



# Perplexity: Evaluating Transcript Abundance Estimation in the Absence of Ground Truth

Jason Fan ✉ 🏠 

University of Maryland, College Park, MD, USA

Skylar Chan ✉ 🏠 

University of Maryland, College Park, MD, USA

Rob Patro ✉ 🏠 

University of Maryland, College Park, MD, USA

---

## Abstract

There has been rapid development of probabilistic models and inference methods for transcript abundance estimation from RNA-seq data. These models aim to accurately estimate transcript-level abundances, to account for different biases in the measurement process, and even to assess uncertainty in resulting estimates that can be propagated to subsequent analyses. The assumed accuracy of the estimates inferred by such methods underpin gene expression based analysis routinely carried out in the lab. Although hyperparameter selection is known to affect the distributions of inferred abundances (e.g. producing smooth versus sparse estimates), strategies for performing model selection in experimental data have been addressed informally at best.

Thus, we derive *perplexity* for evaluating abundance estimates on fragment sets directly. We adapt perplexity from the analogous metric used to evaluate language and topic models and extend the metric to carefully account for corner cases unique to RNA-seq. In experimental data, estimates with the best perplexity also best correlate with qPCR measurements. In simulated data, perplexity is well behaved and concordant with genome-wide measurements against ground truth and differential expression analysis.

To our knowledge, our study is the first to make possible model selection for transcript abundance estimation on experimental data in the absence of ground truth.

**2012 ACM Subject Classification** Applied computing → Computational biology

**Keywords and phrases** RNA-seq, transcript abundance estimation, model selection

**Digital Object Identifier** 10.4230/LIPIcs.WABI.2021.4

**Supplementary Material** *Software (Source Code)*: <https://github.com/COMBINE-lab/perplexity>  
*Software (Source Code)*: <https://github.com/COMBINE-lab/perplexity-paper>

**Funding** This work is supported by NIH R01 HG009937, and by NSF CCF-1750472, and CNS-1763680. Additionally, JF is supported by the NSF GRFP award no. DGE-1840340.

*Conflicts of interest.* RP is a co-founder of Ocean Genomics, Inc.

## 1 Introduction

Due to its accuracy, reproducibility, simplicity and low cost, RNA-seq has become one of the most popular high-throughput sequencing assays in contemporary use, and it has become the *de facto* method for the profiling of gene and transcript expression in many different biological systems. While there are many uses for RNA-seq that span the gamut from *de novo* transcriptome assembly [5, 10] through meta-transcriptome profiling [30], one of the most common uses is to interrogate the gene or isoform-level expression of known (or newly-assembled) transcripts, often with the subsequent goal of performing a differential analysis between conditions of interest.



© Jason Fan, Skylar Chan, and Rob Patro;

licensed under Creative Commons License CC-BY 4.0

21st International Workshop on Algorithms in Bioinformatics (WABI 2021).

Editors: Alessandra Carbone and Mohammed El-Kebir; Article No. 4; pp. 4:1–4:22

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Because of the popularity of gene and transcript expression profiling using RNA-seq, considerable effort has been expended in developing accurate, robust and efficient computational methods for inferring transcript abundance estimates from RNA-seq data. Some popular approaches focus on counting the aligned RNA-seq reads that overlap genes in different ways [1, 19]. However, these approaches have no principled way to deal with reads that align well to multiple loci (e.g. to different isoforms of a gene, or between sequence-similar regions of related genes), and this restricts their use primarily to gene-level analysis, where they may still under-perform more sophisticated approaches that attempt to resolve fragments of ambiguous origin [33].

Alternatively, many approaches offer the ability to estimate transcript-level expression using RNA-seq data (which can, if later desired by a user, be aggregated to the gene-level). The majority of these approaches perform statistical inference over a probabilistic generative model of the experiment based either on sufficient statistics of counts [13, 36] or the set of fragment alignments themselves [17]. Moreover, in addition to methods focused on deriving point estimates for transcript abundances, there has been considerable development of probabilistic Bayesian approaches for this inference problem [9, 11, 22, 23, 24, 26], as well as recent attempts at multi-sample probabilistic models for simultaneous experiment-wide transcript abundance estimation [14, 15]. Bayesian approaches can sometimes offer more accurate or robust inference than methods based strictly on maximum likelihood estimation, but these Bayesian models invariably expose prior distributions, with associated hyperparameters, upon which the resulting inferences depend.

Interestingly, the recommended best practices suggested by the different Bayesian (or variational Bayesian) approaches for selecting hyperparameters differ. Specifically, Nariai et al. [22] evaluate performance varying the prior used in their variational Bayesian expectation maximization (VBEM)-based method, and they conclude that a small prior (i.e.  $\alpha < 1$ ) leads to a sparse solution, which, in turn, results in improved accuracy. On the other hand, Hensman et al. [11] perform inference using a prior of  $\alpha = 1$  read per transcript. They find that, doing so, their method produces the most *robust* estimates (i.e. with the highest concordance between related replicates) that are also more accurate under different metrics that they measure. Their conclusion is that methods adopting a maximum likelihood model inferred using an expectation maximization procedure tend to produce sparse estimates close to the boundary of the parameter space which leads to less robust estimation among related samples. Unfortunately, regardless of how prior studies have argued for a “better” prior, none provide an empirical or practical procedure for model selection. Rather, they show that a value works well across a range of data under some evaluation metric, and set this as the default value for all inference tasks. Given the number of existing methods that can make use of prior information (including methods like those by Srivastava et al. [34] for single-cell data, or those by Liu et al. [20] that use orthogonal modalities of data to set priors), it becomes increasingly important to develop methods that lets one robustly and automatically select an appropriate prior (hyperparameter) for these algorithms.

To perform model (or hyperparameter) selection for transcript abundance estimators, one must be able to evaluate estimated abundances. However, evaluation of abundance estimates remains a challenge for current methods on experimental data where ground truth is completely absent. Notably, evaluation of transcript abundance estimators on experimental data have relied on careful experiment design that enables comparisons to complementary assays (e.g. correlation with qPCR) or measurements (e.g. concordance with known mixing proportions or spike-ins) [35]. Such evaluation procedures vary from study-to-study, and are simply not possible when complementary experiments are not designed or available. Thus, the natural question is then: *can the quality of transcript abundance estimates be meaningfully evaluated on the set of given fragments directly?*

It may initially be unintuitive to think that the “goodness” of a transcript abundance estimate can be evaluated in the absence of ground truth. However, in a related line of research, likelihood-based metrics for assessing the quality of *de novo* assemblies, where ground truth is unavailable, have been explored. For example, Rahman and Pachter [27] developed a method to compute the likelihoods of assembled genomes; Li et al. [18] developed a likelihood-based score to evaluate transcriptome assemblies; Smith-Unna et al. [32] developed a method to assess the quality of assembled contigs in transcriptomes; and Clark et al. [6] developed a method that is applicable to both genome and metagenomic assemblies. Furthermore, if we look to other unsupervised problem settings where ground truth annotations are absent, metrics for measuring the “goodness” of estimated models with latent parameters not only exist, but are regularly used. For example, metrics such as the silhouette score used to evaluate clustering algorithms come to mind [29]. In fact, evaluation of unsupervised probabilistic models, especially language and topic models in natural language processing, is commonplace [4, 12]. Specifically, *perplexity*, the inverse geometric mean per-*word* likelihood of a held-out test set, has been ubiquitously used to compare models [4].

In this work, we derive perplexity for transcript abundance estimation with respect to held-out per-*read* likelihoods. As we shall see, the perplexity of a held-out fragment set given an abundance estimate, computed via a quantify-then-validate approach, is a theoretically and experimentally motivated measure of the quality of the given estimate. Notably, perplexity quantifies an important biologically motivated intuition – that a good abundance estimate ought to generalize and generate the validation set, which is, in a sense, a form of a technical replicate, with high probability.

Perplexity can be used wherever the assessment of the quality of abundance estimates is desired. For example, perplexity can be used to compare different transcript abundance estimation algorithms or, as suggested above, to perform model selection to obtain the most accurate estimates from a given algorithm. In this work, we focus on experimentally assessing perplexity with respect to the latter, model selection for the prior used to estimate abundances with `salmon` [26]. In `salmon`, the reads-per-transcript prior size is a hyperparameter that controls its preference for inferring sparse or smooth abundance estimates. Notably, the problem of model selection offers a succinct assessment and immediately useful application of how perplexity can be computed to evaluate and compare the quality of candidate transcript abundance estimates.

## 1.1 Contributions

Theoretically, we derive and motivate a notion of *perplexity* for transcript abundance estimation – a metric for evaluating inferred estimates in the absence of ground truth. Experimentally, we demonstrate that perplexity for transcript abundance estimates is well behaved, and establish empirical correspondence between perplexity and other metrics that are more commonly used to demonstrate the “goodness” of transcript abundance estimates.

We summarize our experimental contributions as follows:

1. In experimental data from the Sequencing Quality Control (SEQC) consortium [35], we show that transcript abundance estimates with the lowest perplexity (lower is better) achieve the highest correlation with complementary qPCR measurements of biological replicates.
2. In simulated data, perplexity is concordant with respect to three measurements against ground truth: Spearman correlation with respect to expressed transcripts, AUROC with respect to unexpressed transcripts, and downstream differential transcript expression analysis.

Evidenced by these results, we propose perplexity as the first and, to our knowledge, only theoretically and experimentally justified metric for model selection for transcript abundance estimation in *experimental* data where ground truth is entirely absent.

## 2 Preliminaries: (Approximate) Likelihood for transcript abundance estimation

Before deriving *perplexity* for transcript abundance estimation, we shall briefly recall and define the necessary objects that pertain to the *likelihood* of the probabilistic model that underpins transcript abundance estimation (as in [17, 26]).

The transcript abundance estimation problem, or quantification, from short RNA-seq *fragments* (a term used to refer, generically, to either single reads or read pairs), is the problem of assigning each fragment  $f_j$  of an input fragment-set  $\mathcal{F} = \{f_1, \dots, f_N\}$  to its transcript of origin. For this work, we shall only consider quantification with respect to a given reference transcriptome whereby a quantifier maps each input fragment  $f_j$  to a transcript in an input set of reference transcripts  $\mathcal{T} = \{t_1, \dots, t_M\}$ .

Given the sequence of an input fragment, said fragment may align to more than one transcript,  $t_i$ , in the reference transcriptome  $\mathcal{T}$ . Here, the *de facto* method for determining transcript of origin for fragments that multi-map to more than one transcript is to view the true fragment to transcript assignment as a latent variable, and to infer the latent variable's expected value by performing inference in the underlying probabilistic model.

Assuming an appropriate normalization of alignment scores, we write the probability of observing a fragment,  $f_j$ , given that it originates from (or aligns to) transcript  $t_i$  to be  $P(f_j|t_i)$ . The probability that a molecule in a sample that is selected for sequencing is the transcript  $t_i$  is then  $P(t_i|\theta)$ , a multinomial over  $\mathcal{T}$ . Marginalizing over all possible alignments, the *likelihood* of observing the fragment set  $\mathcal{F}$  given model parameters  $\theta$  is,

$$\mathcal{P}(\mathcal{F} | \theta) = \prod_j \sum_{i=1}^M P(t_i | \theta) \cdot \mathcal{P}(f_j | t_i). \quad (1)$$

In this work, we shall work with the *range-factorized* equivalence class approximation of the likelihood that has proven to be effective and is efficient to compute [38]. Here, sets of fragments in  $\mathcal{F}$  that map to the same set of transcripts, and have similar conditional probabilities of arising from these transcripts, are said to belong to the equivalence class  $\mathcal{F}^q$  (indexed by  $q$ ). Instead of working with alignment probabilities  $\mathcal{P}(f_j|t_i)$  of each fragment, fragments in an equivalence class  $\mathcal{F}^q$  are approximated to have the same conditional probability  $\mathcal{P}(f_j|\mathcal{F}^q, t_i)$  for mapping to each transcript  $t_i$ . Let  $\mathcal{C}$  be the set of equivalence classes induced by  $\mathcal{F}$  and  $\Omega(\mathcal{F}^q)$  be the set of transcripts to which  $f \in \mathcal{F}^q$  map. The range-factorized equivalence class approximation of the likelihood  $\mathcal{P}(\mathcal{F} | \theta)$  is,

$$\mathcal{P}(\mathcal{F} | \theta) \approx \prod_{\mathcal{F}^q \in \mathcal{C}} \left( \sum_{t_i \in \Omega(\mathcal{F}^q)} P(t_i | \theta) \cdot \mathcal{P}(f_j | \mathcal{F}^q, t_i) \right)^{N^q}. \quad (2)$$

Here, the approximate likelihood can be computed over the number of unique equivalence classes, which is considerably smaller than the number of all possible alignments for all fragments.

### 3 Methods

We propose a subtle but instructive change in the usual computational protocol for evaluating transcript abundance estimates. We propose a *quantify-then-validate* approach which evaluates the quality of transcript abundance estimates directly on read-sets, analogous to *train-then-test* approaches for evaluating probabilistic predictors common in natural language processing (NLP) and other fields [3, Ch. 1.3]. Instead of quantifying all available fragments and then performing evaluation with respect to complementary measurements downstream, the quantify-then-validate approach validates and evaluates the quality of a given abundance estimate directly on a set of held-out *validation* fragments withheld from inference.

We derive and adapt from NLP, the notion of *perplexity* for transcript abundance estimation for this quantify-then-validate approach [4, 12]. Perplexity is computed given only an abundance estimate, and a held-out validation set of fragments as input. Thus, perplexity evaluates the quality of abundance estimates on fragments directly and can evaluate estimates from experimental data in the absence of ground truth. Most importantly, evaluating perplexity with the quantify-then-validate approach enables quantitative, evidence-based, cross-validated selection of hyperparameters for transcript abundance estimation methods that use them.

Perplexity for transcript abundance estimation quantifies the intuition that an abundance estimate for a given sample ought, with high probability, explain and generate the set of fragments of a technical replicate. The key observation is that the likelihood  $\mathcal{P}(\mathcal{F}|\theta)$  is simply a value that can be computed for any fragment set  $\mathcal{F}$  and any abundance estimate  $\theta$  (model parameters), irrespective of whether  $\theta$  is inferred from  $\mathcal{F}$ . It is the context and application of the likelihood,  $\mathcal{P}(\mathcal{F}|\theta)$ , that yields semantic meaning.

Given a fragment set,  $\mathbf{F}$ , over which one seeks to infer and evaluate abundance estimates, the quantify-then-validate procedure is as follows. First, partition the input set into a *quantified* set,  $\mathcal{F}$ , and a *validation* set,  $\hat{\mathcal{F}}$ . Second, “quantify” and infer abundance estimates (model parameters)  $\theta$  given the quantified set  $\mathcal{F}$ . Third, validate and compute the perplexity,  $PP(\hat{\mathcal{F}}, \theta)$  – the inverse geometric mean held-out per-read likelihood of observing the validation set,  $\hat{\mathcal{F}}$  – given model parameters  $\theta$  and the validation set  $\hat{\mathcal{F}}$ . The lower the perplexity, the better the parameters  $\theta$  describe the held-out fragments  $\hat{\mathcal{F}}$ , and the better the abundance estimate parameterized by  $\theta$  ought to be. In fact, if we believe that the generative model is truly descriptive of the distributions that arise from the underlying biological and technical phenomena, perplexity is, in expectation, minimized when the “true” latent parameters are inferred.

Formally, given an abundance estimate  $\theta$ , and a validation fragment-set  $\hat{\mathcal{F}} = \{\hat{f}_1, \dots, \hat{f}_{\hat{N}}\}$ , the perplexity for transcript abundance estimation is:

$$\begin{aligned} PP(\hat{\mathcal{F}}, \theta) &= \exp \left\{ -\frac{1}{\hat{N}} \log \mathcal{P}(\hat{\mathcal{F}} | \theta) \right\} = \exp \left\{ -\frac{1}{\hat{N}} \sum_{j=1}^{\hat{N}} \log \mathcal{P}(\hat{f}_j | \theta) \right\} \\ &= \exp \left\{ -\frac{1}{\hat{N}} \sum_{j=1}^{\hat{N}} \log \sum_{i=1}^M \mathcal{P}(t_i | \theta) \cdot \mathcal{P}(\hat{f}_j | t_i) \right\}. \end{aligned} \quad (3)$$

Crucially, the probability  $\mathcal{P}(\hat{f}_j | \theta)$  of observing each held out fragment given  $\theta$  is computed and marginalized over two terms,  $\mathcal{P}(\hat{f}_j | t_i)$  that depends only on the validation set of held-out fragments, and  $\mathcal{P}(t_i | \theta)$  that depends only on the given abundance estimate.

One particular application of the perplexity metric, which we explore here, is to select the best abundance estimate out of many candidate estimates arising from different hyperparameter settings for quantifiers. Thus, in this work, we use the range-factorized equivalence

class approximation for perplexity (as in Eq. 2) throughout [38]. Given the range-factorized equivalence classes,  $\hat{\mathcal{C}}$ , induced by the *validation* set,  $\hat{\mathcal{F}}$ , (where  $\hat{N}^q$  is the number of fragments in an equivalence class  $\hat{\mathcal{F}}^q \in \hat{\mathcal{C}}$ ) the approximation is:

$$PP(\hat{\mathcal{F}}, \theta) \approx \exp \left\{ -\frac{1}{\hat{N}} \sum_{\hat{\mathcal{F}}^q \in \hat{\mathcal{C}}} \hat{N}^q \log \left( \sum_{t_i \in \Omega(\hat{\mathcal{F}}^q)} P(t_i | \theta) \cdot \mathcal{P}(\hat{f}_j | \hat{\mathcal{F}}^q, t_i) \right) \right\}. \quad (4)$$

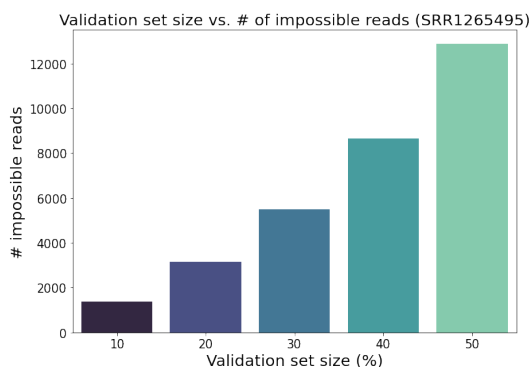
We use `salmon`'s selective-alignment based probabilistic model for conditional probabilities  $\mathcal{P}(\hat{f}_j | \hat{\mathcal{F}}^q, t_i)$  and effective lengths of transcripts, since the model and equivalence class approximation `salmon` uses has proven to be a fast and effective way to approximate the full likelihood [15, 38]. For the scope of this work, `salmon`'s format for storing range-factorized equivalence classes conveniently contains all relevant information and values to compute perplexity with vastly smaller space requirements than would be required to store per-fragment alignment probabilities  $P(f_j | t_i)$ .

### 3.1 “Impossible” fragments under parameter estimates $\theta$

We now address a perplexity-related issue that is unique to evaluating transcript abundance estimates – that an observed event in the validation set may be deemed “impossible” given model parameters  $\theta$ . The marginal probability,  $\mathcal{P}(\hat{f}_j | \theta)$ , for observing a fragment  $\hat{f}_j$  in the validation set given some abundance estimate,  $\theta$ , may actually be zero, even if said validation fragment aligns to the reference transcriptome. This occurs exactly when all transcripts,  $t_i$ , to which the validation fragment  $\hat{f}_j$  map are deemed unexpressed by  $\theta$  (i.e.  $P(t_i | \theta) = 0$  for all such transcripts). Here, we say that  $\hat{f}_j$  is an *impossible fragment* given  $\theta$ , and that  $\theta$  *calls*  $\hat{f}_j$  impossible. When impossible fragments are observed in the validation set, perplexity is not a meaningful measurement.

To illustrate how impossible fragments come to be, consider the toy example in which all fragments in a quantified set that align to transcripts  $A$ ,  $B$ , or  $C$  only ambiguously map to  $\{A, B\}$ , or to  $\{A, C\}$ . That is, no such fragments uniquely map – a phenomenon observed rather frequently for groups of similar isoforms expressed at low to moderate levels. Now, suppose that an abundance estimation model assigns all such fragments to transcript  $A$  and produces an estimate  $\theta$ . The quantifier may be satisfying a prior that prefers sparsity; or prefers to do so because transcript  $A$  is considerably shorter than transcripts  $B$  and  $C$ , which gives it a higher conditional probability under a length normalized model. In this case, the marginal probability,  $\mathcal{P}(\hat{f}_j | \theta)$ , of observing a validation fragment  $\hat{f}_j$  that maps to  $\{B, C\}$  is exactly zero given the parameters  $\theta$ .

As an example, we randomly withhold varying percentages of fragments from one sample (SRR1265495) as validation sets and use all remaining fragments to estimate transcript abundances with `salmon`'s default model (i.e. the VBEM model using prior size of 0.01 reads-per-transcript). Figure 1 shows that at all partitioned percentages, impossible fragments in the validation set are prevalent with respect to estimated abundances. In fact, due to the prevalence of impossible reads, perplexity as written in Eq. 4 is undefined (or infinite) for all estimates and all validation sets in the experiments below. An important observation in both the toy and experimental examples is that there likely exist better abundance estimates that would call fewer fragments impossible, while still assigning high likelihood to the rest of the (possible) fragments. For example, an abundance estimate that reserves even some small probability mass to transcript  $B$  in the toy example would not call the validation fragments in question impossible.



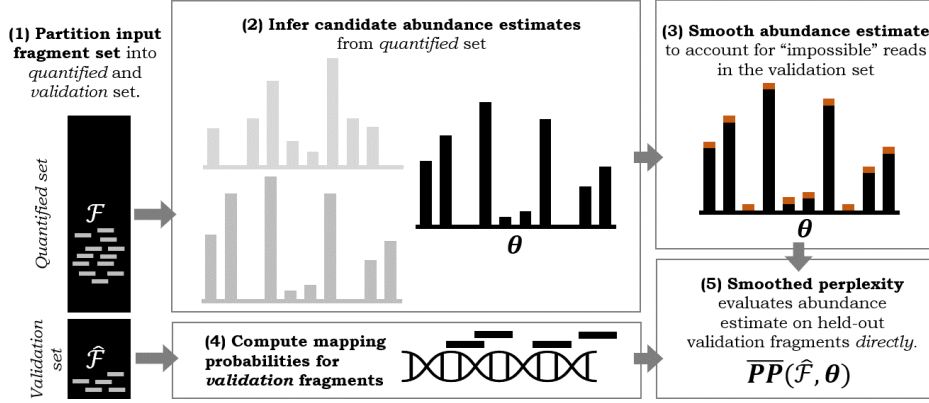
■ **Figure 1** Number of fragments called impossible versus withheld validation fragment set size for sample SRR1265495. All remaining fragments are used to estimate abundances using `salmon`'s VBEM model using default parameters (i.e. using a prior size of 0.01 reads-per-transcript).

### 3.2 Smoothed perplexity: accounting for “impossible” validation fragments

The problem with impossible fragments is not only that they exist. It is that, for a fixed validation fragment set, perplexity deems an abundance estimate that calls even one fragment impossible equally as bad as an abundance estimate that calls all fragments impossible. However, the former is clearly preferable to the latter. Furthermore, as we shall see in the experiments that follow, the number of fragments called impossible by an abundance estimate can actually be indicative of inaccuracies with respect to estimated abundances of transcripts called expressed by  $\theta$ . Thus, one must quantitatively account for impossible fragments to enable the comparison of estimates that call some validation fragments impossible.

Other fields that have adopted and used perplexity (e.g. NLP) usually sidestep the issue of impossible events entirely both by construction and pre-processing, working only with smoothed probabilistic models in which no-event has probability zero, or removing rare words from input language corpora. However, neither strategy is available nor appropriate for evaluating transcript abundance estimates. It is neither reasonable nor useful to amend and modify each of the many modern quantifiers to produce smooth outputs (outputs in which no transcript has truly zero abundance), and fragments and transcripts cannot be pre-processed away since the set of expressed transcripts cannot be identified *a priori*. One may also be tempted to simply remove impossible fragments from a validation set,  $\widehat{\mathcal{F}}$ , before computing a perplexity or hold out fragments – but this also is not a valid strategy. This is because two different abundance estimates  $\theta$  and  $\theta'$  may call different validation fragments in  $\widehat{\mathcal{F}}$  impossible, and comparisons of likelihoods  $P(\widehat{\mathcal{F}}|\theta')$  and  $P(\widehat{\mathcal{F}}|\theta)$  are only meaningful if the validation sets are the same (i.e.  $\widehat{\mathcal{F}} = \widehat{\mathcal{F}}'$ ). Furthermore, there is no straightforward strategy to sample and hold-out validation fragments so that no fragments are impossible. This is because most validation fragments cannot be determined to be impossible prior to abundance estimation, and any non-uniform sampling strategy would alter the underlying distributions that estimators aim to infer.

Thus, we propose a *smoothed perplexity* measure to evaluate the quality of abundance estimates in which a consistent smoothing scheme can be fairly applied to any given abundance estimate. By smoothing an input abundance estimate, impossible fragments result in a penalty instead of immediately shrinking  $P(\widehat{\mathcal{F}}|\theta)$  to zero. More concretely, we define smoothed perplexity given abundance estimate  $\theta$  to be the perplexity evaluated with respect to the



■ **Figure 2** Overview of the *quantify-then-validate* approach using *smoothed perplexity* to evaluate the quality of abundance estimates directly on fragment sets in the absence of ground truth. (1) An input fragment set is first partitioned into a *quantified* and a *validation* set. (2) Abundance estimates for different candidate models (e.g. for explored hyperparameters as part of model selection) are inferred from the *quantified* fragment set only. (3) To account for “impossible” fragments and avoid shrinkage to unbounded perplexities, given abundance estimates are smoothed (see Sections 3.1 and 3.2). (4) Mapping probabilities to the reference transcriptome are computed for fragments in the validation set. (5) *Smoothed perplexity* computed given each input abundance estimate and the held-out validation fragment set can be used to evaluate and perform model selection – the lower the perplexity, the better an abundance estimate describes the held-out set of validation fragments.

smoothed distribution  $\mathcal{P}(t_i | s_\beta(\theta))$ . The Laplacian smoothing scheme  $s_\beta(\theta)$  smooths input abundance estimate  $\theta$  by redistributing a small constant probability mass. Let  $\mathcal{P}(t_i | \theta) = \eta_i$ , and  $M$  be the number of transcripts in the reference, the smoothed distribution  $\mathcal{P}(t_i | s_\beta(\theta))$ , parameterized by  $\beta$ , is defined:

$$\mathcal{P}(t_i | s_\beta(\theta)) = \frac{\eta_i + \beta}{1 + M\beta}. \quad (5)$$

This is equivalent to adding, for each transcript  $t_i$  in the reference,  $\beta \cdot \sum_j^M c_j / \tilde{l}_j$  reads-per-nucleotide to the expected fragment counts  $c_i$  then re-normalizing to obtain TPMs, given the model parameters  $\theta$  and effective transcript lengths  $\tilde{l}_i$  (as defined in `salmon` [26]).

We are now ready to define *smoothed perplexity* in full. Given an abundance estimate  $\theta$  and a validation set of fragments  $\hat{\mathcal{F}}$ , the smoothed perplexity measure  $\overline{PP}(\hat{\mathcal{F}}, \theta)$  is,

$$\overline{PP}(\hat{\mathcal{F}}, \theta) = \exp \left\{ -\frac{1}{\hat{N}} \sum_{\hat{\mathcal{F}}^q \in \hat{\mathcal{C}}} \hat{N}^q \log \left( \sum_{t_i \in \Omega(\hat{\mathcal{F}}^q)} \mathcal{P}(t_i | s_\beta(\theta)) \cdot \mathcal{P}(\hat{f}_j | \hat{\mathcal{F}}^q, t_i) \right) \right\}. \quad (6)$$

We schematically illustrate how smoothed perplexity using the proposed quantify-then-validate protocol is computed to evaluate the quality of transcript abundance estimates in Figure 2.

For all following sections, for brevity, we shall use *perplexity* to mean *smoothed perplexity* unless stated otherwise.

### 3.3 Model selection using perplexity in practice

Arguably, one of the most useful outcomes of being able to evaluate the quality of abundance estimates in the absence of ground truth is the ability to perform model selection for transcript abundance estimation in experimental data. For those familiar with train-then-test



experimental protocols for model selection in machine learning or NLP, model selection for transcript abundance estimation *vis-a-vis* our proposed quantify-then-validate approach is analogous and identical in abstraction. However, since, to our knowledge, this work is the first to propose a quantify-then-validate approach for transcript abundance estimation, we shall briefly detail how perplexity ought to be used in practice.

Let us consider model selection via 5-fold cross-validation using perplexity given some fragment set  $\mathbf{F}$ . First,  $\mathbf{F}$  is randomly partitioned into five equal sized, mutually exclusive validation sets,  $\{\widehat{\mathcal{F}}_1, \dots, \widehat{\mathcal{F}}_5\}$  – and quantified sets are subsequently defined,  $\mathcal{F}_i = \mathbf{F} - \widehat{\mathcal{F}}_i$ . Now, suppose we desire to choose between  $L$  model configurations (e.g. from  $L$  hyperparameter settings). Then for each  $\ell$ -th candidate model, we produce a transcript abundance estimate from each  $i$ -th quantified set,  $\theta_i^{(\ell)}$ . To select the best out of the  $L$  candidate models, one simply selects the model that minimizes the average perplexity over the five folds,  $\frac{1}{5} \sum_i \overline{PP}(\widehat{\mathcal{F}}_i, \theta_i^{(\ell)})$ .

One additional practical consideration should also be noted. Given *any* pair of quantification and validation sets  $\mathcal{F}$  and  $\widehat{\mathcal{F}}$ , a validation fragment,  $\hat{f}_j \in \widehat{\mathcal{F}}$ , can be *necessarily impossible*. A necessarily impossible validation fragment is one that maps to a set of transcripts to which no fragments in the quantified set  $\mathcal{F}$  also map. Such a fragment will always be called impossible given any abundance estimate deriving from the quantified set  $\mathcal{F}$ , since no fragments in  $\mathcal{F}$  provide any evidence that transcripts to which  $\hat{f}_j$  map are expressed.

It is of limited meaning to evaluate estimates with respect to necessarily impossible fragments. For the purposes of this work, we shall consider the penalization of an abundance estimate only with respect to impossible fragments that are recoverable – in other words, fragments that could be assigned non-zero probability given a better abundance estimate inferable from  $\mathcal{F}$ . As such, we remove necessarily impossible validation fragments from  $\widehat{\mathcal{F}}$ , given  $\mathcal{F}$ , prior to computing perplexity.

## 3.4 Data

### 3.4.1 Sequencing Quality Control (SEQC) project data

We downloaded Illumina HiSeq 2000 sequenced data consisting of 100+100 nucleotide paired-end reads from the Sequencing Quality Control (SEQC) project [35]. SEQC samples are labeled by four different conditions  $\{A, B, C, D\}$ , with condition  $A$  being Universal Human Reference RNA and  $B$  being Human Brain Reference RNA from the MAQC consortium [31], with additional spike-ins of synthetic RNA from the External RNA Control Consortium (ERCC) [2]. Conditions  $C$  and  $D$  are generated by mixing  $A$  and  $B$  in 3:1 and 1:3 ratios, respectively.

In this work, we analyze the first four replicates from each condition sequenced at the Beijing Genomics Institute (BGI) – one of three official SEQC sequencing centers. For each sample, we aggregate fragments sequenced by all lanes from the flowcell with the lexicographically smallest identifier.<sup>1</sup> Quantitative PCR (qPCR) data of technical replicates for each sample in each condition are downloaded via the `seqc` BioConductor package.

### 3.4.2 Simulated lung transcript expression data

We simulated read-sets based on 10 sequenced healthy lung samples, with Sequence Read Archive accession number SRR1265{495-504} [16]. Transcript abundance estimates inferred by Salmon using the `--useEM` flag for each sample are used as ground truth abundances for

<sup>1</sup> Scripts to download and aggregate SEQC data are available at [github.com/thejasonfan/SEQC-data](https://github.com/thejasonfan/SEQC-data).

read simulation (expressed in transcripts per million (TPM) and expected read-per-transcript counts). Then, transcript abundances in samples `SRR1265{495-499}`, for 10% of transcripts expressed in at least one of the five samples, are artificially up or down regulated by a constant factor ( $2.0\times$ ) to simulate differential transcript expression. We treat the resulting read-per-transcript counts as ground truth, and generate for each sample a fragments set of 100+100 nucleotide paired-end reads using Polyester at a uniform error rate of 0.001 with no sequence specific bias [8].

### 3.5 Evaluation and experiments

The purpose of the experiments in this work are twofold. First, to establish the relationship and correspondence between perplexity and commonly used measures of goodness or accuracy in transcript abundance estimation. And second, to demonstrate how model and hyperparameter selection can be performed using perplexity. In particular, we perform and evaluate hyperparameter selection for `salmon` with respect to the prior size in the variational Bayesian expectation maximization (VBEM) model used for inference [26]. The user-selected prior size for the VBEM model in `salmon` encodes the prior belief in the number of reads-per-transcript expected for any inferred abundance estimate. This hyperparameter controls `salmon`'s preference for inferring sparse or smooth estimates – the smaller the prior size, the sparser an estimate `salmon` will prefer. As discussed above, prior studies on Bayesian models have not necessarily agreed on how sparse or smooth a good estimate ought to be [11, 22] – the experiments in this work aim to provide a quantitative framework to settle this disagreement.

We perform all experiments according to the proposed quantify-then-validate procedure and report results with respect to various metrics over a 5-fold cross-validation protocol. We set the smoothing parameter for perplexity to  $\beta = 10^{-8}$  for all experiments. We use the Ensembl human reference transcriptome GRCh37 (release 100) for all abundance estimation and analysis [37].

#### 3.5.1 Evaluation versus parallel SEQC qPCR measurements

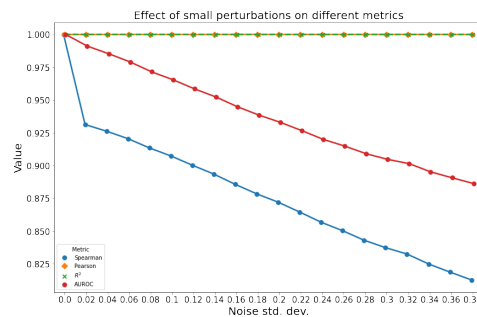
We analyze the relationship between perplexity and accurate abundance estimation in experimental data from the SEQC consortium. In SEQC data, we evaluate accuracy of abundances estimated by `salmon` by comparing estimates to qPCR gene expression data on biological replicates, a coarse proxy to ground truth. We evaluate the Spearman correlation between gene expressions of qPCR probed genes in SEQC replicates versus the corresponding abundance estimates. Gene expression from estimated transcript expression is aggregated via `txImport` [33] with transcript-to-gene annotations from `Ensembl.Hsapiens.v86` [28]. From gene expression data, Ensembl genes are mapped to corresponding Entrez IDs via `biomaRt` [7], and 897 genes are found to have a corresponding qPCR measurement in downloaded SEQC data. Expressions for genes with repeated entries in SEQC qPCR data are averaged.

#### 3.5.2 Evaluation versus ground truth on simulated data

In simulated data, since ground truth abundances are available, we compare estimated TPMs (computed by `salmon`) against ground truth TPMs under two metrics.

First, we consider the Spearman correlation with respect to known expressed transcripts (i.e. transcripts with non-zero expression in ground truth abundances). We choose to evaluate Spearman correlation with respect to ground truth non-zero TPMs because of the presence

of many unexpressed transcripts in the ground truth, meaning a high number of values tied at rank zero. Here, small deviations from zeros can lead to large changes in rank, leading to non-trivial differences in the resulting Spearman correlation metric. We demonstrate this phenomenon with respect to the ground truth abundance of a simulated sample (SRR1265495) with a mean TPM of 5.98, in which 49% of transcripts are unexpressed (82,358 / 167,268). We report the change in Pearson correlation,  $R^2$  score, and Spearman correlation of ground truth TPMs versus ground truth TPMs perturbed with normally distributed noise at varying standard deviations. As we can see from Figure 3, even small perturbations cause non-trivial changes in Spearman rank correlation, while changes in Pearson correlation are entirely imperceptible. The Pearson correlation, however, suffers from the well known problem that, in long-tailed distributions spanning a large dynamic range, like those commonly observed for transcript abundances, the Pearson correlation is largely dominated by the most abundant transcripts.



■ **Figure 3** Spearman correlation, Pearson correlation and  $R^2$  with respect to all transcripts in the reference, and AUROC for recalling ground truth unexpressed transcripts, with respect to added normally distributed noise with varying standard deviations. Plotted lines for Pearson correlation and  $R^2$  overlap.

Second, we complement measuring Spearman correlation of non-zero ground truth TPMs with reporting the area under receiver operating characteristic (AUROC) for recalling ground truth zeros based on estimated abundances. While the measurement of Spearman correlation on the truly expressed transcripts is robust to small changes in predicted abundance near zero, it fails to account for false positive predictions even if they are of non-trivial abundance. The complementary metric of the AUROC for recalling ground truth zeros complements that metric, since it is affected by false positive predictions.

### 3.5.3 Differential expression analysis on simulated data

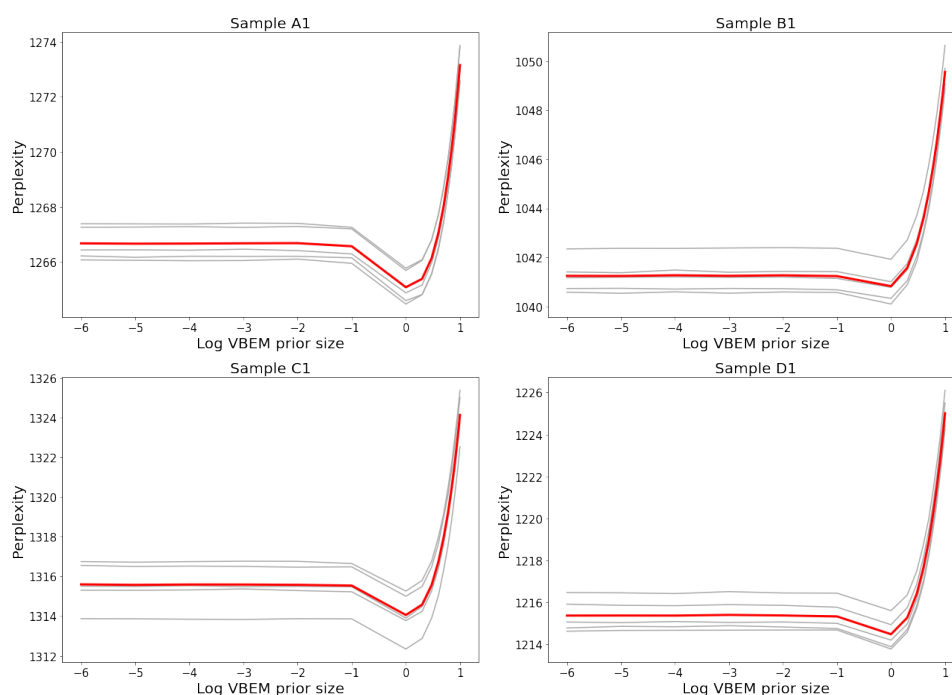
We perform transcript level differential expression analysis and analyze the recall of known differentially expressed transcripts in simulated lung tissue data (See 3.4.2). We perform differential expression analysis at the transcript level using `swish` [39] using 20 inferential replicates from `salmon`. We modified `salmon` to ensure that prior sizes supplied via the `--vbPrior` flag are propagated to the Gibbs sampling algorithm. We plot receiver operating characteristic (ROC) curves and report the mean AUROC for predicting differentially expressed transcripts over multiple folds. We assign  $P = 1$  to transcripts for which `swish` does not assign adjusted P-values.

### 3.6 Implementation

We implement smoothed perplexity in Rust and provide `snakemake` [21] workflows to (1) set up quantified-validate splits of read-sets for K-fold cross-validation, and (2) compute perplexities of `salmon` abundance estimates with respect to validation fragment sets at: <https://github.com/COMBINE-lab/perplexity>. Code to reproduce the experiments and figures for this work is available at <https://github.com/COMBINE-lab/perplexity-paper>.

## 4 Results

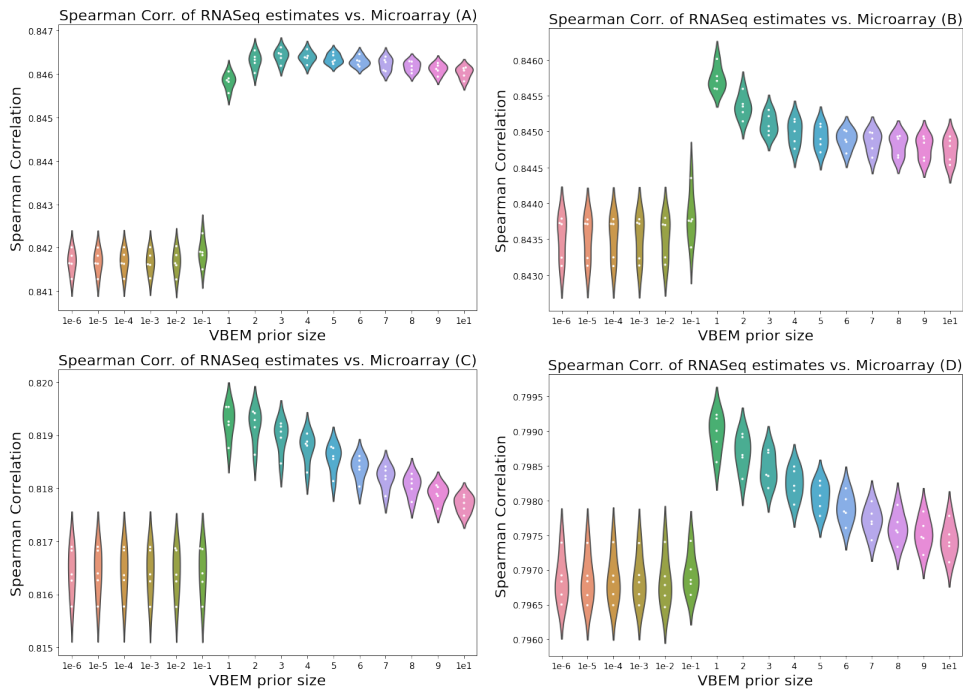
### 4.1 Lower perplexity implies more accurate abundance estimates in experimental SEQC data



**Figure 4** Perplexity plots for SEQC samples. Plots show perplexity versus VBEM reads-per-transcript prior size for SEQC samples – plots only for the first replicate of samples from conditions *A-D* are shown. Perplexity plots for other replicates are consistent within condition and are included in the Appendix. Mean perplexities across five folds are plotted in red, and perplexities for each fold are plotted in gray.

In experimental data from the Sequencing Quality Control (SEQC) project [35], we demonstrate that perplexity can be used to perform parameter selection and select the `salmon` VBEM prior size that leads to the most accurate transcript abundance estimates. We note that perplexity plots for replicates are similar within conditions *A-D*, and thus include only plots for the first replicate in each condition in the main text – plots for other samples are presented in the Appendix, Figure A1, for completeness.

Empirically, perplexity is well-behaved over all samples in the experimental data. As shown in Figure 4 and 5, plots of perplexity against VBEM prior size and Spearman correlation against VBEM prior size both display an empirically convex shape minimized at the same VBEM prior size. This suggests that minimizing perplexity is, at least, locally optimal with respect to the set of explored hyperparameters.



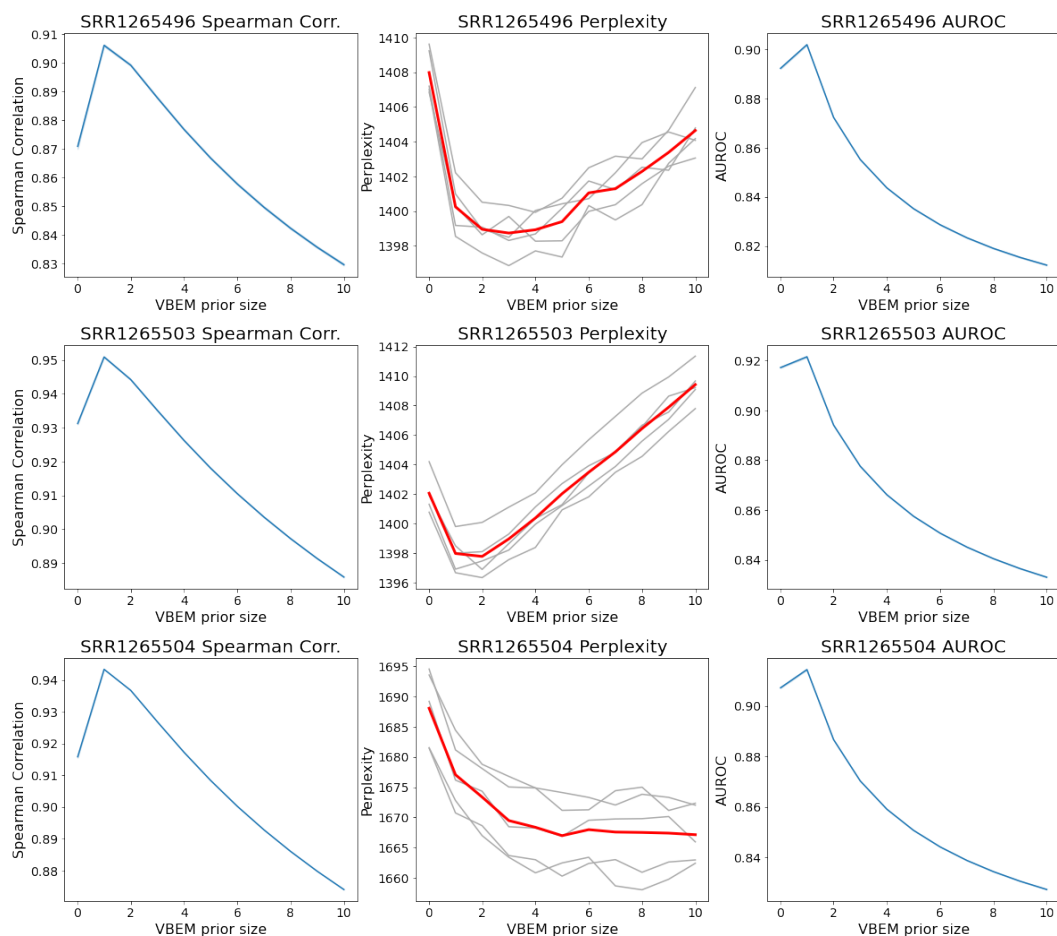
■ **Figure 5** Spearman correlation of abundance estimates at various VBEM reads-per-transcript prior sizes, versus parallel qPCR microarray gene-expression measurements conditions *A-D*. Each point in above plots indicate the mean correlation across replicates for a given fold.

Furthermore, for almost all samples, perplexity is minimized where correlation with qPCR measurements is maximized. For all replicates in conditions  $\{B, C, D\}$ , estimates that minimize perplexity with respect to held-out validation fragments achieve the best correlation with qPCR measured gene expression. For replicates in these conditions, abundances inferred using a prior size of 1 read-per-transcript resulted in estimates with the lowest perplexity. In replicates from condition *A*, estimates with lowest perplexity are significantly better than estimates at default hyperparameter settings (0.01 reads-per-transcript).

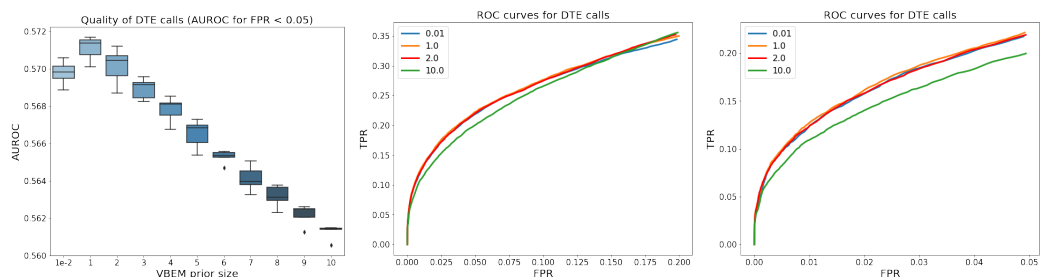
Perhaps surprisingly, both perplexity and correlation against qPCR measurements prefer a reads-per-transcript prior size that is larger than the 0.01 reads-per-transcript that is the current default for the `salmon` VBEM model. Selecting a larger per-transcript prior for transcript abundance estimation with `salmon` results in estimates that are more smooth and less sparse. For smoother abundance estimates, fewer validation time fragments are likely called impossible (compared to sparser estimates). In these cases, the number of impossible reads called by an estimate not only indicates inferential errors with regard to transcripts incorrectly called unexpressed, but likely suggests less accurate inferred abundances with respect to transcripts that are called expressed.

To the best of our knowledge, this experiment is the first to carry out both an effective and ubiquitously applicable quantitative strategy to perform model selection in the context of transcript abundance estimation on experimental data in the absence of ground truth.

## 4:14 Perplexity: Evaluating RNASeq in the Absence of Ground Truth



■ **Figure 6** Quality of transcript abundance estimates as a function of VBEM per-nucleotide prior size for samples SRR1265{496,503,504}. (Left column) Spearman Correlation with respect ground truth expressed transcripts. (Middle column) Perplexity of abundance estimates; perplexities per-fold indicated in gray and mean perplexities in red. (Right column) AUROC for retrieving ground truth unexpressed transcripts. Leftmost plotted points for all plots use default `salmon` VBEM prior size of 0.01 reads-per-transcript.



■ **Figure 7** Accuracy of differential expression analysis with respect to experiment-wide selection of VBEM per-nucleotide prior size. (Left) AUROC with respect to DTE calls at real FPRs up to 0.05. (Middle) ROC curve up to FPR = 0.20. (Right) ROC curve up to FPR = 0.05. To reduce visual clutter, only the ROC curves some representative VBEM prior size settings are plotted.

## 4.2 Perplexity versus ground truth, and differential expression analysis in simulated data

In simulated data, the relationship between perplexity and measurements against ground truth, though well-behaved, is admittedly less direct. In short, under the experimental framework we have chosen, minimizing perplexity does not always find the best performing estimates. Across all 10 samples, perplexity prefers abundance estimates that are smoother than estimates that are most accurate when compared to ground truth. For brevity, we include in the main text perplexity plots of three samples (SRR1265{496,503,504}) that are representative of three main modalities of perplexity plot behaviors (Figure. 6). For completeness, and refer the reader to the appendix for analogous plots for the seven remaining samples (Figures A2 and A3).

In all but one sample (SRR1265504), perplexity plots display an empirically convex shape with a local minima close to the optimal VBEM prior size (1 read-per-transcript). For example, for sample SRR1265503, perplexity is minimized at a VBEM prior setting of 2 reads-per-transcript, the second best performing hyperparameter setting with respect to Spearman correlation (Figure. 6; middle). And for sample SRR1265496, we can clearly see that perplexity prefers VBEM prior setting in a wide local minima ranging from 2 to 4 reads-per-transcript (Figure. 6; top). Sample SRR1265504 is the only sample for which a local minimal perplexity cannot be identified with respect to the range of hyperparameters scanned (Figure. 6; bottom). However, the perplexity plot for SRR1265504 displays a knee-like behavior which suggests that after a certain VBEM prior size, larger VBEM prior sizes are no longer preferred – which is consistent across all perplexity plots and comparisons to ground truth.

These observations in the simulated data could suggest that perplexity may be an imperfect tool, or perhaps that different characteristics and read depths between the experimental and simulated data signal the need for a data-dependent selection mechanism for the smoothing function used to evaluate perplexity. Nonetheless, these observations do offer several insights as to how perplexity ought to be used in practice, especially when careful (albeit qualitative) inspection of perplexity plots reveal inconsistent preferences for hyperparameters across similar samples experiment-wide. First, perplexities may prefer abundance estimations smoother than ideal. In particular, when perplexities between two VBEM prior settings are close, or when perplexities are roughly minimized for a range of values, one ought to select the model that provides the sparsest estimates. Second, our experiments suggest that an optimal hyperparameter setting for a set of samples can be selected experiment-wide and perplexity plots can be used as a rough guide to select said hyperparameter setting. For example, visual inspection of perplexity plots (Figures A2 and A3) experiment-wide show a knee-like behavior and rough local minima for perplexity beginning at a VBEM prior size of 2 reads-per-transcript – the second best hyperparameter setting.

Thus, we note that perplexity can be used to quantitatively screen for bad abundance estimates (or the hyperparameters that generate them). The significance of this observation may be overlooked at first. However, to our knowledge, perplexity is the only metric that can differentiate between a satisfactory and a much more inaccurate abundance estimate when ground truth is absent.

Given the above, we also analyze the accuracy of differential transcript expression (DTE) analysis of estimates with the same VBEM prior size experiment-wide. We report AUROC of DTE calls up to a nominally useful maximum false discovery rate (FDR) of 0.05 (Figure 7). Not surprisingly, AUROC of DTE calls mirror the shape of Spearman correlations of estimates

inferred from different VBEM prior sizes. Again, though minimizing perplexities does not exactly select the best estimates with regard to downstream DTE analysis, perplexity plots begin to exhibit plateaus or knee-like behaviors at VBEM prior size of 2 reads-per-transcript, the second best performing hyperparameter setting with regard to DTE (Figure 7).

## 5 Discussion

In this work, we derive the smoothed perplexity metric, which, to our knowledge, is the first metric that enables the evaluation of the quality of transcript abundance estimates in the absence of ground truth. Though we focus only on performing model selection with respect to one hyperparameter (the VBEM prior size) in `salmon`, model selection for other settings (e.g. choosing the number of bins for the range-factorized likelihood approximation, or selecting between VBEM and EM models and optimizations) are also certainly possible using perplexity.

In experimental data from the Sequencing Quality Control (SEQC) project [35], we show that the most accurate abundance estimates consistently have the lowest perplexity (lower is better) and demonstrate how quantitative model selection can be performed on input fragment sets directly and in the absence of ground truth. In simulated samples, we demonstrate a looser, but still useful, relationship between perplexity and measurements against ground truth. One possible explanation for the more erratic behavior and noisier perplexity plots for our simulated samples is due to these samples consisting of many fewer fragments than SEQC samples. On average, the simulated samples contain 17,410,732 fragments on average while the SEQC samples average 47,589,281 fragments.

Admittedly, the parameterization of the smoothing applied prior to input abundance estimates is somewhat unsatisfying. We do note, however, that at different settings of  $\beta$ , when a minima with regard to perplexity is observed in analyzed samples, the minima remains largely consistent – we demonstrate this for SEQC sample A1 in Figure A4. We plan to address the trade-offs and strategies for selecting smoothing strategies in future work.

Other directions for future work include utilizing perplexity or other metrics based on held-out likelihoods to not only select hyperparameters, but also to compare different abundance estimation models themselves. Furthermore, perplexity can also be adapted and applied to other problem settings in bioinformatics in which abundances are inferred from probabilistic models. For example, in metagenomics where model selection (i.e. choosing confidence cutoffs for taxa identification, or selecting candidate reference genomes) can have a large effect on abundance estimates [25].

In sum, this work demonstrates that evaluation of transcript abundance estimates in the absence of ground truth is possible, and presents a promising new direction in which estimated abundances are evaluated and validated directly on input fragment sets.

---

## References

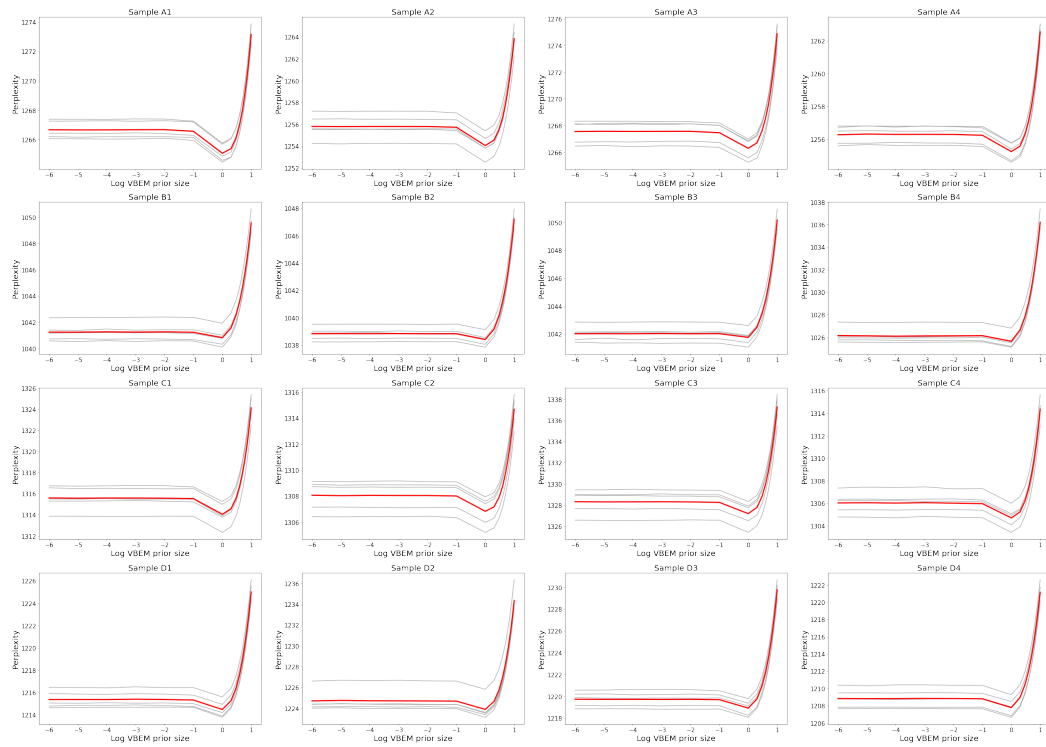
- 1 Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.
- 2 Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, et al. The External RNA Controls Consortium: a progress report. *Nature Methods*, 2(10):731–734, 2005. doi:10.1038/nmeth1005-731.
- 3 Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2016.
- 4 David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, 2003.



- 5 Elena Bushmanova, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9), September 2019. giz100. doi:10.1093/gigascience/giz100.
- 6 Scott C. Clark, Rob Egan, Peter I. Frazier, and Zhong Wang. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29(4):435–443, 2013. doi:10.1093/bioinformatics/bts723.
- 7 Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191, 2009. doi:10.1038/nprot.2009.97.
- 8 Alyssa C. Frazee, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, April 2015. doi:10.1093/bioinformatics/btv272.
- 9 Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.
- 10 Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011. doi:10.1038/nbt.1883.
- 11 James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, and Magnus Rattray. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, 31(24):3881–3889, August 2015. doi:10.1093/bioinformatics/btv483.
- 12 F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976. doi:10.1109/proc.1976.10159.
- 13 Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-seq. *Bioinformatics*, 25(8):1026–1032, 2009.
- 14 Daniel C. Jones, Kavitha T. Kuppasamy, Nathan J. Palpant, Xinxia Peng, Charles E. Murry, Hannele Ruohola-Baker, and Walter L. Ruzzo. Isolator: accurate and stable analysis of isoform-level expression in rna-seq experiments. *bioRxiv*, 2016. doi:10.1101/088765.
- 15 Daniel C. Jones and Walter L. Ruzzo. Polee: RNA-Seq analysis using approximate likelihood. *bioRxiv*, 2020. doi:10.1101/2020.09.09.290411.
- 16 Woo Jin Kim, Jae Hyun Lim, Jae Seung Lee, Sang-Do Lee, Ju Han Kim, and Yeon-Mok Oh. Comprehensive analysis of transcriptome sequencing data in the lung tissues of copd subjects. *International Journal of Genomics*, 2015:206937, March 2015. doi:10.1155/2015/206937.
- 17 Bo Li and Colin N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, August 2011. doi:10.1186/1471-2105-12-323.
- 18 Bo Li, Nathanael Fillmore, Yongsheng Bai, Mike Collins, James A Thomson, Ron Stewart, and Colin N Dewey. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15(12):553, 2014. doi:10.1186/s13059-014-0553-5.
- 19 Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.
- 20 Peng Liu, Rajendran Sanalkumar, Emery H Bresnick, Sündüz Keleş, and Colin N Dewey. Integrative analysis with chip-seq advances the limits of transcript quantification from rna-seq. *Genome research*, 26(8):1124–1133, 2016.
- 21 Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, et al. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, 2021. doi:10.12688/f1000research.29032.1.
- 22 Naoki Nariai, Osamu Hirose, Kaname Kojima, and Masao Nagasaki. TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. *Bioinformatics*, 29(18):2292–2299, July 2013. doi:10.1093/bioinformatics/btt381.

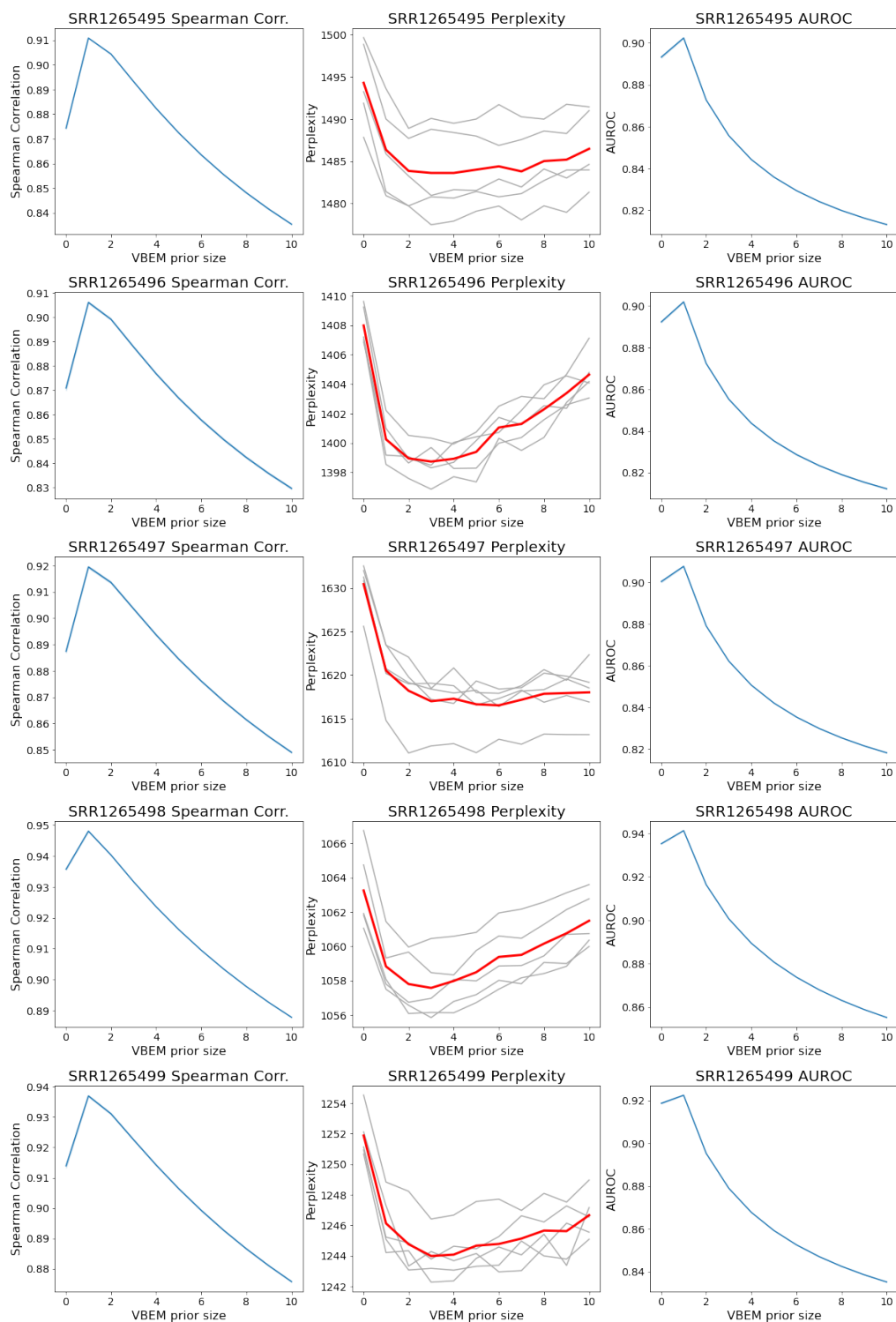
- 23 Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yosuke Kawai, and Masao Nagasaki. A bayesian approach for estimating allele-specific expression from RNA-seq data with diploid genomes. In *BMC genomics*, volume 17(1), pages 7–17. BioMed Central, 2016.
- 24 Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki. Tigar2: sensitive and accurate estimation of transcript isoform expression with longer RNA-seq reads. *BMC genomics*, 15(10):1–9, 2014.
- 25 Daniel J. Nasko, Sergey Koren, Adam M. Phillippy, and Todd J. Treangen. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biology*, 2018. doi:10.1186/s13059-018-1554-6.
- 26 Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, April 2017. doi:10.1038/nmeth.4197.
- 27 Atif Rahman and Lior Pachter. CGAL: computing genome assembly likelihoods. *Genome Biology*, 14(1):R8, 2013. doi:10.1186/gb-2013-14-1-r8.
- 28 Johannes Rainer. *EnsDb.Hsapiens.v86: Ensembl based annotation package*, 2017. R package version 2.99.0.
- 29 Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi:10.1016/0377-0427(87)90125-7.
- 30 Migun Shakya, Chien-Chi Lo, and Patrick S. G. Chain. Advances and challenges in metatranscriptomic analysis. *Frontiers in Genetics*, 10:904, 2019. doi:10.3389/fgene.2019.00904.
- 31 Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006. doi:10.1038/nbt1239.
- 32 Richard Smith-Unna, Chris Boursnell, Rob Patro, Julian M. Hibberd, and Steven Kelly. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26(8):1134–1144, 2016. doi:10.1101/gr.196469.115.
- 33 Charlotte Sonesson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015. doi:10.12688/f1000research.7563.1.
- 34 Avi Srivastava, Laraib Malik, Hirak Sarkar, and Rob Patro. A Bayesian framework for inter-cellular information sharing improves dscRNA-seq quantification. *Bioinformatics*, 36(Supplement\_1):i292–i299, 2020.
- 35 Zhenqiang Su, Paweł P Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, 2014. doi:10.1038/nbt.2957.
- 36 Ernest Turro, Shu-Yi Su, Ângela Gonçalves, Lachlan JM Coin, Sylvia Richardson, and Alex Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome biology*, 12(2):1–15, 2011.
- 37 Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, et al. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, November 2019. doi:10.1093/nar/gkz966.
- 38 Mohsen Zakeri, Avi Srivastava, Fatemeh Almodaresi, and Rob Patro. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics*, 33(14):i142–i151, July 2017. doi:10.1093/bioinformatics/btx262.
- 39 Anqi Zhu, Avi Srivastava, Joseph G Ibrahim, Rob Patro, and Michael I Love. Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Research*, 47(18):e105–e105, 2019. doi:10.1093/nar/gkz622.

**A** Appendix

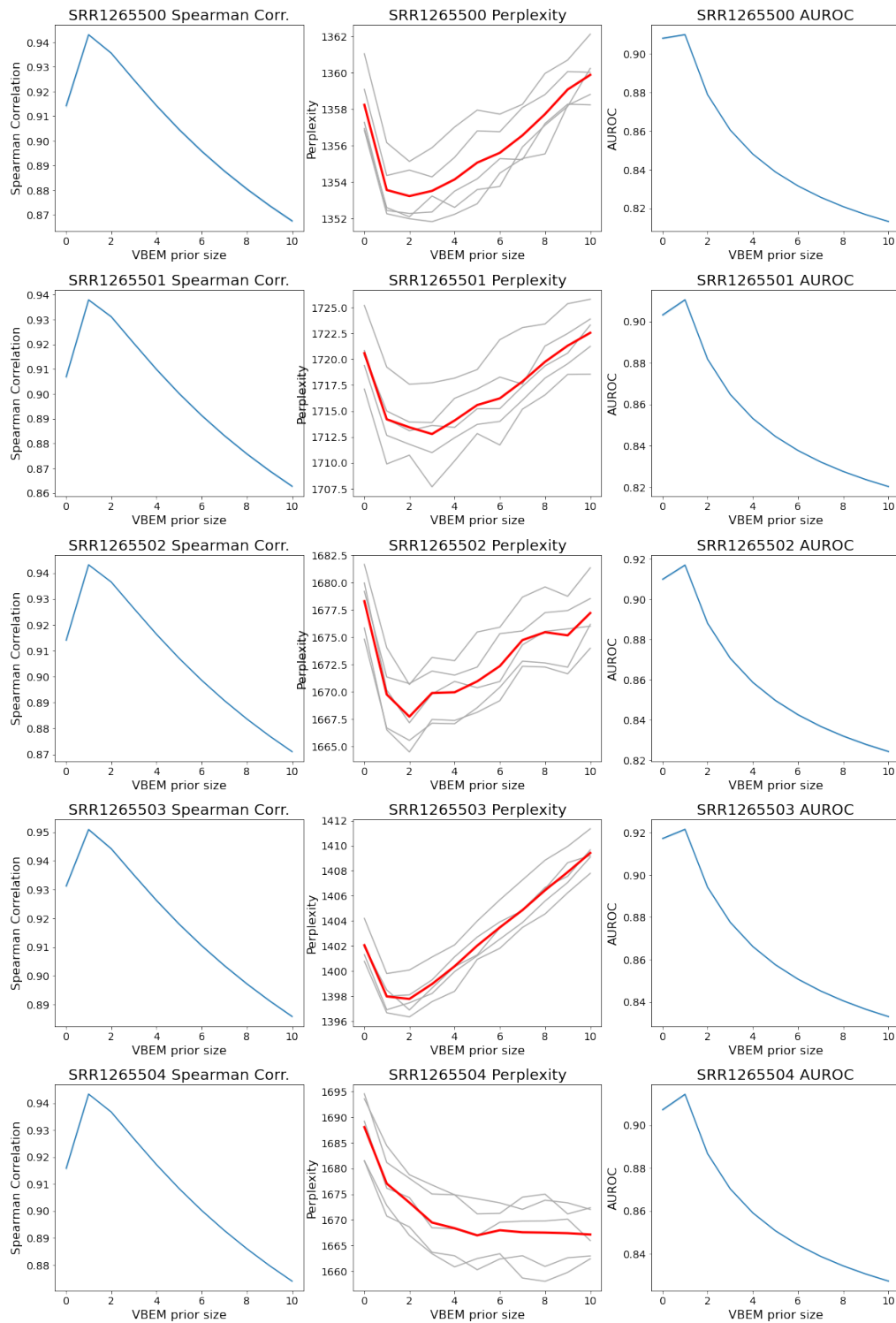


■ **Figure A1** Perplexity plots. Plots show perplexity versus VBEM reads-per-transcript prior size for SEQC samples. Mean perplexities across five folds are plotted in red, and gray perplexities for each fold are plotted in gray.

## 4:20 Perplexity: Evaluating RNASeq in the Absence of Ground Truth

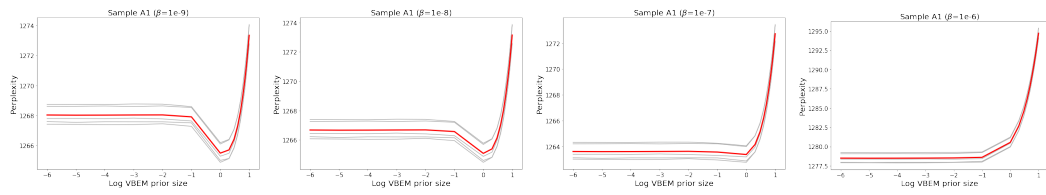


■ **Figure A2** Quality of transcript abundance estimates as a function of VBEM per-nucleotide prior size for samples SRR1265{495-499}. (Left column) Spearman Correlation with respect ground truth expressed transcripts. (Middle column) Perplexity of abundance estimates; perplexities per-fold indicated in gray and mean perplexities in red. (Right column) AUROC for retrieving ground truth unexpressed transcripts. Leftmost plotted points for all plots use default `salmon` VBEM prior size of 0.01 reads-per-transcript.



■ **Figure A3** Quality of transcript abundance estimates as a function of VBEM per-nucleotide prior size for samples SRR1265{500-504}.

## 4:22 Perplexity: Evaluating RNASeq in the Absence of Ground Truth



■ **Figure A4** Perplexity plots for SEQC sample A1 at different smoothing parameter settings. Plots show perplexity versus VBEM reads-per-transcript prior size for SEQC samples. Mean perplexities across five folds are plotted in red, and gray perplexities for each fold are plotted in gray.