Multi-sample, multi-platform isoform quantification using empirical Bayes

Arghamitra Talukder[0009−0001−3469−0295], Shree Thavarekere[0009−0000−3375−9855], and David A. Knowles

Department of Computer Science, Columbia University, New York, NY
{at3836}@columbia.edu

Abstract. Accurate quantification of RNA isoform abundance is crucial for understanding gene regulation, cellular behavior, and disease mechanisms. While short-read (SR) sequencing provides high-throughput and cost-effective transcript quantification, it suffers from read-to-transcript ambiguity. Long-read (LR) sequencing reduces this ambiguity but faces challenges such as high error rates, biases, and lower throughput. Existing methods rely on either SR or LR data and operate on single or merged samples, failing to leverage the variability across multiple samples and the complementary strengths of both technologies. As a result, they struggle to accurately quantify low-abundance and moderate-expressed isoforms and often require complex models for sample-specific bias correction. To address these limitations, we introduce **JOLI**, a hierarchical model that leverages multi-sample learning to enhance transcript quantification by jointly integrating SR and LR sequencing data. By incorporating multi-sample learning, JOLI captures shared transcript structures, corrects for systematic biases, and enhances statistical power, particularly for low- and moderate-abundance isoforms. Our model applies an empirical Bayes framework, learning a shared prior across samples to improve inference consistency. By jointly modeling SR and LR data, it integrates the strengths of both technologies, achieving higher accuracy and reproducibility in transcript quantification. Through benchmarking on simulated and real RNA-seq datasets, we show that JOLI consistently outperforms single-sample EM method by improving ranking consistency, proportional agreement, and estimation accuracy while enhancing reproducibility. Specifically, in simulations, JOLI multi-sample improves Spearman correlation by 9.8% for LR and 7.7% for SR data compared to single-sample method, while for real data, the improvements are 2.56% (LR) and 1.28% (SR), respectively. Multi-sample learning further improves the quantification of isoforms with low to moderate expression levels. Furthermore, JOLI performs competitively with state-of-the-art methods, highlighting its robustness in transcript quantification.

Keywords: Isoform quantification · Hierarchical modeling · Empirical Bayes

1 Introduction

The transcription of DNA into RNA is an intricate process, where individual genes can produce multiple RNA isoforms through mechanisms such as alternative splicing. Accurately quantifying transcript abundance, especially at the isoform level, is fundamental in genomics, shedding light on cellular behavior [29], developmental dynamics [16], and disease mechanisms [10].

Over the past decade, RNA-seq has transformed transcriptomics by enabling high-resolution analysis of gene expression and insights into isoform-level variation. Short-read (SR) sequencing technologies [3], which are cost-effective and extremely high-throughput, enable reproducible quantification. However, its short read length—typically 100-150 base pairs (bp)—is much smaller than the median length of spliced mRNA transcripts (around 3,300bp in humans) [24]. This discrepancy leads to a fundamental issue of read-to-transcript ambiguity, where the origin of many reads cannot be precisely determined. This ambiguity contributes to uncertainty in transcript abundance estimates, which complicates downstream analyses. In contrast, emerging long-read (LR) RNA sequencing technologies, such as those from Pacific Biosciences (PacBio) [6,25] and Oxford Nanopore Technologies (ONT) [20,31], offer reads spanning thousands of base pairs, potentially even covering the entire full-length transcript. This reduces or in the latter case eliminates ambiguity in transcript assignment and facilitates more accurate transcript discovery and quantification. PacBio long reads, especially HiFi reads, are highly regarded for their exceptional accuracy and minimal error rates [11]. Nevertheless, this advantage comes at the expense of reduced throughput and higher costs when compared to ONT sequencing. Conversely, ONT sequencing provides much greater throughput, often making it a more cost-efficient option, but the reads generated typically exhibit lower accuracy and increased error rates ranging from 5 to 20% [12,13]. Indeed, the recent LR-GASP benchmarking effort showed that for isoform quantification (as opposed to discovery), SR analysis currently outperforms LR [21]. This motivates the question: can we get the best of both worlds, combining the throughput and accuracy of SR data with the reduced ambiguity of LR?

Many methods have been developed for isoform quantification for SR platforms—-BitSeqVB [9], RSEM [14], Kallisto [5], Salmon [22]—and LR platforms—LIQA [12], NanoCount [8], Oarfish [13], LR-kallisto [19]. These methods are restricted to a single platform and work with either individual or combined samples, preventing them from capturing variability across multiple samples. Consequently, they are unable to integrate both the distinct advantages of SR and LR technologies for transcript quantification. To address this, we introduce **JOLI**, a hierarchical model for *JOint Long and short Isoform quantification*. JOLI builds upon the standard EM framework by incorporating two key extensions: 1) joint analysis of SR and LR data, and 2) sharing information between multiple similar samples ("multisample" analysis). Our proposed method simultaneously leverages read information from multiple short and long-read samples, exploiting their complementarity to improve accuracy, consistency and reproducibility. It operates on multiple related samples (e.g., from the same tissue type), under the

assumption that the true isoform abundances for these samples will be correlated. Applying an empirical Bayes framework, we learn a common prior across the samples that allows the model to capture shared variability and facilitates robust and consistent inference. To estimate the Maximum A Posteriori (MAP) parameters, we employ the EM algorithm within each sample followed by gradient descent for the shared prior parameters.

Multi-sample variability. Our hierarchical model accounts for sample variability, enabling more accurate estimation of isoform abundances while addressing biases that are not generalizable across different sequencing technologies [18]. SR and LR sequencing technologies are subject to different biases and artifacts, such as PCR amplification and reverse transcription bias [23,1,26]. LR sequencing faces challenges such as 3' bias, caused by technical limitations like pore blockages, incomplete cDNA synthesis, or RNA degradation. As a consequence, LR can exhibits length biases, with short transcripts being disproportionately represented [2]. In comparison, SR sequencing offers more uniform coverage due to the relatively random distribution of reads across fragmented RNA, making it well-suited to counteract these biases when the two technologies are used together [26]. Methods such as LIQA [12] and Oarfish [13] rely on complicated models for bias and error correction. By leveraging both SR and LR, our model is able to provide a simpler and more robust solution.

Enhancing Noise-to-Signal Ratio for Low and moderate Isoform Abundance: By pooling information across samples, our method leverages shared transcript structures, increasing statistical power to detect low and moderate abundance isoforms and distinguish genuine signals from noise, even in the presence of technical biases [28,17]. A common strategy for leveraging multiple samples perform isoform assembly and identification using a merged set of all reads, then quantification of these isoforms independently in each sample [27]. This approach however does not share valuable quantitative information between samples. In contrast, our hierarchical model addresses achieves this by treating observations as noisy realizations of a shared prior to balance variability and consistency across related samples.

2 Methods

We introduce the key notations used in this work in Table 1 and present the generative model underlying our framework in Fig. 1. This section details the generative model, JOLI model fitting, and dataset.

2.1 JOLI Generative Model

Our generative model describes RNA-seq reads originating from multiple samples and builds upon the single-sample case by incorporating sample-specific parameters (Fig. 1). Each sample f has its own transcript proportions θ_f, reflecting differences across biological conditions or individuals. Each θ_f is drawn from a Dirichlet distribution, i.e. $\theta_f|\alpha \sim \text{Dirichlet}(\alpha)$, where α is a hyperparameter

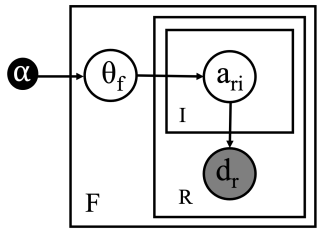

Fig. 1: Generative model for multi-sample reads.

Table 1: Summary of notations

Symbol	Description
R	Number of reads
I	Number of isoforms/transcripts
F	Number of samples
d_r	The r-th read
t_i	The i-th transcript
α	Dirichlet parameter of transcript abundance
θ_{fi}	Abundance of transcript t_i in the sample f
a_{ri}	True binary indicator if the d_r is from t_i.
y_{ri}	Assignment matrix $y_{ri} = 1$ if the d_r is consistent with t_i, or $y_{ri} = 0$ otherwise
z_{ri}	Probability matrix, that the d_r is generated by t_i
n_i	Expected number of reads generated by t_i.

vector. The Dirichlet serves as a shared prior over the transcript proportions, encoding shared information across all samples while allowing sample-specific variations. Defining $\alpha_\circ = \sum_i \alpha_i$ the prior mean is $\alpha/\alpha_\circ$ and larger values of $\alpha_\circ$ lead to more uniform transcript proportions across samples, while smaller values allow for greater variability. For each read d_{fr} in sample f, the generative process is:

1. A transcript t_{fr} is sampled from a categorical distribution parameterized by θ_f, i.e., $t_{fr}|\theta_f \sim \text{Categorical}(\theta_f)$. This step models that the prior probability that a read comes from a specific isoform is proportion to its relative abundance. For convenience, we define the binary variable a_{fri}, where $a_{fri} = 1$ if read d_{fr} originates from transcript i (i.e., if $t_{fr} = 1$) and $a_{fri} = 0$ otherwise. The contribution of this step to the joint probability is, $\prod_r \prod_i \theta_i^{a_{fri}}$.
2. The observed read d_{fr} is generated conditionally on the selected transcript t_{fr} according to,

$$P(d_{fr}|t_{fr} = i) = p_{ri} = \begin{cases} \frac{1}{L_i - l_r + 1}, & \text{if } d_{fr} \text{ maps to } t_{fi}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where l_r is the read length and L_i is the transcript length.

The joint probability of the model is,

$$P(\theta, t, d|\alpha) = \prod_f P(\theta_f|\alpha) \prod_r P(t_{fr}|\theta_f) P(d_{fr}|t_{fr}), \tag{2}$$

where,

$$P(\theta_f|\alpha) \propto \prod_{i=1}^{I} \theta_{fi}^{\alpha_i - 1}, \quad P(t_{fr}|\theta_f) = \prod_{i=1}^{I} \theta_{fi}^{a_{fri}}.$$

2.2 JOLI Model Fitting

We use an expectation-maximization (EM) algorithm to estimate transcript proportions θ_f for each sample f, given observed RNA-seq reads r. Conditional on the prior parameter α, the computations are independent for each sample f, so for notational brevity we drop the indexing by f. The Dirichlet prior, which governs the distribution of θ, is then optimized by leveraging isoform abundance θ_f estimates from all samples. This step serves as the key mechanism for integrating multi-sample information.

E: Soft-assigning isoforms to reads. In the E-step we assign reads to isoforms given the current estimate of isoform proportions θ. For each read r this involves finding the distributions,

$$q(a_r) = P(a_r|\theta, d_r) \propto P(d_r|a_r)P(a_r|\theta) = \prod_i p_{ri}^{a_{ri}} \theta_i^{a_{ri}} \tag{3}$$

$$= \text{Multinomial}(a_r|1, z_r) \quad \text{where} \quad z_{ri} = \frac{p_{ri}\theta_i}{\sum_{i'} p_{ri'}\theta'_i} = \mathbb{E}[a_{ri}]. \tag{4}$$

z_{ri} is our current estimate of the posterior probability that read r came from isoform i.

M: Estimating isoform abundance. For the M-step we need the expected log-likelihood which from Eq. (2) is,

$$\begin{aligned}\mathbb{E}_q[\log P(a, \theta|d)] &= \sum_i (\alpha_i - 1)\log\theta_i + \sum_r \mathbb{E}[a_{ri}]\log\theta_i \\ &= \sum_i (n_i + \alpha_i - 1)\log\theta_i \quad \text{where} \quad n_i = \sum_r \mathbb{E}[a_{ri}] = \sum_r z_{ri}\end{aligned}$$

This is the log pdf of a Dirichlet distribution on θ with parameter vector $\alpha' = (\alpha_1+n_1, \ldots, \alpha_I+n_I)$. Standard EM would set θ to the mode of this distribution, but we find using the posterior mean $\alpha'_i / \sum_{i'=1}^{I} \alpha'_{i'}$ to be more numerically stable.

Empirical Bayes prior update. To optimize the Dirichlet hyperparameter vector α, we maximize the likelihood of θ_f across all samples, i.e.,

$$\mathcal{L}(\alpha) = \sum_{f=1}^{F} \left[\log \frac{1}{B(\alpha)} + \sum_{i=1}^{I} (\alpha_i - 1)\log\theta_{fi} \right] \tag{5}$$

The normalizing constant is $B(\alpha) = \prod_i \Gamma(\alpha_i) / \Gamma\left(\sum_i \alpha_i\right)$. The likelihood in Eq. (5) cannot be optimized analytically so we use gradient ascent to optimze α.

In the next iteration, the updated $\hat{\alpha}$ is used in the posterior expectation calculation of θ, and this process is repeated iteratively until convergence defined as when the difference between successive estimates, $|\theta^{(s)} - \theta^{(s-1)}|$, falls below 0.001.

2.3 Dataset

PacBio and Illumina data: We utilized PacBio HiFi long reads [11] alongside compatible Illumina short reads for our analysis, as illustrated in Fig. 7. The

dataset was collected across various stages (days 0–5) of the differentiation protocol for WTC11-derived primordial endothelial cells, with 2 biological replicates for each day. For multi-sample models, we specifically focused on data from day 0 and day 5. Long reads were aligned to the human transcriptome using `Minimap2 [v2.17]` [15], while short reads were processed with `STAR [v2.5.2a]`[7]. To resemble a more typical LR coverage, the LR alignment files were downsampled to 10% of the original reads. Post-downsampling, the dataset consisted of approximately 100 million short reads and 10 million long reads for each sample, originating from around 150,000 transcripts.

Simulated data. Using the LRGASP-simulation pipeline [21], two samples, each containing PacBio and Illumina reads, were simulated. PacBio reads were generated using `IsoSeqSim [v0.1]` [30], while Illumina reads were simulated with `RSEM [v1.3.0]` [14]. The entire human genome served as the reference for this simulation. Transcript abundance estimates from real PacBio reads (as described in the previous section) were used to create the expression profile. Each simulated sample consisted of 10 million short reads and 1 million long reads. The ground truth isoform abundance file, obtained from the human genome, transcriptome, and annotation files using the LRGASP pipeline, contained approximately 149,500 isoforms per sample.

3 Results

To understand how multi-sample learning provides advantages over single-sample processing, we evaluate our model in four different setups as shown in (Fig. 7): *SS (Single Sample)*, *MS (Multi-Sample)*, *Merged SS (Merged Single Sample)*, and *Merged MS (Merged Multi-Sample)*. Each setup varies in how read data is provided to the model and how abundance estimates are computed.

In SS baseline, each sample is treated completely independently without any information sharing prior optimization. In MS, one LR and one SR dataset, are treated as unique samples (with their own θ_f), while still being part of a unified analysis with a shared Dirichlet prior. This allows us to assess the benefits of multi-sample variability learning over single-sample estimation. In merged SS, SR and LR from the same biological replicate are combined into a single "sample" before being processed through EM without any prior optimization, giving a single abundance estimate per ground truth proportion (θ), allowing the model to integrate complementary information from different read types. Finally, merged MS integrates both sample merging and multi-sample learning. This configuration evaluates whether aggregating information across multiple samples improves the robustness of abundance estimation, even after merging short and long reads within each sample. For simulation data, all four setups were evaluated, whereas for real data, only SS and MS were conducted.

To comprehensively evaluate the accuracy and reliability of isoform quantification, we employ four metrics: Spearman correlation (SC), Pearson correlation (PC), Median Relative Difference (MRD), and Irreproducibility Measure (IM); all definitions can be found in Appendix 5. Each of these metrics cap-

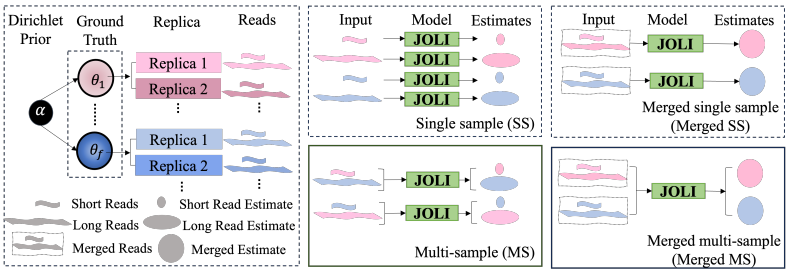

Fig. 2: **Overview of the experimental design.** The leftmost section illustrates the data generation process. Dirichlet prior (α) generates multiple ground truth proportions ($\theta_1 \ldots \theta_f$), each of which produces multiple replicates. These replicates, generate both short and long sequencing reads. The right side of the figure presents the four experimental setups for evaluating the model: *single sample (SS)*, *multi-sample (MS)*, *merged single sample (merged SS)*, and *merged multi-sample (merged MS)*.

tures a distinct aspect of model performance, ensuring a balanced assessment. SC measures rank-based association, making it robust to non-linear scaling and outliers, while PC captures linear relationships, ensuring proportionality with true expression levels or replicates. Higher values of both indicate better performance. For simulated data where true abundances are known, MRD measures the relative deviation of estimated values from the ground truth, capturing error magnitude. Finally, IM measures transcript abundance irreproducibility across replicates in real data. It is computed based on the coefficient of variation (CV) of transcript abundances. Lower MRD indicates better approximation of true expression levels, while lower IM signifies higher reproducibility. For simulation data, evaluation metrics were computed with respect to the ground truth, while for real data, they were assessed based on replicates.

Simulations results: Fig. 3 compares estimated transcript abundances $\log_e(\text{TPM} + 1)$ against ground truth across different evaluation setups. Comparing SS and MS setups, we observe that MS learning consistently improves performance. For long reads, SC and PC improve by approximately 9.8% and 9.9% (Fig. 3, left). For short reads, SC increases by 7.7% and PC by 0.4%(Fig. 3, center). The merged setup, combining long and short reads per sample, shows SC improving by 2.9% and PC by 0.2% (Fig. 3, right).

Fig. 4 highlights the contribution of MS learning in enhancing the noise-to-signal ratio for low- and moderate-expressed isoforms by assessing SC, PC, and MRD across three transcript expression categories—low, moderate, and high. The classification is based on ground truth abundance, with varying mean and standard deviation detailed in Table 2. Correlation is particularly enhanced for low and moderate expressed transcripts. For LR (Fig. 4, left), JOLI MS outperforms JOLI SS at moderate expression levels in SC and PC but struggles at low levels with a higher MRD. For SR (Fig. 4, center), JOLI MS improves MRD

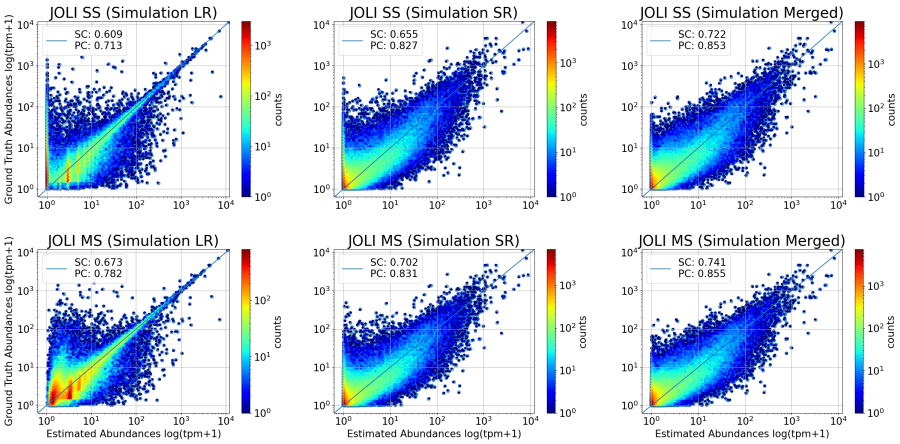

Fig. 3: **Scatter plots of estimated vs. ground truth transcript abundances** ($\log_e(\text{TPM}+1)$) for the simulation dataset under different experimental setups. Multi-sample learning improves abundance estimation, with higher SC and PC, demonstrating the benefits of leveraging cross-sample information.

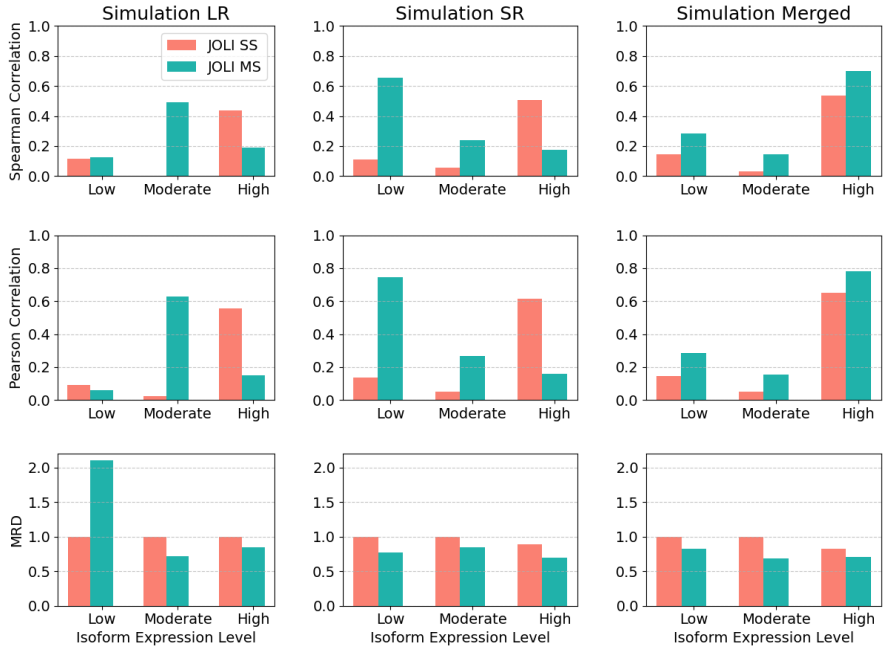

Fig. 4: **SC, PC, and MRD comparisons for low, moderate, and highly expressed isoforms** in simulated data. For LR (left), JOLI MS outperforms JOLI SS in SC and PC at moderate expression but has higher MRD at low levels; for SR (center), JOLI MS improves MRD across all levels and outperforms JOLI SS in SC and PC at low and moderate expression; in merged experiments (right), JOLI MS shows higher SC and PC at low and moderate expression with consistently better MRD at low and moderate levels.

Expression Level	# of Isoforms	Avg. Abundance	Std Dev
Low	50,531	0.087	0.055
Moderate	50,731	0.479	0.185
High	49,831	1.958	1.076

Table 2: Number of isoforms, average abundance ($\log_e(\text{TPM}+1)$), and standard deviation across different abundance classes.

across all expression levels and outperforms JOLI SS in SC and PC for low and moderate expression. In merged experiments (Fig. 4, right), JOLI MS follows the same trend as SR, exhibiting higher SC and PC at both low and moderate expression levels, while consistently achieving better MRD across all expression ranges..

Fig. 5a showcases a performance comparison of JOLI SS and MS against state-of-the-art (SOTA) methods, including `lr-Kallisto [v0.51.1]` and `Oarfish [v0.6.5]` for LR, as well as `Kallisto [v0.51.1]` for SR. The commands used to run the SOTA methods can be found in Appendix 5. For LR in Fig. 5a (left), the lower SC (0.67) and PC (0.78) for JOLI MS compared to lr-kallisto (0.85, 0.93) and Oarfish (0.81, 0.89) indicate that while JOLI MS captures overall isoform abundance, it struggles with precise ranking and consistency in absolute expression values, likely due to increased variance in MS learning. The most notable advantage of JOLI MS is in MRD, where it achieves 0.61, significantly outperforming JOLI SS and Oarfish (1.0) and approaching lr-kallisto (0.33). This highlights that despite lower SC and PC, JOLI MS reduces systematic biases found in single-sample methods and improves reliability in abundance estimates by mitigating quantification errors. On the other hand, JOLI MS demonstrates a much more competitive performance in SR results Fig. 5a (right). It outperforms kallisto in PC (0.83 vs. 0.78), a slightly lower SC (0.70 vs. 0.72), and a comparable MRD (0.78 vs. 0.77). These results indicate that JOLI MS effectively preserves absolute expression values while maintaining ranking accuracy and quantification error at levels similar to kallisto.

PacBio and Illumina results: Fig. 5b presents a performance comparison of JOLI for PacBio LR and Illumina SR. Fig. 5b (left) shows that both for LR and SR, JOLI MS improves over JOLI SS across all metrics. The SC of JOLI MS (0.80) is slightly lower than lr-kallisto (0.84) but higher than Oarfish (0.72), indicating that MS learning enhances rank preservation. PC remains consistently high across methods, with JOLI MS (0.94) performing on par with JOLI SS and lr-kallisto (0.92). The most significant improvement is seen in the IM, where JOLI MS (0.53) substantially reduces irreproducibility compared to JOLI SS (1.0) and even outperforms Oarfish (0.93), though lr-kallisto (0.50) achieves the lowest IM. This suggests that JOLI MS improves reproducibility while maintaining high correlation with the estimated transcript abundances. For SR, Fig. 5b (right) shows JOLI MS performing competitively with kallisto, achieving higher SC (0.91 vs. 0.85), comparable PC (0.98 vs. 0.96), and lower IM (0.53 vs. 0.60), while outperforming JOLI SS across all metrics. Overall, these results demonstrate

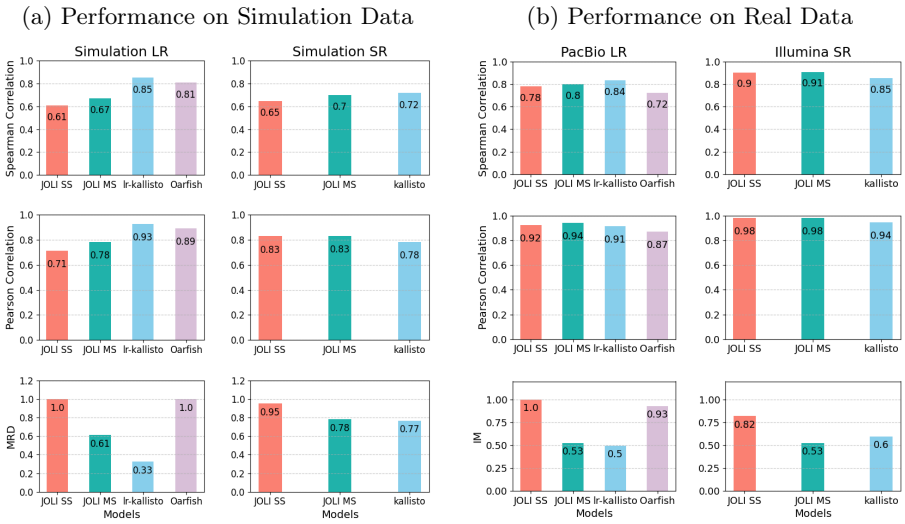

Fig. 5: **Performance comparison of JOLI models with SOTA for both simulation and real data.** (a) Simulation results compare JOLI MS and JOLI SS with lr-kallisto and Oarfish for LR and kallisto for SR across SC (top), PC (middle), and MRD (bottom), showing that JOLI MS reduces MRD while trailing lr-kallisto and Oarfish in SC and PC. (b) Real data results for PacBio LR and Illumina SR indicate that JOLI MS improves over JOLI SS in all metrics, reducing IM while maintaining high correlation, performing similarly to lr-kallisto and outperforming Oarfish. For SR, JOLI MS competes with kallisto, achieving higher SC, comparable PC, and lower IM, emphasizing its reproducibility and ranking consistency.

that JOLI MS provides notable improvements in reproducibility and ranking consistency, making it a strong contender in real-data settings despite minor differences.

4 Discussion

We have introduced JOLI, a novel hierarchical model for isoform quantification that jointly analyzes SR and LR sequencing data from multiple samples. Our approach extends the standard EM framework by incorporating information sharing across multiple related samples through an empirical Bayes framework which learns a common prior to capture shared variability. By integrating both sequencing technologies, JOLI aims to combine the accuracy and throughput of SR data with the reduced ambiguity of LR data. During the development of JOLI, another joint model, MPAQT [4], was proposed to integrate LR and SR. However, it does not model multiple samples, relies on precomputed LR

abundance estimates, requires memory-intensive simulated data, and assumes unambiguous LR alignment.

Through benchmarking on both simulated and real datasets, we have demonstrated that JOLI MS outperforms its SS counterpart, particularly in reducing estimation errors and improving reproducibility. In simulations, JOLI MS consistently outperforms JOLI SS across different sequencing setups. Further analysis indicates that MS learning has the most significant impact on low- and moderate-abundance transcripts, which are typically harder to quantify. Comparison with SOTA methods shows that JOLI MS achieves competitive results, particularly in error reduction, highlighting its ability to generate more reliable abundance estimates. On real datasets, JOLI MS continues to show strong performance improvements over JOLI SS, demonstrating the practical advantages of MS learning in handling biological variability and noise, while performing at the same level as other state-of-the-art methods.

Our findings demonstrate that MS learning is particularly effective in improving reproducibility and mitigating systematic biases, making it an essential component of accurate isoform quantification. This motivates us to integrate joint SR and LR, and MS, learning into existing state-of-the-art methods to fully leverage its benefits.

Code Availability: The source code is available at `https://github.com/ArghamitraT/JOLI`. This code is based on `https://github.com/a-slide/NanoCount`.

References

1. Akbari, M., Hansen, M.D., Halgunset, J., Skorpen, F., Krokan, H.E.: Low copy number dna template can render polymerase chain reaction error prone in a sequence-dependent manner. The Journal of molecular diagnostics **7**(1), 36–39 (2005)
2. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q.: Opportunities and challenges in long-read sequencing data analysis. Genome biology **21**(1), 30 (2020)
3. Ansorge, W.J.: Next-generation dna sequencing techniques. New biotechnology **25**(4), 195–203 (2009)
4. Apostolides, M., Choi, B., Navickas, A., Saberi, A., Soto, L.M., Goodarzi, H., Najafabadi, H.S.: Accurate isoform quantification by joint short-and long-read rna-sequencing. BioRxiv (2024)
5. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic rna-seq quantification. Nature biotechnology **34**(5), 525–527 (2016)
6. Pacific Biosciences of California, I.P.: Pacific biosciences of california, inc. (pacbio) (January 2025)
7. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
8. Gleeson, J., Leger, A., Prawer, Y.D., Lane, T.A., Harrison, P.J., Haerty, W., Clark, M.B.: Accurate expression quantification from nanopore direct rna sequencing with nanocount. Nucleic acids research **50**(4), e19–e19 (2022)

9. Hensman, J., Papastamoulis, P., Glaus, P., Honkela, A., Rattray, M.: Fast and accurate approximate inference of transcript expression from rna-seq data. Bioinformatics **31**(24), 3881–3889 (2015)
10. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V., et al.: Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell **158**(4), 929–944 (2014)
11. Hon, T., Mars, K., Young, G., Tsai, Y.C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C., et al.: Highly accurate long-read hifi sequencing data for five complex genomes. Scientific data **7**(1), 399 (2020)
12. Hu, Y., Fang, L., Chen, X., Zhong, J.F., Li, M., Wang, K.: Liqa: long-read isoform quantification and analysis. Genome biology **22**(1), 182 (2021)
13. Jousheghani, Z.Z., Patro, R.: Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. bioRxiv (2024)
14. Li, B., Dewey, C.N.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics **12**, 1–16 (2011)
15. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018)
16. Li, J.J., Huang, H., Bickel, P.J., Brenner, S.E.: Comparison of d. melanogaster and c. elegans developmental stages, tissues, and cells by modencode rna-seq data. Genome research **24**(7), 1086–1101 (2014)
17. Li, W., Feng, J., Jiang, T.: Isolasso: a lasso regression approach to rna-seq based transcriptome assembly. In: Research in Computational Molecular Biology: 15th Annual International Conference, RECOMB 2011, Vancouver, BC, Canada, March 28-31, 2011. Proceedings 15. pp. 168–188. Springer (2011)
18. Li, W.V., Zhao, A., Zhang, S., Li, J.J.: Msiq: joint modeling of multiple rna-seq samples for accurate isoform quantification. The annals of applied statistics **12**(1), 510 (2018)
19. Loving, R.K., Sullivan, D.K., Reese, F., Rebboah, E., Sakr, J., Rezaie, N., Liang, H.Y., Filimban, G., Kawauchi, S., Oakes, C., et al.: Long-read sequencing transcriptome quantification with lr-kallisto. bioRxiv (2024)
20. (ONT), O.N.T.: Oxford nanopore technologies (January 2025)
21. Pardo-Palacios, F.J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J.E., De María, M., Adams, M.S., Balderrama-Gutierrez, G., et al.: Systematic assessment of long-read rna-seq methods for transcript identification and quantification. Nature methods pp. 1–15 (2024)
22. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**(4), 417–419 (2017)
23. Perocchi, F., Xu, Z., Clauder-Münster, S., Steinmetz, L.M.: Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin d. Nucleic acids research **35**(19), e128 (2007)
24. Pimentel, H., Bray, N.L., Puente, S., Melsted, P., Pachter, L.: Differential analysis of rna-seq incorporating quantification uncertainty. Nature methods **14**(7), 687–690 (2017)
25. Rhoads, A., Au, K.F.: Pacbio sequencing and its applications. Genomics, Proteomics and Bioinformatics **13**(5), 278–289 (2015)
26. Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., Ge, Q.: Bias in rna-seq library preparation: current challenges and solutions. BioMed research international **2021**(1), 6647597 (2021)

27. Song, L., Sabunciyan, S., Yang, G., Florea, L.: A multi-sample approach increases the accuracy of transcript assembly. nat commun. 2019; 10 (1): 5000
28. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology **28**(5), 511–515 (2010)
29. Waks, Z., Klein, A.M., Silver, P.A.: Cell-to-cell variability of alternative rna splicing. Molecular systems biology **7**(1), 506 (2011)
30. Wang, Y.: Isoseqsim: Iso-seq reads simulator for pacbio and ont full-length isoform sequencing technologies. github
31. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., Au, K.F.: Nanopore sequencing technology, bioinformatics and applications. Nature biotechnology **39**(11), 1348–1365 (2021)

5 Appendix

5.1 Performance Evaluation Metric

Spearman Correlation (SC): We denote θ_i as the estimated abundance and $\hat{\theta}_i$ as the ground truth abundance of transcript i $(i = 1, 2, \ldots, I)$. The Spearman correlation between $\log(\theta + 1)$ and $\log(\hat{\theta} + 1)$ is given by:

$$\rho = 1 - \frac{6 \sum_{i=1}^{I} d_i^2}{I(I^2 - 1)}$$

where

$$d_i = R(\log(\theta_i + 1)) - R(\log(\hat{\theta}_i + 1))$$

represents the difference between the ranks of the log-transformed estimated abundance θ and ground truth $\hat{\theta}$.

Pearson Correlation (PC): We denote θ_i as the estimated abundance and $\hat{\theta}_i$ as the ground truth abundance of transcript i $(i = 1, 2, \ldots, I)$. The Pearson correlation coefficient between $\log(\theta + 1)$ and $\log(\hat{\theta} + 1)$ is given by:

$$r = \frac{\sum_{i=1}^{I} \left(\log(\theta_i + 1) - \overline{\log(\theta + 1)}\right) \left(\log(\hat{\theta}_i + 1) - \overline{\log(\hat{\theta} + 1)}\right)}{\sqrt{\sum_{i=1}^{I} \left(\log(\theta_i + 1) - \overline{\log(\theta + 1)}\right)^2} \sqrt{\sum_{i=1}^{I} \left(\log(\hat{\theta}_i + 1) - \overline{\log(\hat{\theta} + 1)}\right)^2}}$$

where

$$\overline{\log(\theta + 1)} = \frac{1}{I} \sum_{i=1}^{I} \log(\theta_i + 1),$$

$$\overline{\log(\hat{\theta} + 1)} = \frac{1}{I} \sum_{i=1}^{I} \log(\hat{\theta}_i + 1)$$

represent the mean of the log-transformed estimated and true abundance values, respectively.

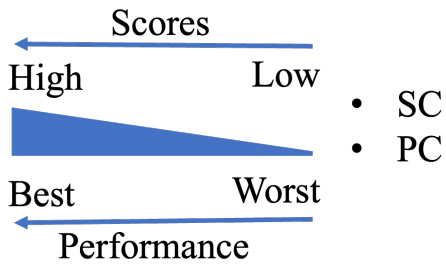

Fig. 6: For SC and PC **higher** values mean **better** performance

Median Relative Difference (MRD): The MRD measures the median percentage of estimated values that deviate from the true ones:

$$MRD = \text{median}\left(\frac{\|\hat{\theta}_i - \theta_i\|_2}{\|\theta_i\|_2}\right)$$

Irreproducibility Measure (IM): Denote θ_{ijk} as the abundance estimation of transcript $(i = 1, 2, \ldots, l)$ in a sample, where $j(j = 1, 2, \ldots, J)$ represents different groups (i.e., conditions or tissues) and $k(k = 1, 2, \ldots, K)$ represents different replicates within the group.

Coefficient of variation $\mathbf{CV_{ij}}$ of $\log \theta_{ijk}$ $(k = 1, 2, \ldots, K)$

$$CV_{ij} = \frac{s_{ij}}{u_{ij}},$$

where s_{ij} and u_{ij} are the sample standard deviation and mean of abundance estimates.

$$s_{ij} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\log(\theta_{ijk} + 1) - u_{ij})^2},$$

$$u_{ij} = \frac{1}{K}\sum_{k=1}^{K}\log(\theta_{ijk} + 1).$$

Irreproducibility measure based on $\mathbf{CV_{ij}}$

$$IM = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} CV_{ij}^2}$$

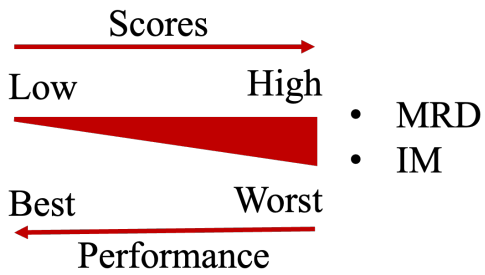

Fig. 7: For MRD and IM **lower** values mean **better** performance

5.2 Commands to run SOTA

lr-Kallisto:

```
kallisto bus -x bulk --threshold 0.8 -t 32 --long --unmapped
-i index_file reads_file -o output_dir

bustools sort -t 30 output_dir/output.bus -o sorted.bus

bustools count sorted.bus -t transcripts.txt -e matrix.ec -o count
--cm -m -g t2g.txt

kallisto quant-tcc -t 32 --long -P PacBio count.mtx
-i index_file -e count.ec.txt -o output_dir
```

Kallisto:

```
kallisto bus -x bulk --threshold 0.8 -t 32 --paired
-i index_file reads_file1 reads_file2 -o output_dir

bustools sort -t 30 output_dir/output.bus -o sorted.bus

bustools count sorted.bus -t transcripts.txt -e matrix.ec -o count
--cm -m -g t2g.txt
```

```
kallisto quant-tcc -t 32 count.mtx
-i index_file -e count.ec.txt -o output_dir
```

Oarfish: Reads were aligned using `Minimap2 [v2.28]`.

```
minimap2 --eqx -N 100 -ax map-pb transcripts.fasta reads.fasta
| samtools view -@4 -b -o aln.bam

oarfish -j 16 -a aln.bam -o output_dir/ --filter-group no-filters --model-coverage
```