding maximum likelihood estimates of parameters in latent variable models, and is used in many RNA-seq quantification methods for parameter inference [12, 47, 48, 54–56]. It consists of a two-stage iterative optimization procedure for maximizing the likelihood. In our setting, we seek to maximize the complete data log-likelihood with respect to $\boldsymbol{\theta}$. However, we observe only the incomplete data $\boldsymbol{R}$. Since we cannot obtain the complete data log-likelihood (Eq. 3.16), we take its expectation under the posterior of the latent variables $\boldsymbol{Z}$ given some setting of the parameters $\boldsymbol{\theta}^{(t)}$: this corresponds to the E step of the EM algorithm. In the M step, we maximize this expectation with respect to the parameters $\boldsymbol{\theta}$ to obtain a new estimate of the parameters $\boldsymbol{\theta}^{(t+1)}$.

We obtain the posterior of the latent variables $\boldsymbol{Z}$ from Eqs. 3.3 and 3.2:

$$p(\boldsymbol{Z} \mid \boldsymbol{R}, \boldsymbol{\theta}) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} [\theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij})]^{z_{ij}=1} \tag{3.17}$$

E-step We take the expectation of the complete data log-likelihood (Eq. 3.16) with respect to the posterior distribution of the latent variables (Eq. 3.17):

$$\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{R},\boldsymbol{\theta}} \left[\log p(\boldsymbol{R}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \right] = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbb{E}[z_{ij}] \cdot \log \left[\theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \right] \tag{3.18}$$

where

$$\mathbb{E}[z_{ij}] = \frac{\theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1)}{\sum_{j'} \theta'_j \cdot a_{ij'} \cdot p(\ell_i \mid d_{j'}, z_{ij'} = 1)} = \gamma_{ij} \tag{3.19}$$

M-step We maximize the expression in Eq. 3.18 with respect to $\boldsymbol{\theta}$ constrained by $\sum_j \theta_j = 1$, which is equivalent to the following:

$$\text{argmax}_{\boldsymbol{\theta}} \sum_i \sum_j \gamma_{ij} \cdot \log \theta_j \tag{3.20}$$

The Lagrangian is

$$\mathcal{L}(\theta_j, \lambda) = \sum_i \sum_j \gamma_{ij} \cdot \theta_j + \lambda \left[\sum_j \theta_j - 1 \right] \tag{3.21}$$

Taking the derivative with respect to θ_j and setting to 0 we have

$$\frac{\partial \mathcal{L}(\theta_j, \lambda)}{\partial \theta_j} = \sum_i \frac{\gamma_{ij}}{\theta_j} + \lambda = 0 \implies \theta_j = -\frac{\sum_i \gamma_{ij}}{\lambda} \tag{3.22}$$

Obtain λ by observing that

$$1 = \sum_j \theta_j = \sum_j -\frac{\sum_i \gamma_{ij}}{\lambda} = \sum_i -\frac{\sum_j \gamma_{ij}}{\lambda} = \sum_i -\frac{1}{\lambda} = -\frac{N}{\lambda} \implies \lambda = -N \tag{3.23}$$

From Eqs. 3.22 and 3.23 we find that

$$\theta_j = \frac{\sum_i \gamma_{ij}}{N} \tag{3.24}$$

An intuitive interpretation of Eq. 3.24 is that the relative abundance of isoform j is the fraction of all reads that are soft-assigned to isoform j by the EM algorithm. In summary, the EM algorithm for our model is as follows:

1. Initialize $\boldsymbol{\theta}^{(0)}$
2. E step: compute $\gamma_{ij}^{(t)} = \frac{\theta_j^{(t)} \cdot a_{ij} \cdot p(\ell_i | d_j, z_{ij}=1)}{\sum_{j'} \theta_{j'}^{(t)} \cdot a_{ij'} \cdot p(\ell_i | d_{j'}, z_{ij'}=1)}$
3. M step: compute $\theta_j^{(t+1)} = \frac{\sum_i \gamma_{ij}^{(t)}}{N}$
4. Repeat steps 2 and 3 until convergence.

To initialize $\boldsymbol{\theta}$, we assign equal relative abundances to all isoforms by setting $\theta_j = 1/M \ \forall j$. We test for convergence by computing changes in the parameters or the log-likelihood in Eq. 3.14 after each iteration of the algorithm. Once the change in the parameters or log-likelihood falls below a certain threshold, we terminate the algorithm. Proof of the concavity of the log-likelihood function, and thus the optimality of the inferred parameters, can be found in the Appendix D.

3.4 Discussion

In this chapter, we formulated a generative process for long-read RNA-seq. For modeling degradation, we introduced the read length-isoform agreement model, which models the expected read length distribution of an isoform given its degradation rate or curve. In particular, we developed two variations of the bias model: the first is an exact model that assumes constant degradation rate. This is useful for testing our full model on simulation datasets and serves as a sanity check. The next is an empirical model that is based on the degradation curve. This approach does not place constraints on the degradation, and can be used flexibly on real data. We incorporated these models into an EM algorithm for bias-aware transcript quantification. In the following chapter, we evaluated our model and inference algorithm on simulated and real datasets, and benchmark its performance against that of other methods.

Chapter 4

Model evaluation and results

In this chapter, we evaluate our quantification model and inference algorithms on simulated datasets with constant expected degradation rates and real direct RNA-seq datasets with sequencing spike-ins.

4.1 Read alignment

Simulated and real reads were aligned with minimap2 (v.2.17-r941) [57, 58], a popular aligner for long reads. For alignment to the genome, we used the flags `-ax splice -uf -k14` that take into account splicing and forward strand alignment for ONT direct RNA-seq as recommended by minimap2 developers. For alignment to the transcriptome, we use default flags `-ax map-ont -N10` for mapping ONT reads, keeping 10 secondary alignments due to similarities in transcript isoform sequences. We used the same genomic or transcriptomic alignments across the tools, depending on which is required as input.

4.2 Model variations

We compare two variations of our model based on the read length-isoform agreement:

Deg. EM (exact) We ran our model with the exact read length-isoform agreement model (Eqn. 3.8) and provided the known degradation rate to the model for inference. This is of course unrealistic, since in real data, the degradation rate is not known *a priori* and there is no bound on the maximum read length (i.e., the degradation rate is not constant). Nevertheless, this serves as a useful sanity check that our approach works.

Deg. EM (emp.) The second variation of our model uses the empirical read length-isoform agreement model (Eqn. 3.9) and estimates the degradation rate per base from the data. This variation has no limitation on the maximum read length nor constraints on the degradation rate.

4.3 Methods for benchmarking

We benchmark our model against three existing methods for transcript quantification from long-read RNA-seq.

Bambu Bambu [48] is a method for reference guided transcript discovery and quantification for long read RNA-seq data. Crucially, Bambu is one of the few existing long-read quantification methods that models degradation bias. For benchmarking, we ran Bambu (v.2.0.6) without transcript discovery. For transcript quantification, we ran Bambu with bias correction (default) and without bias correction (`degradationBias=FALSE`). We used the defaults for all other parameters.

FLAIR Full-Length Alternative Isoform analysis of RNA (FLAIR) [46] is a method for correction, isoform definition and alternative splicing analysis of long reads. For benchmarking, we ran FLAIR (v.1.5.1) with the modules `correct`, `collapse` and `quantify`. When running FLAIR on simulated data, we set `--support 1` for FLAIR `collapse`, keeping isoforms that are supported by minimally one read (default=3).

NanoCount NanoCount [47] is a method for quantifying isoform abundance from ONT direct RNA-seq data. Of the three methods listed here, it is the most comparable to our model as it is tailored for direct RNA-seq and uses an expectation maximization algorithm for estimating isoform abundance estimates. For benchmarking, we ran NanoCount (v.1.0.0.post6) with default parameters.

4.4 Evaluations on simulated data

To evaluate our model's ability to correct for degradation bias, we simulated five datasets for a range of degradation rates $\mathbb{E}[d] \in \{0.05, 0.1, 0.2, 0.4, 0.5\}$ (Appendix A). In addition, we simulated reads for artificial novel isoforms that are modified by dropping exons from the 5' end of selected reference isoforms, termed *subset* isoforms (Appendix B, Fig. B.1). This increases the proportion of multi-mapping reads and makes correcting for degradation bias crucial for accurate transcript quantification.

The read counts for simulation follow a negative binomial distribution, which is often used for modeling RNA-seq counts and other count data that is over-dispersed, i.e., where the assumption of equal mean and variance is not held [59–61]. We parameterise the distribution with the mean μ and a dispersion parameter α such that the variance is given by $\mu + \alpha\mu^2$. This is the same parameterisation used in [61]. To ensure that the negative binomial is a valid choice of distribution, we fit discrete distributions to the counts returned by existing methods on real long-read RNA-seq data, and find that the negative binomial provides a better fit to the data compared to the Poisson distribution (Appendix C).

4.4.1 Comparisons between model variations

In this section, we compare the Deg. EM (exact) and Deg. EM (emp.) models based on the isoform abundance estimates obtained. We compare these estimates against the simulated ground truth based on Spearman correlation (SCC), normalized root-mean-squared error (NRMSE) and median relative difference (MRD). Explanations of these metrics can be found in Appendix E.

Both variations of the model perform comparably well on the simulated data, achieving SCC > 0.775 across all five simulated datasets with low NRMSE and MRD. Table 4.1 shows the mean of each metric across the five datasets for all and subset isoforms. We first note that performance on the subset isoforms is slightly poorer compared to that on all isoforms, for both variations and across metrics. This is expected as there is more ambiguity in read assignment with the subset isoforms.

	All isoforms			Subset isoforms		
Method	SCC	NRMSE	MRD	SCC	NRMSE	MRD
Deg. EM (exact)	0.848	0.429	**0.015**	**0.813**	**0.44**	**0.127**
Deg. EM (emp)	**0.849**	**0.427**	0.016	0.801	0.434	0.191

Table 4.1: Summary of metrics across simulated datasets for model variations. We report the mean SCC, NRMSE and MRD across the five datasets for all isoforms and subset isoforms separately. Bold values indicate the best performance for each column.

Next, we examine the performance on each dataset separately for all and subset isoforms (Fig. 4.1). Across all isoforms, the variations perform extremely similarly. In contrast, on the subset isoforms, the exact model dominates the empirical model in MRD. This is expected, as estimations with the empirical cumulative distribution function introduce noise to the estimates. Conversely, the exact model is formulated exactly to work on the data.

To investigate these differences further, we visualised estimates and fitted a kernel density estimate on the subset isoforms only for datasets with degradation rates 0.2 and 0.5 (Fig. 4.2). On a dataset with relatively lower degradation rate (0.2), the exact model performs qualitatively well, with the density of points over the diagonal (Fig. 4.2a), while the empirical model underestimates counts for certain isoforms (Fig. 4.2b). We see similar results in datasets with higher degradation rates (0.5) where the variance in the estimates returned by the exact model is lower (Fig. 4.2c) compared to the empirical model, where the counts are more diffuse about the diagonal (Fig. 4.2d). This deviation from the diagonal also explains the increase in MRD in the empirical model for larger degradation rates compared to the exact model (Fig. 4.1).

Based on our analyses in section 2.1, we observed that average degradation rates in real direct RNA-seq data tend to be in the range of 0.1-0.2. In this range, the two variations perform comparably, and have a high SCC with each other ($\mathbb{E}[d]$=0.1, SCC=0.975, $\mathbb{E}[d]$=0.2, SCC=0.979).

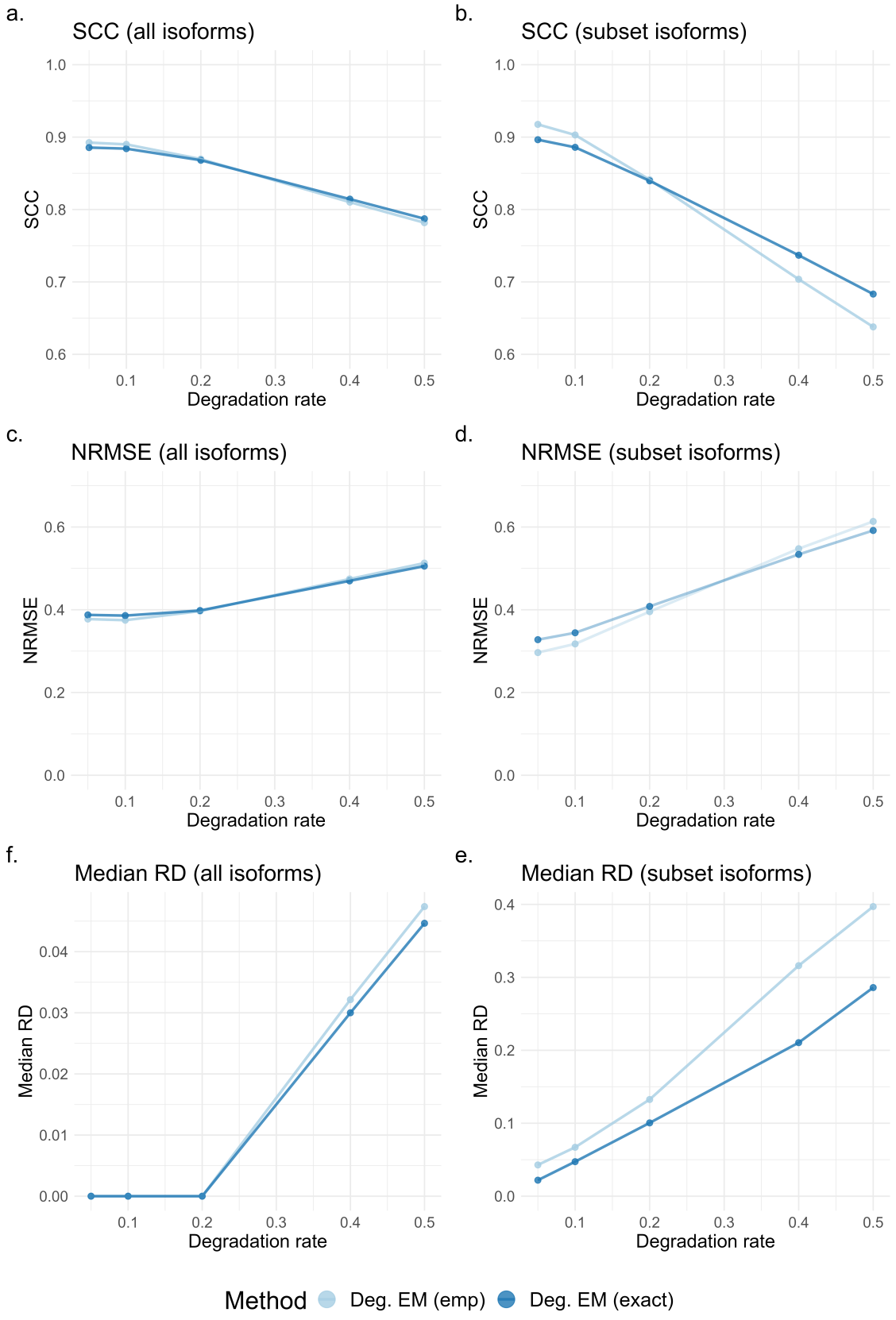

Figure 4.1: SCC, NRMSE and MRD across simulated datasets for model variations. Here, each point is a dataset with constant expected degradation $\mathbb{E}[d] = \{0.05, 0.1, 0.2, 0.4, 0.5\}$. **a.** SCC on all isoforms. **b.** SCC on subset isoforms. **c.** NRMSE on all isoforms. **d.** NRMSE on subset isoforms. **f.** MRD on all isoforms. **e.** MRD on subset isoforms.

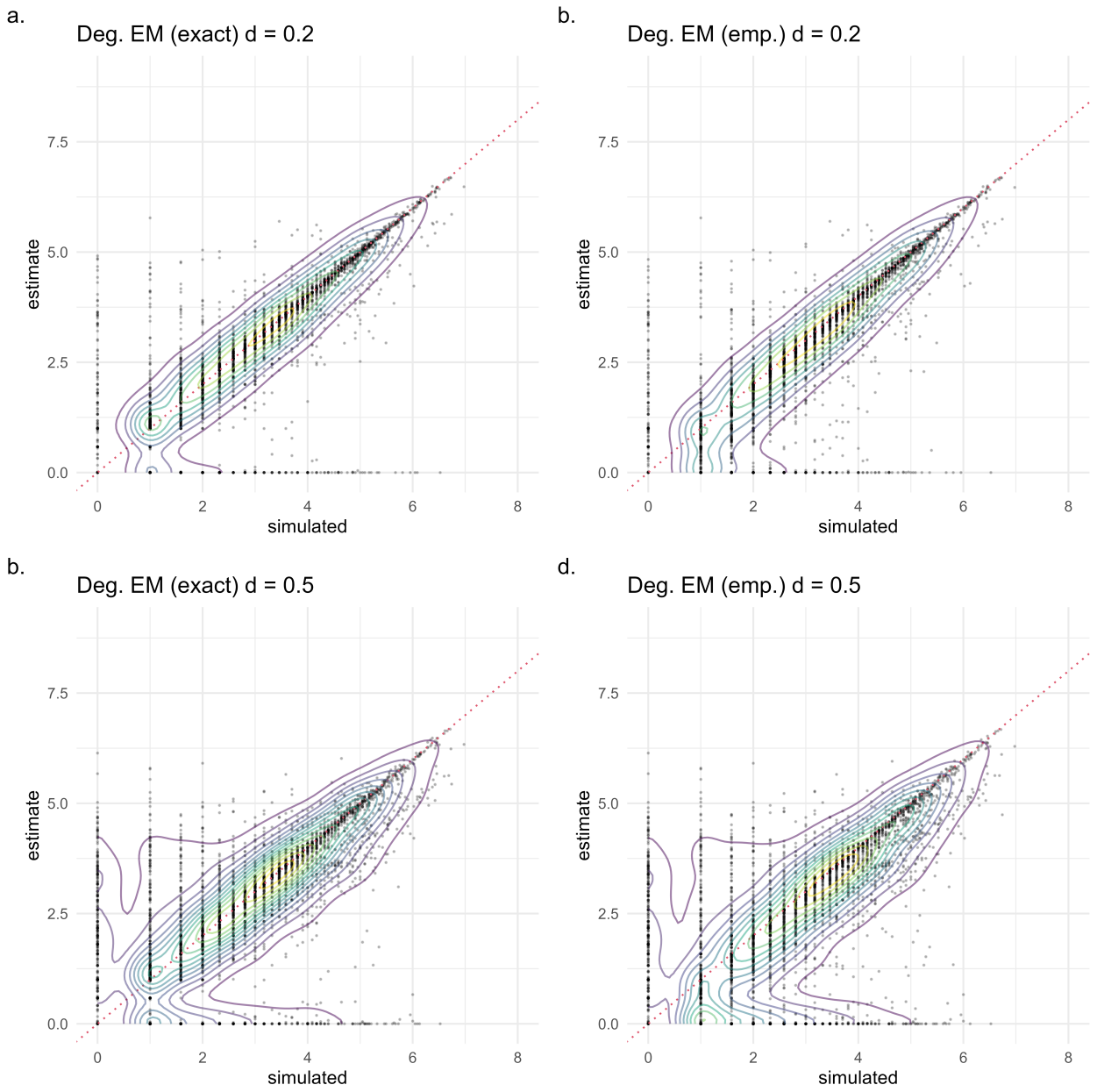

Figure 4.2: Scatter plots of the simulated and estimated counts on the log2 scale across simulated datasets with $\mathbb{E}[d] = 0.2$ and $\mathbb{E}[d] = 0.5$ for the exact and empirical model. **a.** Exact model with degradation 0.2. **b.** Empirical model with degradation 0.2. **c.** Exact model with degradation 0.5. **d.** Empirical model with degradation 0.5.

4.4.2 Comparisons with existing methods

We now benchmark our model against Bambu with and without bias modeling, FLAIR and NanoCount by comparing the estimates returned by each method with the simulated ground truth. We analyse counts for all isoforms that we included in the simulation and where the sum of counts across the methods was greater than zero. In addition, we make a distinction between methods that are *bias-aware* (Deg. EM (exact), Deg. EM (emp.), Bambu) and methods that are *bias-unaware* (Bambu (no bias), FLAIR, NanoCount).

All methods perform reasonably well on the simulated data across all isoforms, achieving SCC > 0.69 across all five simulated datasets (Table 4.2). NRMSE across the methods are comparable, but MRD for the other methods are about one order of magnitude larger than those attained by our model. However, we observe stark differences between the bias-aware and bias-unaware methods on the subset isoforms. In particular, bias-unaware methods perform considerably poorer on SCC and MRD than bias-aware ones (Table 4.2).

	All isoforms			Subset isoforms		
Method	SCC	NRMSE	MRD	SCC	NRMSE	MRD
Deg. EM (exact)	0.86	0.423	**0.011**	**0.834**	**0.431**	**0.104**
Deg. EM (emp.)	**0.861**	**0.421**	0.012	0.826	0.424	0.154
Bambu	0.771	0.599	0.101	0.671	1.009	0.36
Bambu (no bias)	0.739	0.637	0.107	0.432	1.165	0.926
FLAIR	0.696	0.697	0.081	0.153	1.176	0.974
NanoCount	0.759	0.566	0.105	0.408	1.056	0.931

Table 4.2: Summary of metrics across simulated datasets for different methods. We report the mean SCC, NRMSE and MRD across the five datasets for all isoforms and subset isoforms separately. Bold values indicate the best performance for each column.

Next, we examine the performance on each dataset separately for all the methods on all and subset isoforms separately (Fig. 4.3). Both Deg. EM (exact) and Deg. EM (empirical) show better performance on all metrics compared to other methods across the datasets. We also note that Bambu with bias modeling improves the performance over Bambu without bias modeling across the whole range of degradation rates. In addition, Bambu's performance drops quickly as the degradation rate increases, as its degradation rate is purposefully calibrated or moderated to lower degradation rates. The authors note that this is to allow for comparisons across samples or replicates, as correction for different degradation rates across samples may result in technical variation which impact downstream analyses. We examine this claim based on reproducibility metrics on real data in the following section. For degradation rates commonly observed in real data (0.1-0.2), Bambu performs better compared to FLAIR and NanoCount on both SCC and NRMSE (Fig. 4.1a,c).

Bias-unaware methods perform poorly over the subset isoforms across the whole range of degradation rates. Interestingly, the MRD on subset isoforms for bias-unaware methods remained consistently high and was invariant across the range of degradation rates, while for bias-aware methods, the MRD was always lower compared to bias-unaware methods and increased with the degradation rate (4.3e). This demonstrates the importance of bias correction and illustrates the difficulty of correcting for degradation bias when the degradation rate is high.

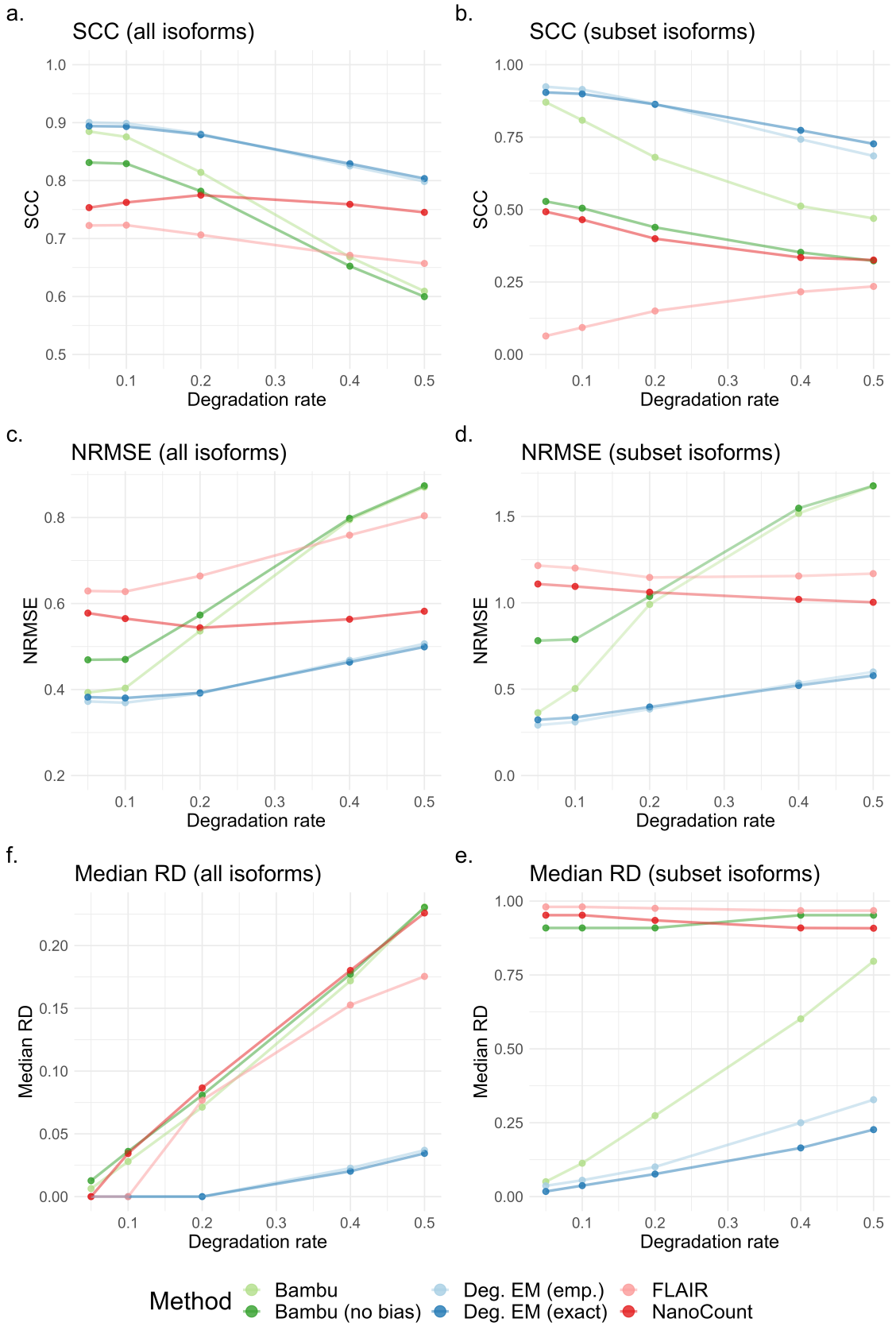

Figure 4.3: SCC, NRMSE and MRD across simulated datasets for different methods. Here, each point is a dataset with constant expected degradation $\mathbb{E}[d] = \{0.05, 0.1, 0.2, 0.4, 0.5\}$. **a.** SCC on all isoforms. **b.** SCC on subset isoforms. **c.** NRMSE on all isoforms. **d.** NRMSE on subset isoforms. **f.** MRD on all isoforms. **e.** MRD on subset isoforms.

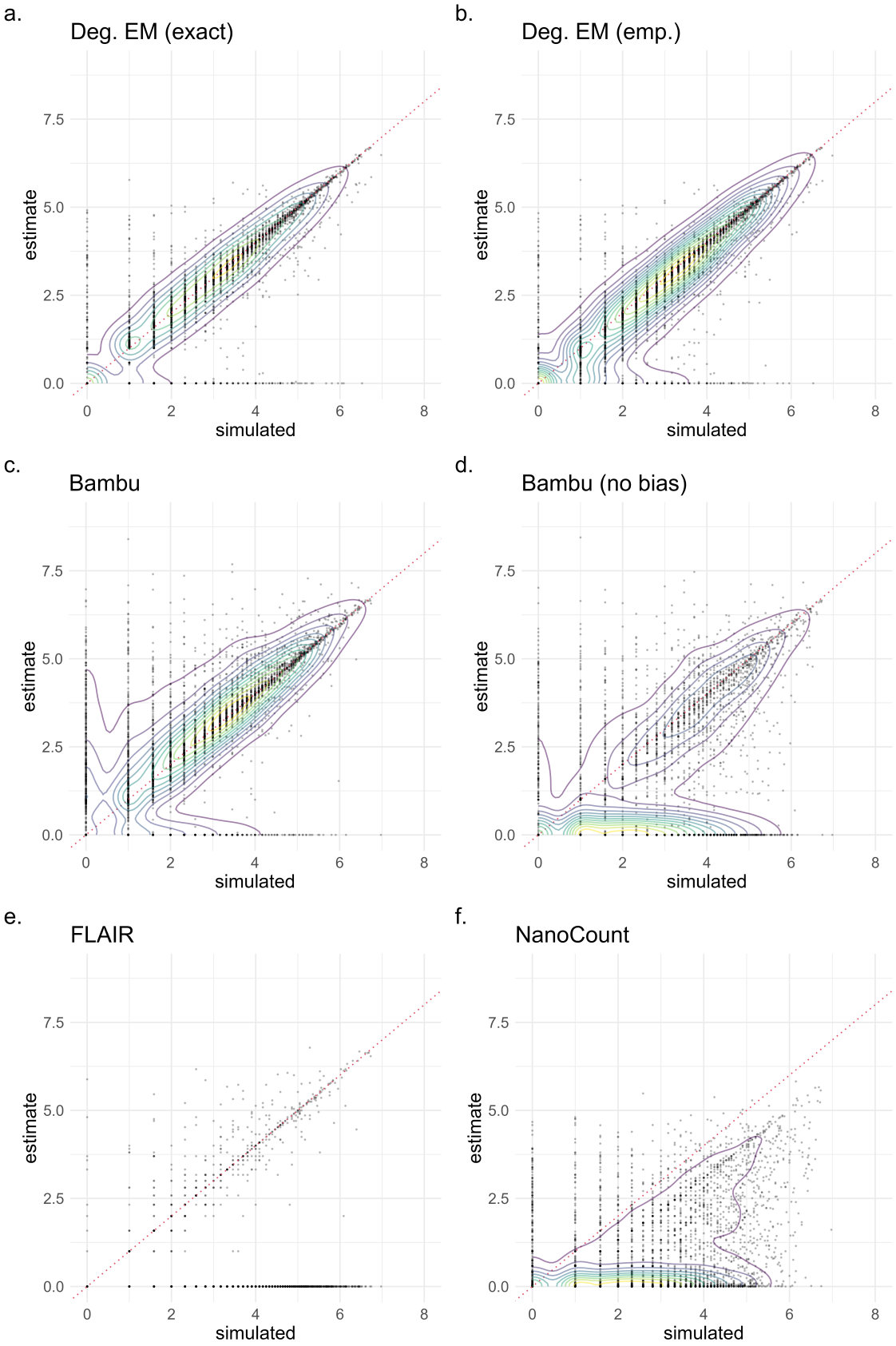

Figure 4.4: Scatter plots of the simulated and estimated counts on the log2 scale across simulated datasets with $\mathbb{E}[d] = 0.2$ for all methods. **a.** Deg. EM (exact) model. **b.** Deg. EM (emp.) model. **c.** Bambu with bias modeling. **d.** Bambu with no bias modeling. **e.** FLAIR. **f.** NanoCount.

Finally, we visualised estimates and fitted a kernel density estimate on the subset isoforms for all methods on the dataset with a degradation rate of 0.2 (Fig. 4.4), which we observe to be common in real data. On this dataset, bias-aware methods significantly improve over bias-unaware ones qualitatively (Fig. 4.4a,b,c). In particular, without correcting for bias, Bambu (no bias), FLAIR and NanoCount (Fig. 4.4d,e,f) severely underestimate the counts for subset isoforms to different extents, with FLAIR performing the worst.

4.4.3 Runtime analysis

We measured the end-to-end runtime taken for all methods (Fig. 4.5) excluding the read alignment step. Bambu and FLAIR were run with 12 threads, while NanoCount and the current version of our model do not implement parallelisation. Across the five datasets, NanoCount was the fastest, followed by Deg. EM (exact). Bambu and FLAIR performed similarly, though FLAIR had a much larger variance in the runtime. Even though the time complexity for both Deg. EM (exact) and Deg. EM (emp.) is $\mathcal{O}(N_A)$, where N_A is the number of alignments, Deg. EM (emp.) is much slower than all other methods because calculation of the empirical read length-isoform agreement probabilities is computationally intensive. Even though the runtimes for the empirical model are longer, they are still tolerable in absolute terms, and come with improvements in accuracy for accurately quantifying degraded reads. Nevertheless, this is one area for improvement for future iterations of our model.

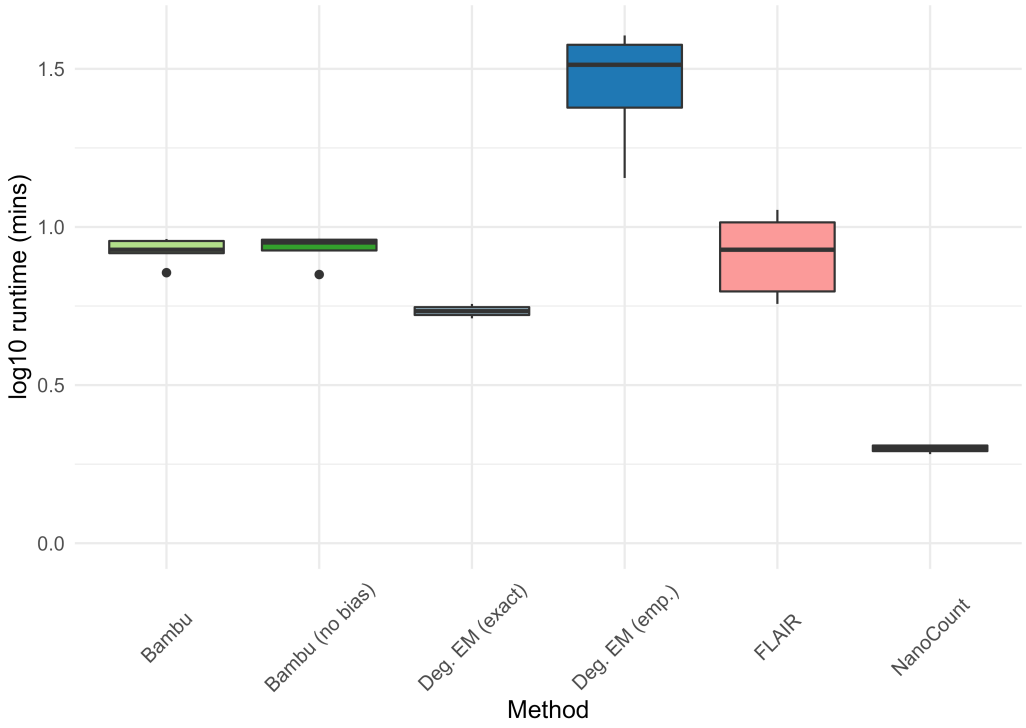

Figure 4.5: Runtime (mins) across simulated datasets for different methods. Times are log10 transformed.

4.5 Evaluations on real data

In this section, we evaluate Deg. EM (emp.), Bambu with bias modeling, and NanoCount on direct RNA-seq data from the SG-NEx project. In particular, we examine the performance of these methods on samples containing RNA sequencing spike-ins (Table 4.3). A subset of samples contains Spike-In RNA Variants (SIRVs) [51] while another contains RNA sequins [62]. The expression of these spike-ins are supposedly known and can be used as a reference.

Cell Line	Run	Replicate	Spike-in
Hct116	1	3	Sequin Mix A v1.0
K562	1	4	Sequin Mix A v1.0
K562	1	5	Sequin Mix A v1.0
MCF7	1	4	Sequin Mix A v1.0
H9	1	2	SIRV-4
H9	1	3	SIRV-4
H9	1	4	SIRV-4
H9	2	2	SIRV-4
H9	2	3	SIRV-4
H9	2	4	SIRV-4

Table 4.3: Description of SG-NEx direct RNA-seq samples. A subset of samples is spiked-in with Sequins while another is spiked-in with SIRV-4.

4.5.1 Empirical results on spike-ins

First, we examine results on the RNA sequin spike-ins (n = 41 transcripts). For each of the four samples, we computed the SCC, NRMSE and MRD on the abundance estimates returned by each method (Fig. 4.6). Deg. EM (emp.) achieved the highest median SCC, while NanoCount perform reasonably well with a median SCC > 0.7. Bambu's performance was weighed down by an outlier (Fig. 4.6a). NRMSE and MRD metrics show similar trends, with Deg. EM (emp.) achieving the lowest median NRMSE and MRD (Fig. 4.6b,c).

We visualize the estimates and supposed concentrations of the sequins for the sample K562 replicate 5 run 1 (Fig 4.7) which obtained close to median performance across the metrics for all methods. Qualitatively, Deg. EM and NanoCount achieve similar results, though in the latter, there is a slight overestimation of counts. Bambu underestimates the counts for a subset of lowly expressed sequins that results in a decrease in performance over the three metrics.

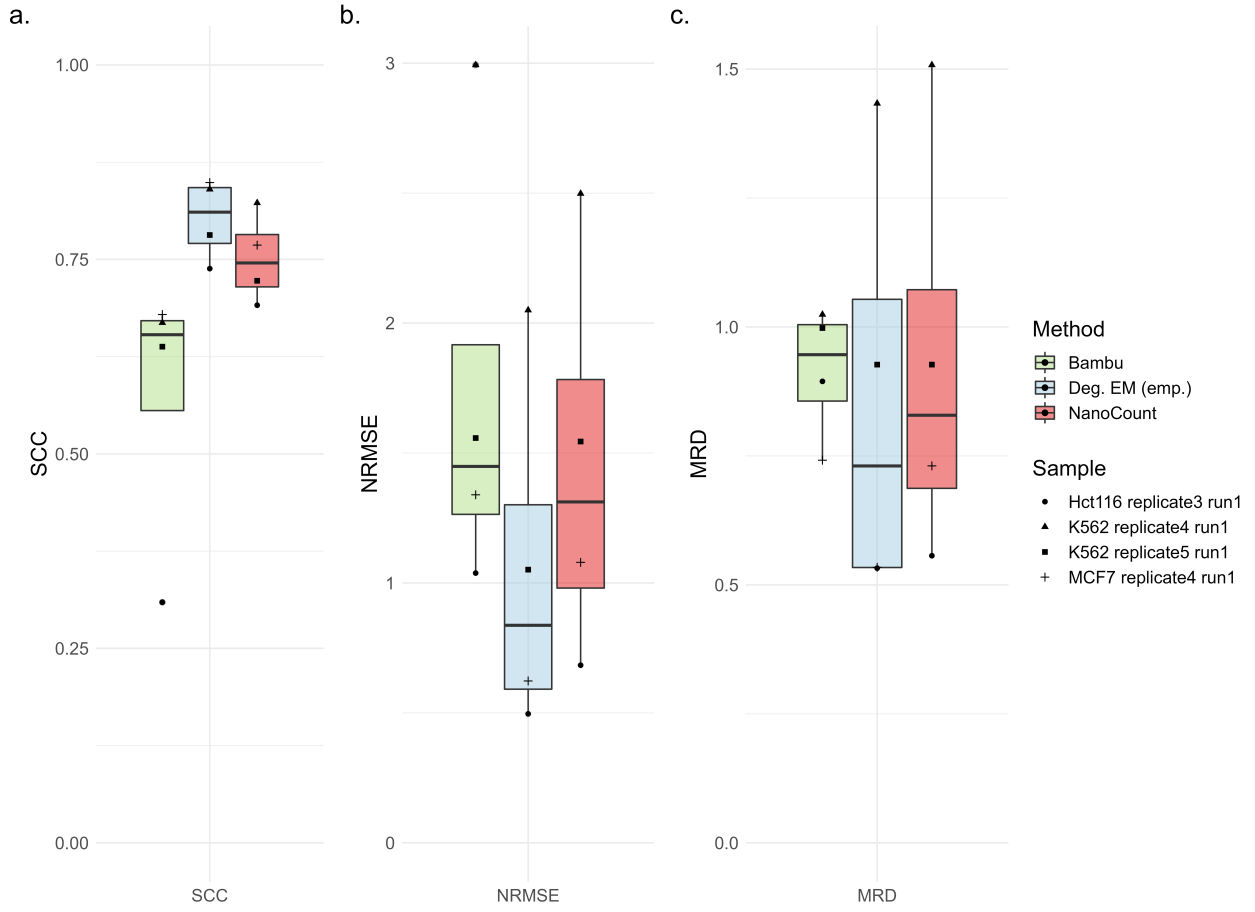

Figure 4.6: SCC, NRMSE and MRD on RNA sequins in SG-NEx data. **a.** SCC, **b.** NRMSE and **c.** MRD for each of the three methods across four samples.

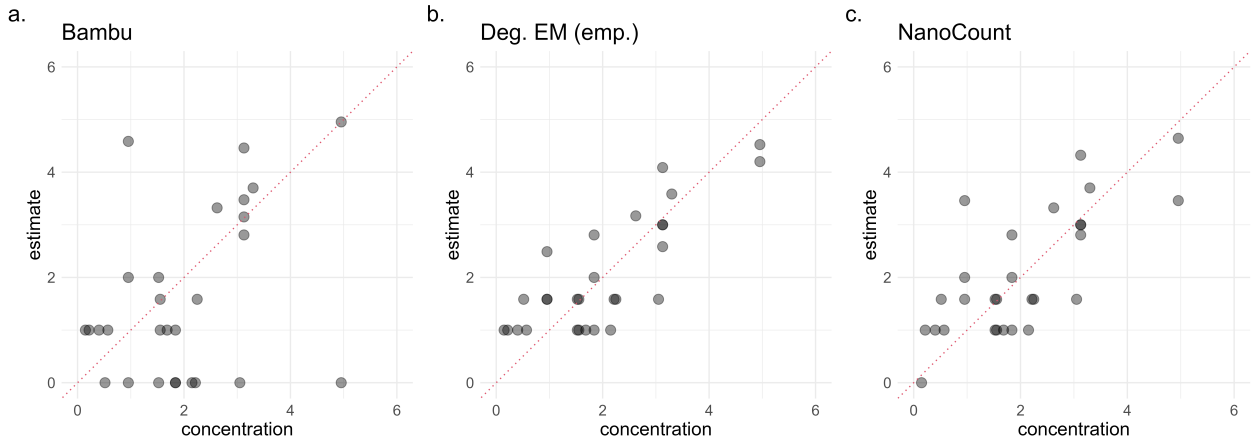

Figure 4.7: Scatter plots on RNA sequins in sample K562 replicate 5 run 1 for **a.** Bambu, **b.** Deg. EM (emp.) and **c.** NanoCount. Plots for rest of the samples are qualitatively similar.

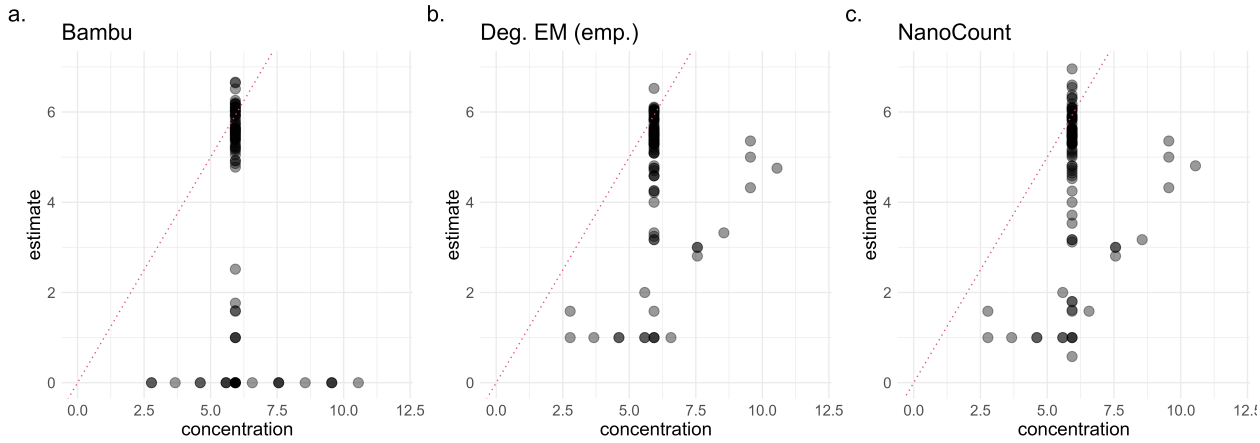

Figure 4.8: Scatter plots on SIRVs in SG-NEx data in H9 replicate 4 run 1 for **a.** Bambu, **b.** Deg. EM (emp.) and **c.** NanoCount. Plots for rest of the samples are qualitatively similar.

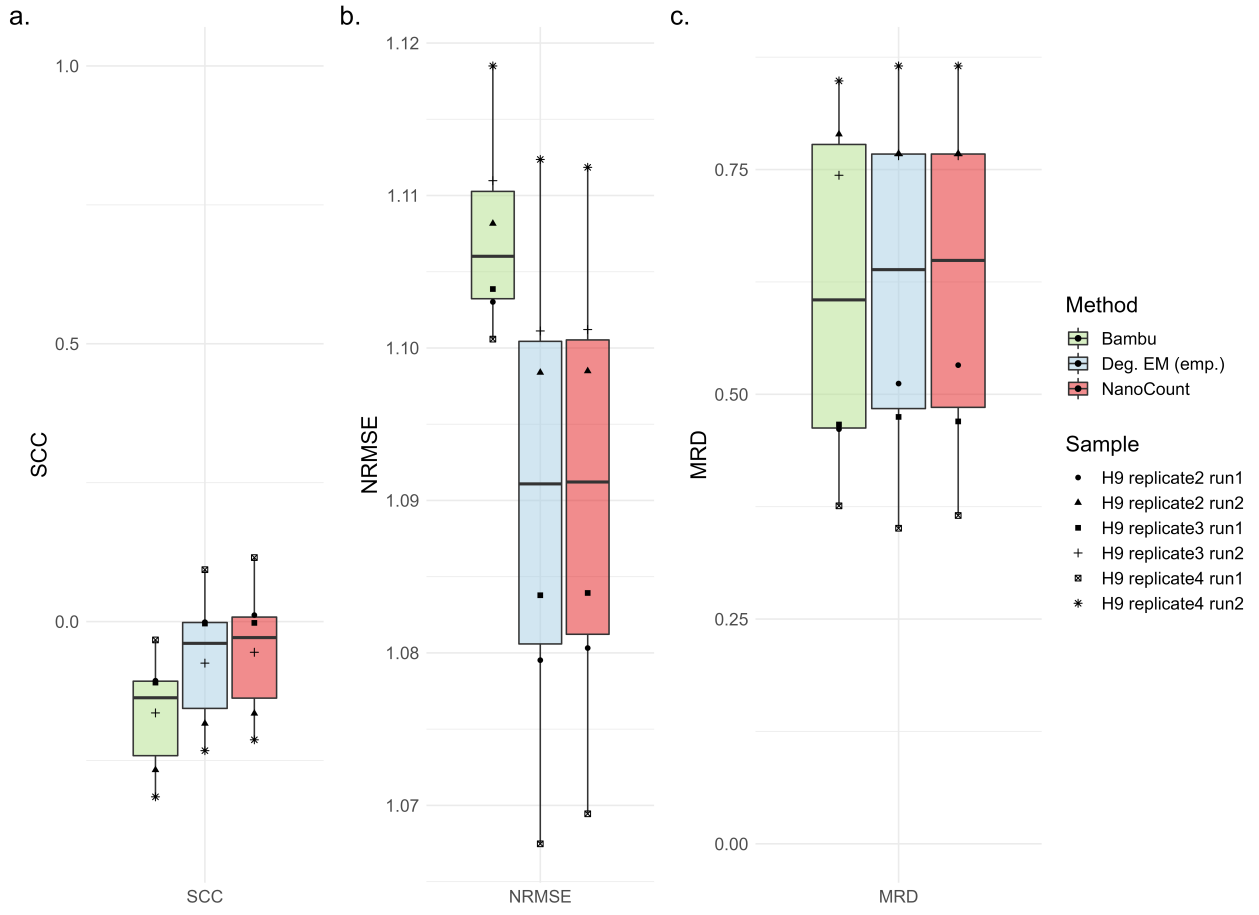

Figure 4.9: SCC, NRMSE and MRD on SIRV-4 in SG-NEx data. **a.** SCC, **b.** NRMSE and **c.** MRD for each of the three methods across six H9 samples.

In addition to the sequin spike-in samples, we examined the SIRV spike-ins in six H9 direct RNA-seq samples (Table 4.3). For all methods and across the samples, the abundance estimates

for SIRVs did not reflect the supposed concentration (Fig. 4.8). However, the estimates obtained across the methods correlated well with each other (median of median pairwise SCC = 0.809, p-value < 2.2e-16). This might suggest that the supposed SIRV concentrations do not accurately represent what is actually present in the samples, diminishing their utility as a control for RNA-seq in this context. Nevertheless, we computed the SCC, NRMSE and MRD across the six samples for each method (Fig. 4.9). The performance on all metrics is similar across the methods, but is unlikely to be informative due to reasons discussed above.

4.5.2 Reproducibilty measures

Besides quantifying accuracy, we can also measure the reproducibility of abundance estimates for each method. We do so with the reproducibility metric (RM), which is a measure of the standard deviation of abundance estimates across replicates (Appendix E). We use H9 samples in the SG-NEx data that were sequenced in two runs, with three replicates in each run (Table 4.3). Within each run, we calculated the RM across the replicates for each isoform called as expressed in at least one method and obtained the median reproducibiltiy metric (MRM) (Table 4.4). For all methods, MRM in the first run was higher than that in the second by virtue of the number of isoforms the MRM was calculated over (Run 1 = 30,720, Run 2 = 21,536). In both runs, Deg. EM (emp.) achieved the lowest MRM suggesting higher reproducibility compared to NanoCount or Bambu.

Method	Run 1	Run 2
Deg. EM (emp.)	**0.886**	**0.772**
Bambu	1.193	0.943
NanoCount	0.943	0.816

Table 4.4: MRM across different runs for SG-NEx H9 samples.

The incorporation of bias modeling in Deg. EM improves reproducibility. We can infer this by comparing the MRMs obtained by Deg. EM and NanoCount since the latter is most comparable to our Deg. EM. By modeling the bias, read-to-isoform assignment is improved across replicates with potentially different degradation rates, thus stabilising variance in abundance estimates across the replicates.

4.5.3 Comparisons of estimated degradation

Lastly, we compare the estimated average degradation rates obtained by Bambu and Deg. EM (emp.) across samples from four cell lines. We note that we do not expect the absolute values of the degradation rates to be similar; rather, we expect the estimates to be linearly related. Indeed,

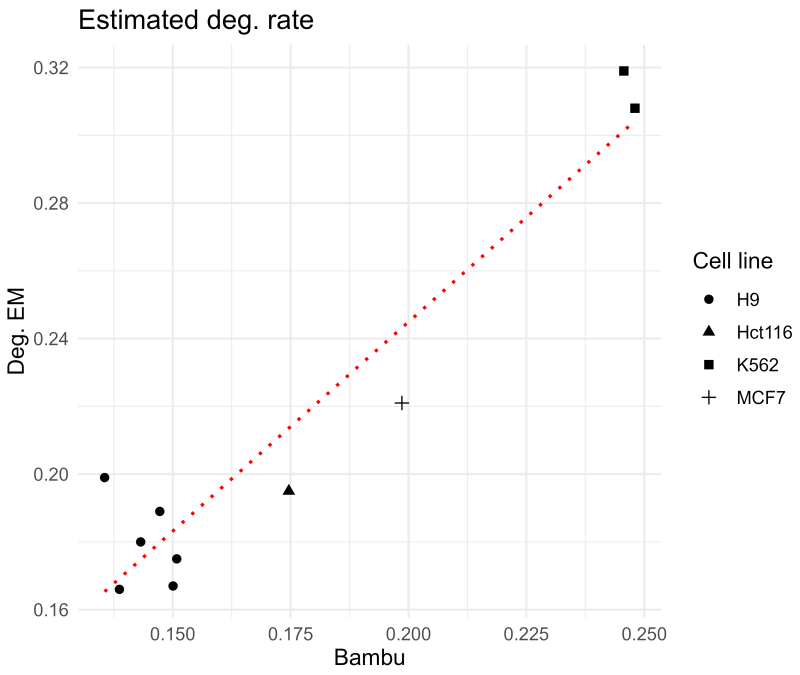

Figure 4.10: Comparison of estimated degradation rates on SG-NEx data across four cell lines.

by regressing the degradation rate estimates from Bambu against Deg. EM (Fig. 4.10) in a linear model, we find that the linear relationship is significant (p-value = 2.63e-05, adjusted R^2 = 0.890). Even though the sample size here is small, the concordance between the degradation rates estimated by Bambu and Deg. EM (emp.) on real data provides a useful sanity check for our model.

4.6 Discussion

In this chapter, we evaluated our model and inference algorithm (Deg. EM) on simulated datasets with known constant degradation rates and real datasets with sequencing spike-ins from the SG-NEx project. First, we compared the exact and empirical variations of our model, finding that they perform comparably in the range of realistic degradation rates.

Next, we benchmarked Deg. EM against Bambu, Bambu without bias, FLAIR and NanoCount. We showed on simulated datasets that Deg. EM achieves the best performance on SCC, NRMSE and MRD compared to existing methods. In particular, methods that do not model for bias perform poorly on subset isoforms. The results obtained on real datasets with spike-ins were similar, with Deg. EM showing comparable, if not slightly better performance compared to NanoCount and Bambu. We also found that the variance in estimates obtained by Deg. EM was lower compared to those of other methods across replicates, showing that modeling degradation can help increase reproducibility *in silico*.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we examined degradation bias in long-read direct RNA-seq. We first characterised degradation by formalising the notion of the degradation rate, and estimated degradation rates across different human cell lines. We showed that degradation rates are consistent across isoforms and their features. By examining degradation in sequencing spike-ins, we find that degradation in endogenous RNA is likely to be a combination of *in vivo* RNA decay and extraneous *in vitro* factors. Next, we developed a bias-aware model for transcript quantification that models the probability of observing a read originating from an isoform based on a read length-isoform agreement model. We derive an expectation maximization algorithm for inferring degradation-aware isoform abundance estimates. To evaluate our model, we perform benchmarking against existing methods for transcript quantification. On simulated datasets with known degradation, our model outperforms other methods on both reference and subset isoforms. On real datasets with sequencing spike-ins, our model achieves results of comparable accuracy to those of existing methods and admits greater reproducibility.

5.2 Further work

To conclude, we highlight four areas of future work.

5.2.1 Gene-specific degradation

Currently, our bias model fits the degradation rate per base globally, with the assumption that the degradation rate per base is the same across all isoforms. The validity of this assumption depends on the source of the degradation - if the degradation is mostly a product of RNA decay *in vivo*, then this assumption is unlikely to hold, since differential rates of RNA decay have been extensively described. However, if the degradation is mostly an artifact of the library preparation or sequencing, then a global degradation rate may apply across the isoforms. Nevertheless, it might still be of interest to fit *gene-level degradation rates*, as this would likely improve the accuracy of abundance estimates and reflect the underlying biology more faithfully.

5.2.2 Unobserved degradation for short isoforms

While we characterised degradation from the data via the read length distribution of the *observed* degraded reads, it is possible that through the entire process of sequencing, transcripts or reads

from shorter isoforms would have been fully degraded. This would lead to an underestimation of the counts for short isoforms, and possibly warrant some form of length normalisation for the data. Developing experiments to examine degradation over the course of a sequencing experiment for modeling *unobserved degraded reads* is one potential area of future work.

5.2.3 Novel isoform discovery

One attractive benefit of longer reads from ONT sequencing is the potential to discover and characterise novel isoforms. Currently, our model admits transcriptome alignments and does not perform discovery. It is possible to couple our model with a tool capable of transcript discovery, such as Bambu, by first producing extended transcriptome annotations with transcript discovery, and running our model for bias-aware transcript quantification.

5.2.4 Read position-isoform agreement

To model bias, we introduced the read length-isoform agreement, which models the probability of observing a read of a certain length originating from an isoform and its degradation rate. Our current model applies however only to direct RNA-seq, as we assume that the 3' end of all reads align within a close proximity to the annotated 3' end of the isoform. However, it is well known that annotations at both the 5' and 3' end are not clearly well-defined and unreliable. In addition, for direct cDNA and PCR-cDNA protocols, a fraction of reads may not align at the 3' end, resulting in an underestimation of the count. To mitigate these issues and generalise our model to cDNA protocols, we can consider a *read position-isoform agreement* that models the probability of observing a read with a certain relative end position $e_i \in (0, 1)$ within the body of an isoform. Extending our current model (Eq. 3.6), we have:

$$p(r_i \mid z_{ij} = 1) = a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \cdot p(e_i \mid \ell_i, z_{ij} = 1) \tag{5.1}$$

Appendix A

Simulating degraded reads

In this appendix, we describe our approach for simulating reads based on constant expected degradation. We simulate reads for protein-coding isoforms and processed transcripts from GRCh38 annotation. The distribution of read counts follows a negative binomial distribution (see Appendix C). To generate reads for an isoform of length $\mathrm{len}(j)$ given a degradation rate d, we first compute the maximum read length $\ell_{\max} = 1/d$. The probability of generating a degraded read is $p_d = \min(\mathrm{len}(j)/\ell_{\max}, 1)$ and the probability of generating a full-length read is $1 - p_d$. We now generate a read length with the following algorithm:

1. Generate $U \sim \mathcal{U}(0, 1)$.
2. If $U < p_d$, generate degraded read length $\ell_i \sim \mathcal{U}_d(0, \min(\mathrm{len}(j), \ell_{\max}) \cdot 10^3)$.
3. Otherwise, generate full read length $\ell_i = \mathrm{len}(j)$.

Here, $\mathcal{U}$ and $\mathcal{U}_d$ are the continuous and discrete uniform distributions respectively. For constant degradation, the length of degraded reads are uniformly distributed over the integers from 0 to the length of the transcript isoform or the maximum read length, whichever is smaller. Once the read length ℓ_i is generated, we simply slice the transcript isoform sequence from the 3' end such that the resulting sequence is of length ℓ_i. The reads simulated are thus perfect reads with 0% error rate and no indels. We obtain a high fraction of reads (>99%) mapped to the genome and transcriptome over simulated datasets with different degradation rates.

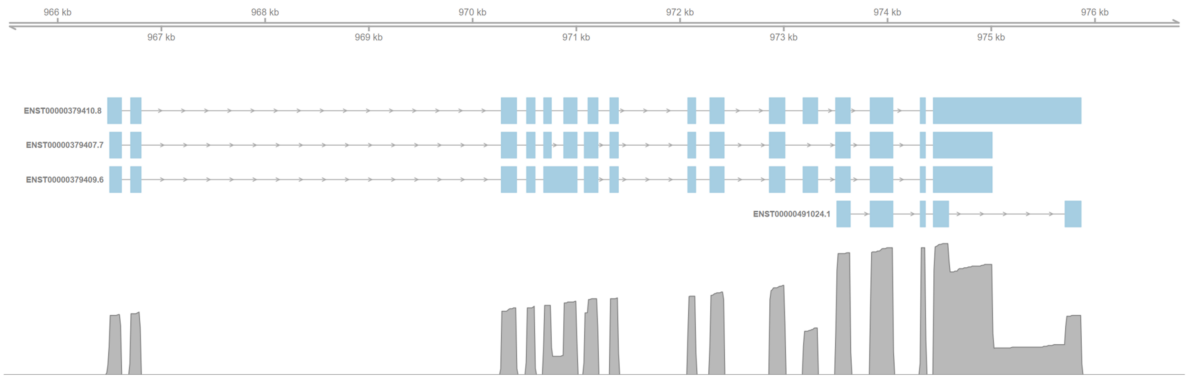

Figure A.1: Simulated reads with constant degradation aligning to the genomic locus of the gene PLEKHN1.

Appendix B

Generating novel isoform models

In this appendix, we describe our approach for generating novel gene isoform models. We first select a set of seed reference isoforms and introduce modifications that are well known to occur via alternative isoform regulation *in vivo*. These modifications include the use of alternative 5' start sites, alternative 3' end sites, alternative splice donors and acceptors, exon skipping, intron retention, and the introduction of new exons and introns. In addition, we also introduce a new modification, termed *subset* isoform, that drops exons from the reference isoform from the 5' end (Fig. B.1). Introduction of these novel subset isoforms increases the number of multi-mapping reads, making the process of assigning these reads to the correct isoform more complex.

Figure B.1: Subset isoform modification. Exons from the 5' end of the reference isoform (blue) are removed to produce a truncated isoform comprising a subset of the exons of the original isoform.

Because some aligners (e.g. `minimap2`) use splice site signals to map reads, we ensure that novel isoform models conform to these splice site signals. We do so by performing splice site correction whenever necessary, i.e., for the alternative splice donor and acceptor, new exon and intron modifications (Fig. B.2).

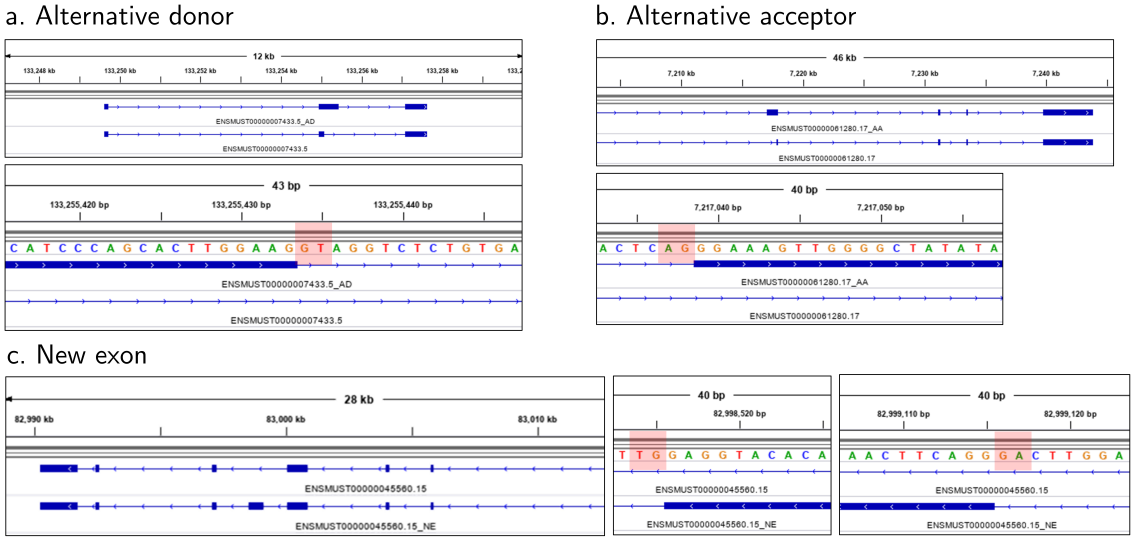

Figure B.2: Splice site correction for novel isoform models for the GT-AG motif highlighted in red. Genomic sequence is reverse complemented where necessary.**a.** Alternative donor. **b.** Alternative acceptor. **c.** New exon.

Appendix C

Count distribution analysis

To identify an appropriate distribution for simulating counts that best reflects real data, we ran NanoCount and Bambu on six direct RNA-seq samples obtained from sequencing of a human embryonic stem cell line (H9). Using the fitdistrplus R package (v1.1.6) [63], we fitted discrete distributions (negative binomial and Poisson) to the counts returned by NanoCount and Bambu after filtering the counts in the top 0.001 quantile and analysed the skew and kurtosis of the fitted distribution. Since skewness and kurtosis are known not to be robust [63], we also fit distributions to 100 bootstrap replicates. In all samples, the negative binomial provided a better fit compared to the Poisson distribution (Fig. C.1). This is unsurprising, as the negative binomial allows more flexible modeling of counts with an additional parameter that allows for greater variation, while the equal mean and variance assumption of the Poisson is somewhat restrictive. In fact, the negative binomial is equivalent to a mixture of Poisson distributions with the rate parameter distributed according to a gamma distribution.

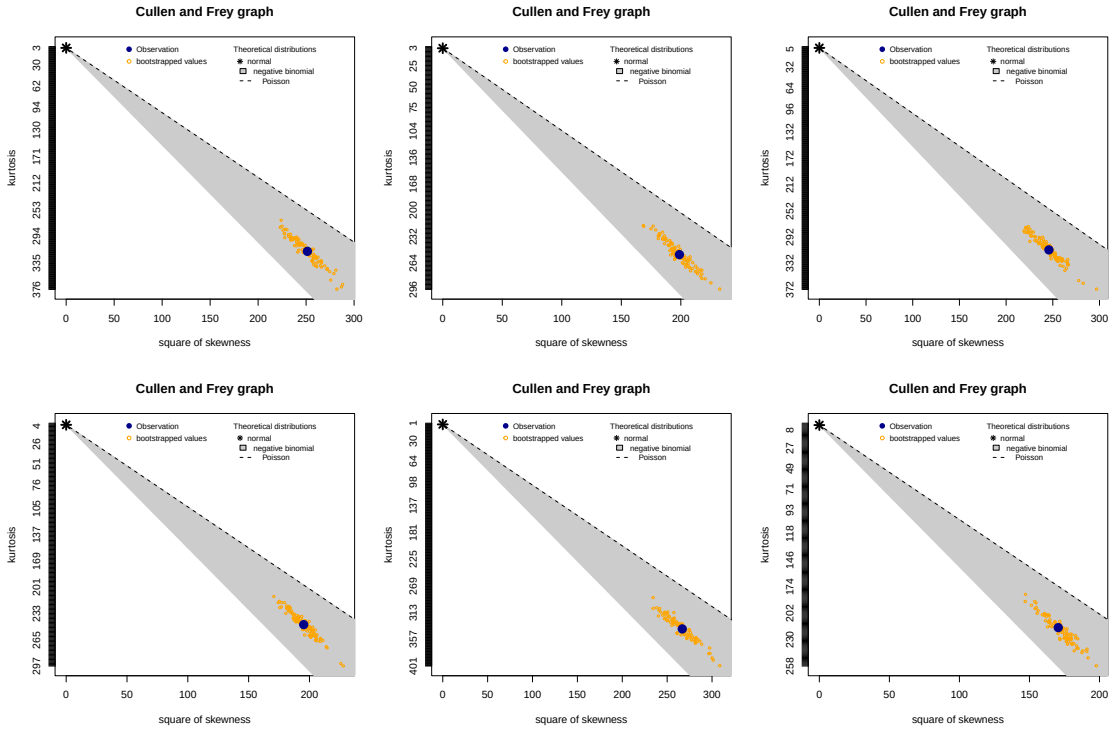

Figure C.1: Distributional analysis of RNA-seq counts returned by NanoCount (results for Bambu are similar). Each panel corresponds to one H9 sample, and each dot is a set of counts. Dark blue dots are real samples while bright orange dots are bootstrapped samples. The grey region indicates combinations of skew and kurtosis parameters for the negative binomial, while the dotted line indicates the same for the Poisson distribution. The real and bootstrap samples fall within the grey region and away from the dotted line implying that the negative binomial offers a better fit compared to the Poisson.

Appendix D

Proof of concavity of log-likelihood function

Here, we show that the log-likelihood function is concave. This implies that the local maximum corresponding to the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ returned by the EM algorithm is a global maximum. We follow the proof used in [64] and [65].

Recall that the log-likelihood formulated in Section 3.3.1 is given by

$$\log p(\boldsymbol{R} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \right] \tag{D.1}$$

Since the sum of concave functions is also concave, we need only prove that each term in this sum is concave. Let the i^{th} term be

$$f(\boldsymbol{\theta}) = \log \left[\sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p(\ell_i \mid d_j, z_{ij} = 1) \right] \tag{D.2}$$

Let $\mathbf{H}_f$ be the Hessian matrix for $f(\boldsymbol{\theta})$. The (x, y) entry of $\mathbf{H}_f$ is

$$\begin{aligned}
(\mathbf{H}_f)_{x,y} &= \frac{\partial^2 f}{\partial \theta_x \partial \theta_y} \\
&= \frac{\partial}{\partial \theta_x} \frac{\partial}{\partial \theta_y} \log \left[\sum_{j=1}^{M} \theta_j \cdot a_{ij} \cdot p\left(\ell_i \mid d_j, z_{ij} = 1\right) \right] \\
&= \frac{\partial}{\partial \theta_x} \frac{a_{iy} \cdot p\left(\ell_i \mid d_y, z_{iy} = 1\right)}{\sum_j \theta_j \cdot a_{ij} \cdot p\left(\ell_i \mid d_j, z_{ij} = 1\right)} \\
&= -\frac{a_{iy} \cdot p\left(\ell_i \mid d_y, z_{iy} = 1\right) \cdot a_{ix} \cdot p\left(\ell_i \mid d_x, z_{ix} = 1\right)}{\left(\sum_j \theta_j \cdot a_{ij} \cdot p\left(\ell_i \mid d_j, z_{ij} = 1\right)\right)^2}
\end{aligned} \tag{D.3}$$

$\mathbf{H}_f$ can thus be expressed as

$$\mathbf{H}_f = -g(\boldsymbol{\theta}) u' u \tag{D.4}$$

where

$$g(\boldsymbol{\theta}) = \frac{1}{\left(\sum_j \theta_j \cdot a_{ij} \cdot p\left(\ell_i \mid d_j, z_{ij} = 1\right)\right)^2} \tag{D.5}$$

and

$$u = \left[a_{i1} \cdot p\left(\ell_i \mid d_1, z_{i1} = 1\right) \ldots a_{iM} \cdot p\left(\ell_i \mid d_M, z_{iM} = 1\right) \right] \tag{D.6}$$

Thus, for all v, we have

$$\begin{aligned}
v \mathbf{H} v' &= v \left(-g(\boldsymbol{\theta}) u' u\right) v' \\
&= -g(\boldsymbol{\theta}) \left(v u'\right) \left(v u'\right)' \\
&= -g(\boldsymbol{\theta}) \left(v u'\right)^2 \leq 0
\end{aligned} \tag{D.7}$$

The inequality holds since $g(\boldsymbol{\theta}) > 0$. Therefore, $\mathbf{H}_f$ is negative semi-definite, and thus $f(\boldsymbol{\theta})$ and $\log p(\boldsymbol{R} \mid \boldsymbol{\theta})$ are concave.

We show a plot of the log-likelihood against EM iterations:

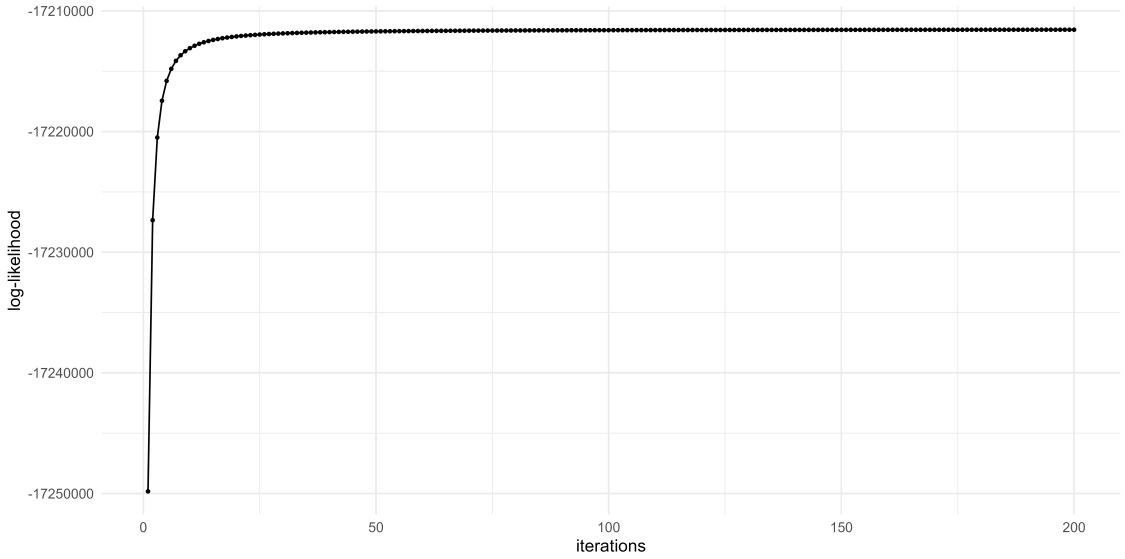

Figure D.1: Log-likelihood against EM iterations

Our EM algorithm converges quickly, with the log-likelihood non-decreasing.

Appendix E

Evaluation metrics

In this appendix, we describe evaluation metrics used for model evaluation. Let $\boldsymbol{\theta}$ be the reference and $\hat{\boldsymbol{\theta}}$ be the estimates.

The **Spearman correlation coefficient** (SCC) is a reference-based metric that measures the strength of monotonic relationships between the reference and estimate. Let $\boldsymbol{R}$ be the ranks for $\boldsymbol{\theta}$ and $\hat{\boldsymbol{R}}$ be the ranks for $\hat{\boldsymbol{\theta}}$. The SCC is given by

$$SCC = \frac{\mathrm{cov}(\boldsymbol{R}, \hat{\boldsymbol{R}})}{\mathrm{sd}(\boldsymbol{R}) \cdot \mathrm{sd}(\hat{\boldsymbol{R}})} \tag{E.1}$$

The **relative difference** (RD) is a reference-based metric that measures relative difference of an estimate against the reference. The RD is given by

$$RD = \frac{\mid \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \mid}{\boldsymbol{\theta}} \tag{E.2}$$

The **normalized root-mean-squared error** (NRMSE) is a reference-based metric that measures the deviation from a linear relationship between the estimate and the reference. The NRMSE is given by

$$NRMSE = \frac{\sqrt{\frac{1}{M}\sum\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^2}}{\mathrm{sd}\left(\hat{\boldsymbol{\theta}}\right)} \tag{E.3}$$

The **reproducibility metric** (RM) is a reference-free metric that measures the variance across replicates. Let $\bar{\boldsymbol{\theta}}$ be the mean of estimates across K replicates. For isoform i, the RM is given by

$$RM = \sqrt{\frac{1}{K}\sum_{k}\left(\hat{\theta}_i - \bar{\theta}_i\right)^2} \tag{E.4}$$

Appendix F

Data and code availability

Data We used long-read RNA-seq data from the SG-NEx project available at `https://github.com/GoekeLab/sg-nex-data`. All simulated datasets were generated with the software below.

Code All software was developed in `python3`. Software for generating novel isoform models and simulating degraded reads are available at `https://github.com/jleechung/noviso` and `https://github.com/jleechung/shamread` respectively. These are not implemented as CLI tools but use a yaml file for configuring inputs. Software for our quantification model is available at `https://github.com/jleechung/daiso` and is implemented as a CLI tool with simple installation and user interface, with only one required input. We display the manual page here for reference:

manual

```
usage: daiso -a alignment.bam [options]

Degradation-aware isoform quantification

Required arguments:
  -a ALIGNMENT, --alignment ALIGNMENT
                        BAM file (sorted and indexed) containing reads aligned
                        to the transcriptome

Optional arguments:
  -o OUTPUT, --output OUTPUT
                        Output prefix. The count and degradation plot files
                        will be written out to <output>_counts.tsv and
                        <output>_survival.png respectively (default: out)
  --no_quant            Estimate degradation only. No quantification is run.
                        (default: False)
  --seed SEED           Set seed (default: 42)

Parameters for alignment filtering:
  --filter_distance FILTER_DISTANCE
                        Filter alignments where end position is further than
                        this distance away from the annotated 3 prime end
                        (default: 50 bp)
  --filter_score FILTER_SCORE
                        Filter alignments where the alignment score is lower
                        than this fraction of the best alignment score for
```

```
                        this read (default: 0.95)

Parameters for degradation estimation:
  --deg_rate DEG_RATE   Degradation rate in (0,1). If this argument is
                        supplied, degradation rate will not be estimated from
                        the data and deg_const will be True
  --deg_const           Uses exact constant degradation model for read length-
                        isoform agreement (default:False)
  --bin_size BIN_SIZE   Bin isoform lengths into bins of this size (default:
                        500 bp)
  --min_read_count MIN_READ_COUNT
                        Only estimate degradation with isoforms with at least
                        this many reads (default: 5)
  --min_iso_count MIN_ISO_COUNT
                        Only estimate degradation with length bins supported
                        by this many isoforms (default: 1)
  --full_len_tol FULL_LEN_TOL
                        Consider a read full length if it is within this many
                        bp of the annotated isoform length (default: 50 bp)
  --delta DELTA         Left and right shift for estimating exact
                        probabilities for ecdf (default: 50 bp)
  --return_survival     Return the survival function (1-ecdf). Writes values
                        to <output>_survival.tsv (default: True)

Parameters for inference:
  --inference {EM,VB}   Inference algorithm - expectation maximization or
                        variational Bayesian inference (default: EM)
  --max_iter MAX_ITER   Maximum number of iterations (default: 200)
  --return_loglik       Return the log likelihood over iterations for EM.
                        Writes values to <output>_loglik.tsv (default: True)
  --prior {symmetric,gamma_hyper}
                        Choice of prior - symmetric prior or a gamma
                        hyperprior (default: gamma_hyper)
  --alpha_zero ALPHA_ZERO
                        Value of the concentration parameter for the symmetric
                        prior (default: 1)
  --gamma_rate GAMMA_RATE
                        Rate parameter for the gamma hyperprior (default: 5)
  --gamma_scale GAMMA_SCALE
                        Scale parameter for the gamma hyperprior (default: 5)
  --return_cred         Return credible interals for parameters (default:
                        False)
  --cred_int CRED_INT   Width of credible interval (default: 0.95)

General help:
  -h, --help            Show this help message and exit
```

```
  -v, --version         Print version
```

Bibliography

[1] Martin O Pollard et al. "Long reads: their purpose and place". In: *Human Molecular Genetics* 27.R2 (2018), R234–R241. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy177. eprint: https://academic.oup.com/hmg/article-pdf/27/R2/R234/25229925/ddy177.pdf. URL: https://doi.org/10.1093/hmg/ddy177.

[2] Mohan T. Bolisetty, Gopinath Rajadinakaran, and Brenton R. Graveley. "Determining exon connectivity in complex mRNAs by nanopore sequencing". In: *Genome Biology* 16.1 (2015), p. 204. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0777-z. URL: https://doi.org/10.1186/s13059-015-0777-z.

[3] Ashley Byrne et al. "Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells". In: *Nature Communications* 8.1 (2017), p. 16027. ISSN: 2041-1723. DOI: 10.1038/ncomms16027. URL: https://doi.org/10.1038/ncomms16027.

[4] Wouter De Coster et al. "Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome". eng. In: *Genome research* 29.7 (2019). gr.244939.118[PII], pp. 1178–1187. ISSN: 1549-5469. DOI: 10.1101/gr.244939.118. URL: https://doi.org/10.1101/gr.244939.118.

[5] Huanle Liu et al. "Accurate detection of m6A RNA modifications in native RNA sequences". In: *Nature Communications* 10.1 (2019), p. 4079. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11713-9. URL: https://doi.org/10.1038/s41467-019-11713-9.

[6] Tuomo Mantere, Simone Kersten, and Alexander Hoischen. "Long-Read Sequencing Emerging in Medical Genetics". In: *Frontiers in Genetics* 10 (2019), p. 426. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00426. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00426.

[7] Sergey Nurk et al. "The complete sequence of a human genome". In: *bioRxiv* (2021). DOI: 10.1101/2021.05.26.445798. eprint: https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.full.pdf. URL: https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.

[8] Rohan Lowe et al. "Transcriptomics technologies". eng. In: *PLoS computational biology* 13.5 (May 2017). PCOMPBIOL-D-17-00431[PII], e1005457–e1005457. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005457. URL: https://doi.org/10.1371/journal.pcbi.1005457.

[9] Samuel Marguerat and Jürg Bähler. "RNA-seq: from technology to biology". In: *Cellular and Molecular Life Sciences* 67.4 (Feb. 2010), pp. 569–579. ISSN: 1420-9071. DOI: 10.1007/s00018-009-0180-6. URL: https://doi.org/10.1007/s00018-009-0180-6.

[10] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. "Biases in Illumina transcriptome sequencing caused by random hexamer priming". In: *Nucleic Acids Research* 38.12 (Apr. 2010), e131–e131. ISSN: 0305-1048. DOI: 10.1093/nar/gkq224. eprint: https://academic.oup.com/nar/article-pdf/38/12/e131/14120852/gkq224.pdf. URL: https://doi.org/10.1093/nar/gkq224.

[11] Jun Li, Hui Jiang, and Wing Hung Wong. "Modeling non-uniformity in short-read rates in RNA-Seq data". In: *Genome Biology* 11.5 (May 2010), R50. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-5-r50. URL: https://doi.org/10.1186/gb-2010-11-5-r50.

[12] Bo Li and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1 (Aug. 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323.

[13] Zhengpeng Wu, Xi Wang, and Xuegong Zhang. "Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq". In: *Bioinformatics* 27.4 (Dec. 2010), pp. 502–508. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq696. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/4/502/16903047/btq696.pdf. URL: https://doi.org/10.1093/bioinformatics/btq696.

[14] Adam Roberts et al. "Improving RNA-Seq expression estimates by correcting for fragment bias". In: *Genome Biology* 12.3 (2011), R22. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-3-r22. URL: https://doi.org/10.1186/gb-2011-12-3-r22.

[15] Yuval Benjamini and Terence P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing". In: *Nucleic Acids Research* 40.10 (Feb. 2012), e72–e72. ISSN: 0305-1048. DOI: 10.1093/nar/gks001. eprint: https://academic.oup.com/nar/article-pdf/40/10/e72/25335311/gks001.pdf. URL: https://doi.org/10.1093/nar/gks001.

[16] Nicholas F. Lahens et al. "IVT-seq reveals extreme bias in RNA sequencing". In: *Genome Biology* 15.6 (June 2014), R86. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-6-r86. URL: https://doi.org/10.1186/gb-2014-15-6-r86.

[17] Michael I. Love, John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation". In: *Nature Biotechnology* 34.12 (2016), pp. 1287–1291. ISSN: 1546-1696. DOI: 10.1038/nbt.3682. URL: https://doi.org/10.1038/nbt.3682.

[18] Erwin L. van Dijk, Yan Jaszczyszyn, and Claude Thermes. "Library preparation methods for next-generation sequencing: Tone down the bias". In: *Experimental Cell Research* 322.1 (2014), pp. 12–20. ISSN: 0014-4827. DOI: https://doi.org/10.1016/j.yexcr.

2014.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0014482714000160.

[19] Lira Mamanova et al. "FRT-seq: amplification-free, strand-specific transcriptome sequencing". In: *Nature Methods* 7.2 (Feb. 2010), pp. 130–132. ISSN: 1548-7105. DOI: 10.1038/nmeth.1417. URL: https://doi.org/10.1038/nmeth.1417.

[20] Michael A. Quail et al. "Optimal enzymes for amplifying sequencing libraries". In: *Nature Methods* 9.1 (Jan. 2012), pp. 10–11. ISSN: 1548-7105. DOI: 10.1038/nmeth.1814. URL: https://doi.org/10.1038/nmeth.1814.

[21] Andreas Tuerk, Gregor Wiktorin, and Serhat Güler. "Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates". In: *PLOS Computational Biology* 13.5 (May 2017), pp. 1–25. DOI: 10.1371/journal.pcbi.1005515. URL: https://doi.org/10.1371/journal.pcbi.1005515.

[22] Zhengpeng Wu, Xi Wang, and Xuegong Zhang. "Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq". In: *Bioinformatics* 27.4 (Dec. 2010), pp. 502–508. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq696. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/4/502/16903047/btq696.pdf. URL: https://doi.org/10.1093/bioinformatics/btq696.

[23] JL Weirather et al. "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; peer review: 2 approved]". In: *F1000Research* 6 (2017). DOI: 10.12688/f1000research.10571.2.

[24] Mauricio O. Carneiro et al. "Pacific biosciences sequencing technology for genotyping and variation discovery in human data". In: *BMC Genomics* 13.1 (Aug. 2012), p. 375. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-375. URL: https://doi.org/10.1186/1471-2164-13-375.

[25] Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. "High-Throughput Sequencing Technologies". In: *Molecular Cell* 58.4 (May 2015), pp. 586–597. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.05.004. URL: https://doi.org/10.1016/j.molcel.2015.05.004.

[26] Miten Jain et al. "Improved data analysis for the MinION nanopore sequencer". In: *Nature Methods* 12.4 (Apr. 2015), pp. 351–356. ISSN: 1548-7105. DOI: 10.1038/nmeth.3290. URL: https://doi.org/10.1038/nmeth.3290.

[27] Miten Jain et al. "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biology* 17.1 (Dec. 2016), p. 239. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1103-0. URL: https://doi.org/10.1186/s13059-016-1103-0.

[28] Yunhao Wang et al. "Nanopore sequencing technology, bioinformatics and applications". In: *Nature Biotechnology* 39.11 (Nov. 2021), pp. 1348–1365. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01108-x. URL: https://doi.org/10.1038/s41587-021-01108-x.

[29] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. "From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy". In: *Genome Biology* 19.1 (2018), p. 90. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1462-9. URL: https://doi.org/10.1186/s13059-018-1462-9.

[30] Daniel R. Garalde et al. "Highly parallel direct RNA sequencing on an array of nanopores". In: *Nature Methods* 15.3 (Mar. 2018), pp. 201–206. ISSN: 1548-7105. DOI: 10.1038/nmeth.4577. URL: https://doi.org/10.1038/nmeth.4577.

[31] *Quantitative RNA-seq: PCR-cdna, PCR-free direct cdna and direct RNA sequencing*. June 2020. URL: https://nanoporetech.com/resource-centre/quantitative-rna-seq-pcr-cdna-pcr-free-direct-cdna-and-direct-rna-sequencing.

[32] Daniel P. Depledge et al. "Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen". In: *Nature Communications* 10.1 (Feb. 2019), p. 754. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08734-9. URL: https://doi.org/10.1038/s41467-019-08734-9.

[33] Charlotte Soneson et al. "A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes". In: *Nature Communications* 10.1 (2019), p. 3359. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11272-z. URL: https://doi.org/10.1038/s41467-019-11272-z.

[34] *Quantitative RNA-seq: PCR-cdna, PCR-free direct cdna and direct RNA sequencing*. June 2020. URL: https://nanoporetech.com/resource-centre/quantitative-rna-seq-pcr-cdna-pcr-free-direct-cdna-and-direct-rna-sequencing.

[35] *Transcriptional landscapes analysis through direct RNA sequencing*. Dec. 2017. URL: https://nanoporetech.com/resource-centre/transcriptional-landscapes-analysis-through-direct-rna-sequencing.

[36] Patrick Denis Browne et al. "GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms". In: *GigaScience* 9.2 (Feb. 2020). giaa008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa008. eprint: https://academic.oup.com/gigascience/article-pdf/9/2/giaa008/32449353/giaa008.pdf. URL: https://doi.org/10.1093/gigascience/giaa008.

[37] Rachael E. Workman et al. "Nanopore native RNA sequencing of a human poly(A) transcriptome". In: *Nature Methods* 16.12 (Dec. 2019), pp. 1297–1305. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0617-2. URL: https://doi.org/10.1038/s41592-019-0617-2.

[38] Elena Conti and Elisa Izaurralde. "Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species". In: *Current Opinion in Cell Biology* 17.3 (2005). Nucleus and gene expression, pp. 316–325. ISSN: 0955-0674. DOI: https://doi.org/10.1016/j.ceb.2005.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0955067405000475.

[39] Nicole L. Garneau, Jeffrey Wilusz, and Carol J. Wilusz. "The highways and byways of mRNA decay". In: *Nature Reviews Molecular Cell Biology* 8.2 (Feb. 2007), pp. 113–126. ISSN: 1471-0080. DOI: 10.1038/nrm2104. URL: https://doi.org/10.1038/nrm2104.

[40] Jonathan Houseley and David Tollervey. "The Many Pathways of RNA Degradation". In: *Cell* 136.4 (2009), pp. 763–776. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2009.01.019. URL: https://www.sciencedirect.com/science/article/pii/S0092867409000671.

[41] Jack D. Keene. "Minireview: Global Regulation and Dynamics of Ribonucleic Acid". In: *Endocrinology* 151.4 (Apr. 2010), pp. 1391–1397. ISSN: 0013-7227. DOI: 10.1210/en.2009-1250. eprint: https://academic.oup.com/endo/article-pdf/151/4/1391/9007493/endo1391.pdf. URL: https://doi.org/10.1210/en.2009-1250.

[42] David Gatfield and Elisa Izaurralde. "Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila". In: *Nature* 429.6991 (June 2004), pp. 575–578. ISSN: 1476-4687. DOI: 10.1038/nature02559. URL: https://doi.org/10.1038/nature02559.

[43] E. Yang et al. "Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes". In: *Genome Res* 13.8 (Aug. 2003), pp. 1863–1872.

[44] Kaiwen Shi et al. "Genome-wide analysis of lncRNA stability in human". In: *PLOS Computational Biology* 17.4 (Apr. 2021), pp. 1–25. DOI: 10.1371/journal.pcbi.1008918. URL: https://doi.org/10.1371/journal.pcbi.1008918.

[45] Shanika L. Amarasinghe et al. "Opportunities and challenges in long-read sequencing data analysis". In: *Genome Biology* 21.1 (2020), p. 30. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1935-5. URL: https://doi.org/10.1186/s13059-020-1935-5.

[46] Alison D. Tang et al. "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns". In: *Nature Communications* 11.1 (Mar. 2020), p. 1438. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15171-6. URL: https://doi.org/10.1038/s41467-020-15171-6.

[47] Josie Gleeson et al. "Accurate expression quantification from nanopore direct RNA sequencing with NanoCount". In: *Nucleic Acids Research* 50.4 (Nov. 2021), e19–e19. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1129. eprint: https://academic.oup.com/nar/

article-pdf/50/4/e19/42617894/gkab1129.pdf. URL: https://doi.org/10.1093/nar/gkab1129.

[48] Ying Chen et al. *bambu*. R package version 2.0.6. 2022. URL: https://github.com/GoekeLab/bambu.

[49] Ying Chen et al. "A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines". In: *bioRxiv* (2021). DOI: 10.1101/2021.04.21.440736. eprint: https://www.biorxiv.org/content/early/2021/04/22/2021.04.21.440736.full.pdf. URL: https://www.biorxiv.org/content/early/2021/04/22/2021.04.21.440736.

[50] Michael B. Clark et al. "Genome-wide analysis of long noncoding RNA stability". eng. In: *Genome research* 22.5 (May 2012). gr.131037.111[PII], pp. 885–898. ISSN: 1549-5469. DOI: 10.1101/gr.131037.111. URL: https://doi.org/10.1101/gr.131037.111.

[51] *SIRVs (spike-in RNA variant control mixes)*. URL: https://www.lexogen.com/sirvs/.

[52] Z. Wang and M. Kiledjian. "The poly(A)-binding protein and an mRNA stability protein jointly regulate an endoribonuclease activity". eng. In: *Molecular and cellular biology* 20.17 (Sept. 2000). PMC86108[pmcid], pp. 6334–6341. ISSN: 0270-7306. DOI: 10.1128/MCB.20.17.6334-6341.2000. URL: https://doi.org/10.1128/MCB.20.17.6334-6341.2000.

[53] *Sequence Alignment/Map Format Specification*. URL: https://samtools.github.io/hts-specs/SAMv1.pdf.

[54] Nicolas L. Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (May 2016), pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: https://doi.org/10.1038/nbt.3519.

[55] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4 (Apr. 2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197. URL: https://doi.org/10.1038/nmeth.4197.

[56] Yu Hu et al. "LIQA: long-read isoform quantification and analysis". In: *Genome Biology* 22.1 (June 2021), p. 182. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02399-8. URL: https://doi.org/10.1186/s13059-021-02399-8.

[57] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (May 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3094/25731859/bty191.pdf. URL: https://doi.org/10.1093/bioinformatics/bty191.

[58] Heng Li. "New strategies to improve minimap2 alignment accuracy". In: *Bioinformatics* 37.23 (Oct. 2021), pp. 4572–4574. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab705. eprint: https://academic.oup.com/bioinformatics/article-pdf/37/23/4572/41641731/btab705.pdf. URL: https://doi.org/10.1093/bioinformatics/btab705.

[59] A. Colin Cameron and P. K. Trivedi. *Regression analysis of Count Data*. Cambridge University Press, 2013.

[60] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (Oct. 2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106. URL: https://doi.org/10.1186/gb-2010-11-10-r106.

[61] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". eng. In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 2010). btp616[PII], pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: https://doi.org/10.1093/bioinformatics/btp616.

[62] *Sequins, synthetic DNA and RNA standards for next-generation sequencing*. URL: https://www.sequinstandards.com/.

[63] Marie Laure Delignette-Muller and Christophe Dutang. "fitdistrplus: An R Package for Fitting Distributions". In: *Journal of Statistical Software* 64.4 (2015), pp. 1–34. URL: https://www.jstatsoft.org/article/view/v064i04.

[64] Hui Jiang and Wing Hung Wong. "Statistical inferences for isoform expression in RNA-Seq". In: *Bioinformatics* 25.8 (Feb. 2009), pp. 1026–1032. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp113. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/8/1026/16892810/btp113.pdf. URL: https://doi.org/10.1093/bioinformatics/btp113.

[65] Bo Li et al. "RNA-Seq gene expression estimation with read mapping uncertainty". In: *Bioinformatics* 26.4 (Dec. 2009), pp. 493–500. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp692. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/4/493/16897170/btp692.pdf. URL: https://doi.org/10.1093/bioinformatics/btp692.