Bioinformatics, 33, 2017, i142–i151
doi: 10.1093/bioinformatics/btx262
ISMB/ECCB 2017

OXFORD

Improved data-driven likelihood factorizations for transcript abundance estimation

Mohsen Zakeri, Avi Srivastava, Fatemeh Almodaresi and Rob Patro*

Department of Computer Science, Stony Brook University, Stony Brook, NY 11790, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Many methods for transcript-level abundance estimation reduce the computational burden associated with the iterative algorithms they use by adopting an approximate *factorization* of the likelihood function they optimize. This leads to considerably faster convergence of the optimization procedure, since each round of e.g. the EM algorithm, can execute much more quickly. However, these approximate factorizations of the likelihood function simplify calculations at the expense of discarding certain information that can be useful for accurate transcript abundance estimation.

Results: We demonstrate that model simplifications (i.e. factorizations of the likelihood function) adopted by certain abundance estimation methods can lead to a diminished ability to accurately estimate the abundances of highly related transcripts. In particular, considering factorizations based on transcript-fragment compatibility alone can result in a loss of accuracy compared to the per-fragment, unsimplified model. However, we show that such shortcomings are not an inherent limitation of approximately factorizing the underlying likelihood function. By considering the appropriate conditional fragment probabilities, and adopting improved, data-driven factorizations of this likelihood, we demonstrate that such approaches can achieve accuracy nearly indistinguishable from methods that consider the complete (i.e. per-fragment) likelihood, while retaining the computational efficiently of the compatibility-based factorizations.

Availability and implementation: Our data-driven factorizations are incorporated into a branch of the *Salmon* transcript quantification tool: https://github.com/COMBINE-lab/salmon/tree/factorizations.

Contact: rob.patro@cs.stonybrook.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Shortly after the RNA-seq assay became popular as a tool for transcriptome profiling and quantification, the computational community began developing principled inference methodologies to allow accurate transcript-level quantification in the presence of multi-mapping reads. Tools such as *Cufflinks* (Trapnell *et al.*, 2010), *RSEM* (Li *et al.*, 2010), *mmseq* (Turro *et al.*, 2011) and *IsoEM* (Nicolae *et al.*, 2011) provided statistical models by which transcript-level abundance estimates could be inferred. These methodologies principally rely on maximum likelihood estimation to infer the transcript abundances that would be most likely given the observed data (i.e. the alignments of the sequenced fragments to the underlying genome or transcriptome). Bayesian methodologies such as *BitSeq* (Glaus *et al.*, 2012) and *Tigar* (Nariai *et al.*, 2013) were also developed and adopt different inferential approaches varying from fully Bayesian approaches like collapsed Gibbs sampling (Glaus *et al.*, 2012) to approximate inference approaches like variational Bayesian optimization (Hensman *et al.*, 2015; Nariai *et al.*, 2013, 2014).

These methods vary widely in their details, though adopt a similar generative model of the underlying RNA-seq experiment; one which is well-represented by the generative model of *RSEM* (Li *et al.*, 2010; Li and Dewey, 2011). In this paper, we shall refer to this as the *full model*. It is a generative model of an RNA-seq experiment that considers the likelihood of observing a collection of alignments as dependent upon the parameters of interest (i.e. the transcript abundances), as well as the details of each alignment of a sequenced fragment to the reference transcriptome. In this way, the full model provides very high fidelity, and is capable of incorporating a tremendous amount of information into the inference procedure (e.g. the

implied fragment length under each alignment, details about the alignment and the fragment's quality values, the probability of different start positions for the sampled fragment, etc.).

Unfortunately, however, this means that straightforward inference procedures that adopt this full model scale in the number of considered alignments *per-iteration*. For example, a 30 million fragment RNA-seq experiment may produce 100 million fragment alignments, all of which are considered by the inference procedure in each of its (typically) hundreds to thousands of iterations. This approach, then, poses two problems. First, inference is typically slow since each iteration must consider a large number of independent probabilities. Second, so as to prevent the inference algorithm from becoming even slower, these per-alignment probabilities are typically retained in memory, which can lead memory requirements to scale linearly with the number of alignments. One approach to mitigate the cost associated with optimizing the *full model* is to alter the actual inference algorithm that is used. For example, *eXpress* (Roberts and Pachter, 2013) uses an online-EM algorithm, rather than a batch-EM algorithm (by default), to infer transcript abundances. This eliminates the need to cache alignments in memory for efficiency, resulting in constant memory usage. However, a single pass over the data is not always sufficient to achieve the same accuracy as methods that run batch algorithms to convergence.

One of the more popular approaches for reducing the computational burden and speeding up the inference procedure is to form an approximate factorization of the likelihood function (see Section 2.1). For example, *mmseq* introduced a notion of fragment equivalence classes, which treats as equivalent any fragments that align to exactly the same set of transcripts. This leads to a likelihood function in which the counts of fragments compatible with subsets of transcripts serve as sufficient statistics. The likelihood defined over these counts is typically orders of magnitude faster to evaluate, but it can discard certain fragment-level information encoded in the alignments. Distinct but related notions of equivalence classes were also introduced by Salzman *et al.* (2011) and Nicolae *et al.* (2011).

Because of the computational economy of this approximate factorization, it (or similar variants) were later adopted by new lightweight approaches for transcript quantification like *Sailfish* (Patro *et al.*, 2014; Srivastava *et al.*, 2016) and *kallisto* (Bray *et al.*, 2016). By coupling a very fast inference approach with techniques that removed the requirement of computing traditional alignments for each sequenced fragment, such approaches reduced the time required to obtain transcript-level quantification estimates by orders of magnitude over existing approaches. These lightweight methods have proven an important and popular development. Recently, Patro *et al.* (2017) introduced a new lightweight approach, *Salmon*, that uses a 'dual-phase' inference algorithm, which combines an online stochastic inference method with an efficient offline inference algorithm. While adopting a similar approximate factorization as *mmseq*, *Sailfish* and *kallisto*, *Salmon* also maintains aggregate (i.e. average) weights per equivalence class that allow retaining some information about fragment-level probabilities during the offline inference algorithm. However, this information is restricted to a single scalar value per transcript-equivalence class pair, and so is necessarily limited in its ability to represent the full model with complete fidelity.

In this paper, we argue that the dual-phase algorithm introduced by *Salmon* allows one to derive a data-driven approximate factorization of the full model likelihood function. The online phase of the algorithm assesses each individual fragment probability, and uses this information to build a highly reduced but accurate proxy for the *full model* likelihood that can be efficiently optimized during the offline phase. While only slightly increasing the per-iteration cost of the underlying inference algorithm, this data-driven factorization can represent the fragment-level likelihood function with much higher fidelity. In fact, we demonstrate that a data-driven likelihood factorization can produce transcript-level abundance estimates that display essentially no loss in accuracy compared to what is obtained under the full model. Thus, such a factorization is preferable to the more common compatibility-based approximate factorization, since it can provide a substantial improvement in accuracy while introducing only a small increase in the computational burden. We note that we focus in this paper on how to factorize the likelihood function, and not, specifically, the algorithm by which this function is best optimized. Thus, we expect the approaches we introduce here to easily translate to other likelihoods or optimization approaches; e.g. to variational Bayesian optimizations (Nariai *et al.*, 2013), or natural gradient-based optimization algorithms (Hensman *et al.*, 2015).

2 Approach

2.1 The likelihood function and its factorizations

We begin by considering the basic generative model laid out by Li *et al.* (2010). We consider a transcriptome T to consist of a set of M transcripts, $t_1, t_2, \ldots, t_M$. In a given sample, there are c_i copies of the ith transcript. Further, we can assign to each transcript its length, such that the length of t_i is given by ℓ_i. The generative model of an RNA-seq experiment states that the expected number of fragments sequenced from each transcript type t_i is proportional to the total number of sequencable nucleotides that it constitutes in the underlying mixture—that is we expect that $\alpha_i \propto \eta_i = \frac{c_i \cdot \ell_i}{\sum_j c_j \cdot \ell_j}$—where α_i is the number of fragments drawn from transcripts of type t_i. Assuming that each fragment is drawn independently, the likelihood of a collection $\mathcal{F}$ of fragments can be written as:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{F}) = \prod_{f_j \in \mathcal{F}} \sum_{i=1}^{M} \Pr(t_i|\boldsymbol{\theta}) \Pr(f_j|t_i), \tag{1}$$

where $\boldsymbol{\theta}$ denotes the parameters of the model, which are the underlying transcript abundances. We note that, throughout this manuscript, we use the term 'fragment' as a generic term which is represented by a single read (in single-end protocols) and a read pair (in paired-end protocols). The methods we propose in Section 3 work only in terms of the conditional fragment probabilities, and so are equally applicable in both single and paired-end protocols, though the definition of these conditional probabilities will be protocol dependent.

The primary quantity of interest, with respect to the factorizations being proposed in this paper, are the $\Pr(f_j|t_i)$ terms—that is, the conditional probability of drawing a particular fragment f_j, given transcript t_i. This term encodes, given parameters of the model and experiment, how likely it is to observe a specific fragment f_j arise from transcript t_i. Many terms can be included in such a conditional probability, some common terms include:

$$\Pr(d_j|f_j, t_i) = \frac{\Pr_D(d_j)}{\sum_{k=1}^{\ell_i} \Pr_D(k)}, \tag{2}$$

the probability of observing a mapping of implied length d_j for f_j given that it derives from t_i, where $\Pr_D(k)$ is the probability of

observing a fragment of length k under the empirical fragment length distribution D;

$$\Pr(p_j|d_j, f_j, t_i) = \frac{1}{\ell_i - d_j + 1}, \quad (3)$$

the probability of a observing a mapping starting at position p_j for fragment f_j given that it has implied length d_j and is derived from t_i;

$$\Pr(o_j|f_j, t_i) = \begin{cases} \begin{cases} 0.5 & \text{if unstranded} \\ \begin{cases} 1.0 & \text{if compatible orientation} \\ \varepsilon & \text{if incompatible orientation} \end{cases} & \text{if strand-specific} \end{cases} \end{cases}, \quad (4)$$

the probability of observing a mapping with a specific orientation o_j (i.e. forward or antisense) with respect to the underlying transcript for f_j, given t_i, ε (a user-defined constant), and knowledge of the underlying protocol, and

$$\Pr(a_j|f_j, o_j, d_j, p_j, t_i), \quad (5)$$

the probability of observing the particular alignment (e.g. CIGAR string) a_j for f_j given it is sampled from transcript t_i, has orientation o_j, implied length d_j and starts at position p_j—such a probability is calculated from a model of alignments, like those presented in (Li *et al.*, 2010; Patro *et al.*, 2017; Roberts and Pachter, 2013).

In fact, one can conceive of many such general models of 'fragment-transcript agreement' (Patro *et al.*, 2017). The framework we propose in Section 3 can naturally account for such conditional probabilities that one might consider as part of $\Pr(f_j|t_i)$. However, in this manuscript, we consider that $\Pr(f_j|t_i)$ is simply the product of the conditional probabilities defined in Equations (2) to (5), appropriately normalized.

2.2 Equivalence classes and approximate likelihood factorizations

Here, we describe the most common definition of fragment equivalence classes, and explain how they are used to derive an approximate factorization of the likelihood function, we adopt a notation similar to Patro *et al.* (2017).

Let $A(\mathcal{T}, f_j)$ be the set of all alignments of fragment f_j to the transcriptome $\mathcal{T}$ and let $\Omega(f_j) = \{\langle i, t_i \rangle | t_i \in A(\mathcal{T}, f_j)\}$ be the tuple of transcripts to which f_j maps—considering the t_i are ordered by their index i. Fragment equivalence classes are defined in terms of the equivalence relation $\sim$, such that $f_m \sim f_n$ if and only if $\Omega(f_m) = \Omega(f_n)$. Thus, fragment equivalence classes consider as equivalent (for the purposes of inference), sequenced fragments that align to the same set of transcripts. We will refer to $\Omega(f_j)$ as the *label* of f_j for all $f_j \in \mathcal{F}^q$, where $\mathcal{F}^q$ is the equivalence class to which f_j belongs. We will also refer to $\Omega(\mathcal{F}^q) = \Omega(f_j), \forall f_j \in \mathcal{F}^q$ as the label of f_j's equivalence class. Finally, it will be convenient to define the total size of each such equivalence class as $N^q = |\mathcal{F}^q|$, which is the total number of equivalent fragments in the class $\mathcal{F}^q$.

Now, we can write the equivalence class-based approximation to the likelihood function as:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{F}) \approx \prod_{\mathcal{F}^q \in \mathcal{C}} \left(\sum_{\langle i, t_i \rangle \in \Omega(\mathcal{F}^q)} \Pr(t_i|\boldsymbol{\theta}) \cdot \Pr(f|\mathcal{F}^q, t_i) \right)^{N^q}, \quad (6)$$

where $\mathcal{C}$ is the set of all equivalence classes, and $\Pr(f|\mathcal{F}^q, t_i)$ is the probability of generating a fragment f given that it comes from equivalence class $\mathcal{F}^q$ and transcript t_i. The key to the efficiency of likelihood evaluation (or optimization) under this factorization, is that the probability $\Pr(f|\mathcal{F}^q, t_i)$ is assumed to be identical for each of the N^q fragments in each equivalence class $\mathcal{F}^q$—hence, we do not subscript f in Equation (6). This allows one to replace the product over all fragments f_j in Equation (1) with a product over all equivalence classes in Equation (6). The approximation, of course, stems from the fact that, under the full model, a fragment f_j may have a probability $\Pr(f_j|t_i)$ that is arbitrarily different from $\Pr(f|\mathcal{F}^q, t_i)$. Moreover, the most common approximations, like those adopted in *mmseq*, *Sailfish* and *kallisto* consider this probability to be fixed and essentially independent of any fragment-level information (e.g. it is set to one divided by the effective length of t_i).

2.3 What approximate factorizations elide

Figure 1 provides an illustrative example why considering conditional fragment probabilities can be important. Consider a multi-isoform gene, and a single fragment f_j, which aligns equally well (i.e. the sequence of both ends of the fragment match the sequence of the underlying transcripts equally well) to isoforms A and B of this gene. If we consider only transcript-fragment compatibility, then both of the alignments illustrated in Figure 1 are delineated only in that isoform A has fewer potential start locations. However, considering the implied length of this fragment, given the expected insert size distribution of the experiment (either provided as input to the model, or inferred from the collection of previously aligned fragments), can provide strong evidence that one or the other of these isoforms was more likely to have generated f_j. For example, were the mean of the fragment length distribution 250, then we would expect isoform A to be much more likely to have generated f_j. Conversely, were the mean of the fragment length distribution 400, then we would expect that, in fact, isoform B might have been more likely to generate this fragment. Standard (i.e. compatibility-based) approximate factorizations of the full likelihood function into equivalence classes discard (or collapse) this fragment-level information. For example, compatibility-only factorizations of the likelihood into equivalence classes simply treat $\Pr(d_j|f_j, t_i)$ as equal for all transcripts in the equivalence class to which fragment f_j belongs. The factorization adopted by *Salmon* attempts to maintain slightly more information by computing these conditional probabilities and averaging them; maintaining a single extra scalar per transcript-equivalence class pair, that represents the conditional probability that any fragment coming from a particular equivalence class would derive from a particular transcript. Though this maintains some extra information, it is not always

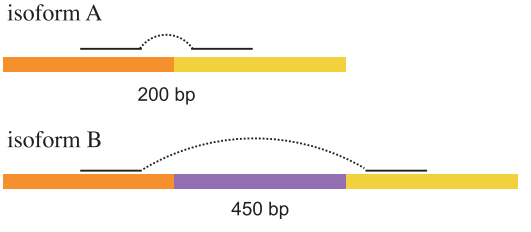

Fig. 1. A fragment multimapping between two different isoforms (**A,B**) of a gene. Depending on the parameters of the fragment length distribution of the underlying sample, either multi mapping locus could be more probable *a priori*. Under the approximate likelihood factorization that considers only compatibility-based equivalence classes, such information is necessarily hidden from the resulting inference algorithm. We note that, of course, such multi-mapping can also happen between different genes (e.g. paralogs)

enough to faithfully approximate the full-model likelihood function.

Below, we describe a data-driven approach that allows for a much more faithful representation of the *full model* likelihood function, while still greatly reducing the amount of information that must be maintained for inference. A broad overview of how these factorizations relate to each other is given in Figure 2, and the specific factorizations are described in more detail below.

3 Materials and Methods

As illustrated in Figure 2 and described above, the approximations that rely on *compatibility-based* factorizations can discard information that may be useful for correct transcript abundance estimation. Specifically, such notions of equivalence classes sacrifice per-fragment information encoded in the conditional probabilities $\Pr(f_j|t_i)$. We propose here alternative notions of equivalence classes that take into account both the transcripts with which a fragment is compatible, as well as the vector of conditional probabilities that encodes how likely the fragment is to have been sequenced from each such transcript. That is, these factorizations account both for the set of transcripts $t_1, \ldots, t_k$ to which a fragment f_j maps, as well as the conditional probabilities $\Pr(f_j|t_1), \ldots, \Pr(f_j|t_k)$ that f_j was sampled from each of these transcripts. Our approach is agnostic to how $\Pr(f_j|t_i)$ is computed, but, as stated in Section 2.1, we consider here each conditional probability to be the product of Equations (2) to (5), appropriately normalized.We accomplish this by defining new equivalence relations over fragments that consider and summarize these conditional probabilities in a data-driven manner.

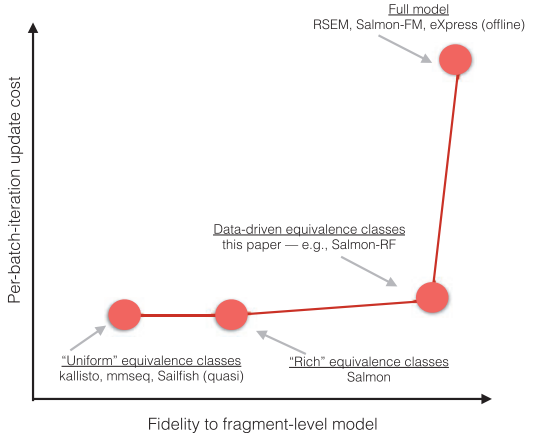

Fig. 2. There is a conceptual tradeoff between the computational efficiency of an inference technique, and the fidelity with which it models the full, fragment-level likelihood function. *kallisto*, *Sailfish* (using quasi-mapping (Srivastava *et al.*, 2016)) and *mmseq* simply consider the compatibility of fragments with transcripts, and thereby discard the conditional fragment-level probabilities completely. *Salmon* collapses the fragment-level conditional probabilities to a single scalar (their average value) per-equivalence class; this recovers some of the fidelity lost in the other approaches, but can still discard useful fragment-level information. Approaches that consider each fragment independently in each round of the optimization algorithm (e.g. *RSEM* and *Salmon-FM* and *eXpress* (offline)) sacrifice no fidelity, but each iteration scales with the total number of aligned/mapped fragments. Our proposed data-driven clustering approach (*Salmon-RF*) captures most of the important fragment-level probabilities of the *full model*, while retaining an update time very similar to *Salmon* 's standard model in its offline rounds. The online rounds of *Salmon* and *eXpress* are not directly comparable to the batch rounds considered in this figure (they update the parameters more frequently), but they do consider the conditional probability of each fragment individually

As one divides each equivalence class into smaller sub-classes of fragments, the factorized likelihood approaches the likelihood (and hence fidelity) of the *full model*. Conversely, as the number of equivalence classes increases so does the complexity of evaluating and optimizing the likelihood.

Here, we introduce two different factorization methods that refine the *compatibility-based* notion of equivalence classes. These approaches are a refinement in the strict sense that each sub-cluster of fragments that fall within the newly defined equivalence classes align to the same set of transcripts as all other fragments in the original, *compatibility-based* definition of the equivalence class. However, in these factorizations, the conditional fragment probabilities (with respect to the set of transcripts) tend to exhibit smaller distance to mean; i.e. the approximate weight used to summarize the conditional probability of all fragments within these refined equivalence classes is much closer to the individual conditional probabilities of all the fragments placed in the class. Subsequently, this leads to a more accurate approximation of the likelihood function. Moreover, we find that only a small number of such refined equivalence classes is required to approximate the full likelihood very closely, meaning that the computational complexity of evaluating and optimizing the likelihood function is very close to what is required when considering the original *compatibility-based* equivalence class factorization (see Results).

3.1 Rank-based factorization

We call the first factorization method that we consider to refine the notion of equivalence classes the 'rank-based factorization'. We consider all transcripts to which a fragment aligns, and sort the transcripts based on the conditional probability values of the fragment given each transcript. Then, the equivalence class for a fragment is determined by the set of transcripts to which it maps, *and* the rank-order of the conditional probabilities for this fragment given those transcripts. For instance, consider 1000 fragments which all align to the transcripts t_1 and t_2, where 250 of these fragments align to t_1 with a higher conditional probability than that with which they align to t_2 (and vice-versa for the rest). In this case, the rank-based equivalence relation will induce two equivalence classes (whereas the *compatibility-based* relation would have induced 1), the first 250 fragments will become members of one equivalence class with label $\{\langle 1, t_1\rangle, \langle 2, t_2\rangle\}$ and the rest will be assigned to another equivalence class with the label $\{\langle 1, t_2\rangle, \langle 2, t_1\rangle\}$. As with the original notion of *rich* equivalence classes in Salmon (Patro *et al.*, 2017), a single scalar value per transcript is saved in each equivalence class, which is the mean of all conditional probabilities of the fragments given each transcript. Of course, in this factorization, the total number of equivalence classes is typically larger than the number of *compatibility-based* equivalence classes. Formally, we define the rank-based equivalence relation $\sim_<$ as follows: let $r(f, \{\langle i_1, t_{i_1}\rangle, \langle i_2, t_{i_2}\rangle, \ldots, \langle i_j, t_{i_j}\rangle\})$ be a function that returns a permutation σ of $(t_{i_1}, t_{i_2}, \ldots, t_{i_j})$ such that $\Pr(f|t_{\sigma_1}) \leq \Pr(f|t_{\sigma_2}) \leq \ldots \leq \Pr(f|t_{\sigma_j})$, with ties broken arbitrarily in favor of the transcript having the smaller index. We define two fragments f_m and f_n to be equivalent ($f_m \sim_< f_n$) if and only if $\Omega(f_m) = \Omega(f_n)$ and $r(f_m, \Omega(f_m)) = r(f_n, \Omega(f_n))$.

3.2 Range-based factorization

We consider a second factorization approach that we call 'range factorization' (*Salmon-RF*). In this approach, we seek equivalence classes that have fragments which both align to the same set of transcripts *and* which have similar conditional probabilities with respect

to these transcripts. To motivate this approach, consider, first, the case of two fragments that have exactly the same conditional probabilities for the same set of transcripts, then one can safely group them together without any loss of accuracy with respect to the original likelihood function. In fact, this is the equivalence relation proposed by Nicolae *et al.* (2011). However, this particular factorization can have a negative impact on performance since most of the time probabilities of fragments are not exactly proportional. Hence, this can lead to a model similar to the full model that considers all fragment-transcript likelihood values. However, we can compromise the 'exact' proportionality of probabilities for the sake of performance. Instead of clustering fragments that have exactly proportional probabilities, we place fragments with the same conditional probability 'range' into the same equivalence class. We first divide the valid range of probabilities $[0, 1]$ into k bins, and then consider two conditional probabilities equal if their values are in the same bin. Two fragments are considered equivalent under this definition, denoted $\sim_r$, if they fall into the same set of bins with respect to all transcripts to which they align. Formally, let $b_k(f, \{\langle i_1, t_{i_1}\rangle, \langle i_2, t_{i_2}\rangle, \ldots, \langle i_j, t_{i_j}\rangle\})$ be a function that returns a vector of bin values (one for each transcript, and each between 0 and $k-1$). We define two fragments f_m and f_n to be equivalent ($f_m \sim_r f_n$) if and only if $\Omega(f_m) = \Omega(f_n)$ and $b_k(f_m, \Omega(f_m)) = b_k(f_n, \Omega(f_n))$.

We can tune the parameter k to tradeoff of the number of such equivalence classes versus the accuracy they provide. As k approaches infinity (or, rather, machine precision), the fidelity provided by this factorization approaches that of the full model, because all fragments will end up in either single-member equivalence classes, or in equivalence classes of fragments having conditional probabilities exactly proportional to theirs. On the other hand, as k gets smaller, the number of clusters gets closer to a small constant times the number of *compatibility-based* equivalence classes, but each cluster consists of fragments with the wider range of conditional probabilities. In this approach, we do not simply replace each conditional probability with the center of the bin into which it falls. Rather, for each bin, we record the sum and a total number of conditional probabilities stored in this bin. After processing all fragments, the centroid of each bin is computed and used as the representative conditional probability for this bin. This model is a natural extension of the rich equivalence class model used in *Salmon*, and the models coincide when $k = 1$. Throughout this paper, range-based equivalence classes have a number of bins equal to $4 + \lceil\sqrt{|\Omega(\mathcal{F}^q)|}\rceil$.

Figure 3 provides a good example of this factorization and its impact on the average of conditional probabilities for each transcript. There are 225 fragments that all are aligned to the two transcripts in this equivalence class. Each dot represents a fragment with its x value equal to $\Pr(f|t_1)$ and y value equal to $\Pr(f|t_2)$. c_{all} shows the average value of conditional probabilities of all fragments for transcript t_1 and t_2. As can be observed, the deviation of c_{all} from many of the conditional probabilities is large since the conditional probabilities are widely distributed over the range from zero to one. However, when we divide the range into three bins and then separate fragments based on the bin into which their conditional probabilities fall, we obtain three clusters containing fragments whose within-cluster conditional probabilities fall into much smaller ranges. So, in this case, all fragments that have the same bin for their conditional probability given t_1 and their conditional probability given t_2 end up in the same cluster. Lines show the borders of each bin and colored circles show the centroids used to represent the conditional probabilities in each bin. In this case,

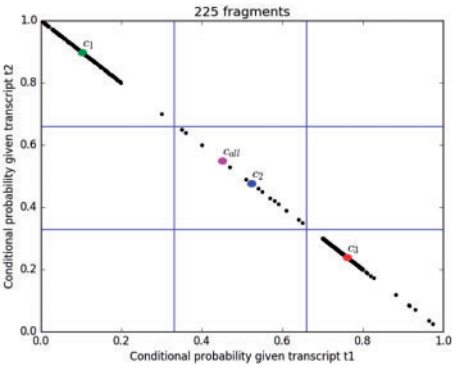

Fig. 3. Factorizing an equivalence class consisting of 225 fragments and 2 transcripts into k = 3 bins. Each dot represents one fragment. The vertical lines indicate borders of bins for transcript t_1 and the horizontal lines show borders of bins for transcript t_2. The purple circle with label c_{all} shows the center for original equivalence class. The rest of the circles are indicators of the centers for each cluster after the factorization

we expect to obtain results closer to the *full model*; yet, the number of clusters over which one must iterate to apply the EM algorithm is still much smaller than the total number of fragments (see Results).

Though we have implemented and experimented with both of these alternative factorizations, in this paper we will focus on the *range-based* factorization, as we observe that it almost always provides a better approximation of the likelihood than the *rank-based* factorization.

4 Results

We test the ability of our proposed factorization to improve the approximation of the *full model* likelihood on both synthetic and experimental data. We demonstrate that, as expected, the *range-based* factorization almost always provides a very good approximation of the *full model* likelihood. Interestingly, we also observe that it sometimes leads to a slightly more accurate solution than when no factorization is applied (i.e. when the likelihood is evaluated for each fragment independently). Though we have not investigated this in depth, it is likely that, in some cases, a small degree of smoothing of the conditional probabilities can lead to a more stable and accurate solution.

We consider both small-scale and transcriptome-wide simulated data. In Section 4.1 we consider simulations over the transcripts from families of paralogous genes. Such situations represent the most challenging abundance estimation problems for transcript quantification tools since high levels of multi-mapping are prevalent. We conduct the simulations over many random settings of the abundances of these transcripts, and look at how well different methods are able to recover the true abundances at different average coverage levels. We directly observe how, in the most adversarial situations, the proposed factorization allows us to recover important information that leads to improved abundance estimates.

In Section 4.2 we explore the effect that different factorizations have on abundance estimates transcriptome-wide. Here, we observe that, while the data-driven factorizations lead to improved abundance estimates, the differences between methods becomes much smaller, since the statistics are aggregated over the entire transcriptome and since many transcripts fall into the 'easy' case of abundance estimation. The differences between methods, while still

moderate, are larger when we restrict our assessment to a more difficult subset of transcripts.

Finally, in Section 4.3, we examine the effect of different factorization methods over experimentally sequenced data. We explore how closely different factorizations approach the abundance estimates derived by *RSEM*—though we note (as observed in some of the simulated data) that *RSEM* is not necessarily more accurate than the alternative methods or factorizations.

In Sections 4.1–4.3 we consider the transcript abundance estimates generated by *RSEM*, *eXpress* (both in default mode and with 50 batch EM rounds) and variants of *Salmon* (using different factorizations). We focus on the performance of these tools when quantifying abundances using alignments, instead of mappings (Srivastava *et al.*, 2016). We keep the input data as close as possible, since the purpose of this paper is not an investigation of the effect of alignment versus mapping on expression estimation, but rather the effect of the factorization of the likelihood and how that factorization affects inference. We noticed that, regardless of the factorization used, there was a small but persistent gap between non-alignment-based tools (*kallisto* and mapping-based variants of *Salmon*) compared to *RSEM* and alignment-based variants of *Salmon* on the *RSEM-sim* data. It is not clear that this is due to any fundamental superiority of alignment compared to mapping, but rather, may be a result of the fact that the specific error model, learned by *RSEM* and used to simulate reads in *RSEM-sim*, acts as a 'side-channel' of information for alignment-based approaches. However, this question, though outside the scope of this paper, deserves further consideration and analysis.

In the Supplementary Material, we explore the effect of these factorizations on mapping-based solutions by comparing different variants of mapping-based *Salmon* with *kallisto* (which only allows using pseudoalignment for quantification).

Alternative factorization variants:

Salmon (i.e. without any modification) uses a *compatibility-based* notion of equivalence classes called 'rich' equivalence classes. Under this notion, the equivalence classes themselves are *compatibility-based*, but each transcript-equivalence class pair is associated with a scalar weight which is computed as the mean conditional probability of all fragments in this equivalence class to derive from this transcript. We also consider a variant of *Salmon* (denoted as *Salmon-U* herein) that adopts a purely *compatibility-based* notion of equivalence classes. That is, it stores no extra information about the conditional probability of deriving the fragments in each equivalence class from the different transcripts, and during inference considers only that $\Pr(f|t) = \Pr(p|f,t) = 1/\tilde{\ell}_t$, where $\tilde{\ell}_t$ is the effective length of transcript t and is defined as $\tilde{\ell}_t = \ell_t - \mu_d^{\ell_t}$. $\mu_d^{\ell_t}$ is the mean of the truncated empirical fragment length distribution as described in Patro *et al.*, 2017.

We also consider a variant of *Salmon*, *Salmon-FM*, that performs no additional factorization. Instead, like *RSEM*, it considers each fragment and its relevant conditional probabilities independently. In this case, the only difference between *Salmon-FM* and *RSEM* is that the former computes the conditional fragment probabilities using an *online* stochastic inference algorithm, while *RSEM* recomputes the conditional fragment probabilities after updating auxiliary model parameters during the first 10 iterations of an offline (i.e. batch) EM procedure.

Finally, we consider a variant of *Salmon*, *Salmon-RF*, that uses the range-factorization described in Section 3.2 to generate equivalence classes based on $\sim_r$ and compute the associated weights.

We use both the mean absolute relative difference (MARD) and Spearman correlation to assess performance. We define the absolute relative difference (ARD) as:

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i + y_i = 0 \\ \dfrac{|x_i - y_i|}{(x_i + y_i)} & \text{otherwise} \end{cases}, \tag{7}$$

Where y_i is the estimated number of reads originating from t_i and x_i is the true (or assumed) number of reads originating from t_i. The MARD is simply defined as $\mathrm{MARD} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{ARD}_i$, where M is the total number of transcripts.

Experimental setup and software parameters:

In the tests below, *Salmon* v0.8.0 was run in alignment mode with the `–useErrorModel` flag. *Salmon-RF* consists of *Salmon* run with `–useRangeClusterEqClasses 4`. *Salmon-U* consists of *Salmon* run with `–noRichEqClasses`. *RSEM* v1.3.0 was run with default parameters. *eXpress* v1.5.1 was run with `–no-bias-correct` and other parameters were left as default (the extra parameter `–additional-batch 50` was used to produce the *eXpress (+50)* results). All alignments were generated using *Bowtie 2* version 2.2.9 with the default parameters chosen by *RSEM*. We note that these default parameters disallow indels in the resulting alignments (though *Salmon* and *eXpress* allow indels in the alignments they process, *RSEM* does not). Further, we note that since we examine simulated data without bias and since we compare against *RSEM* (which does not model sequence-specific or fragment-GC bias) in the experimental data, we run all other methods without bias correction. On experimental RNA-seq data, one might expect bias correction alone to substantially improve the accuracy of a given method. Though those accuracy improvements should be orthogonal to those obtained by improving the fidelity of the likelihood function. All the tests are performed on a 64-bit Linux server with 256GB of RAM and 4 x 6-core Intel Xeon E5-4607 v2 CPUs running at 2.60GHz.

4.1 Small-scale simulations on `RAD51` and its paralogs

We first consider a few small-scale simulations to motivate how the conditional probabilities considered by the *full model* (and approximated closely by the *range-based* equivalence classes) might improve abundance estimates. We note that these simulations are specifically constructed to represent adversarial and difficult-to-quantify mixtures of highly related isoforms. We consider the transcripts from large families of paralogous genes, under many random distributions of abundances. Often, the fragments will align to many different transcripts with few-or-no nucleotide differences, and sometimes even with similar implied insert sizes. Thus, we expect that closely approximating the conditional fragment probabilities might have a large effect in this case. We note, however, that such adversarial abundance configurations are likely rare in experimental data.

We consider two different, small-scale tests focusing around the gene `RAD51` and members of its paralogous family in *Homo Sapiens*. The `RAD51` family includes eight paralogous genes including `RAD51` itself. `RAD51` codes for a 339-amino acid protein that has a significant role in repairing double strand breaks of DNA (Yates *et al.*, 2015).

In the first experiment we apply *RSEM* and all varieties of *Salmon* on all isoforms of the `RAD51` gene. We extracted all (10) reference transcripts of `RAD51` from the Ensembl (release 80) reference transcriptome. True reads counts for all transcripts were generated by sampling a read count for each transcript uniformly over $[1, 200]$; these counts represent base-depth coverage (left) in Figure 4a. These counts were multiplied by 10 to derive the input read counts at $10X$

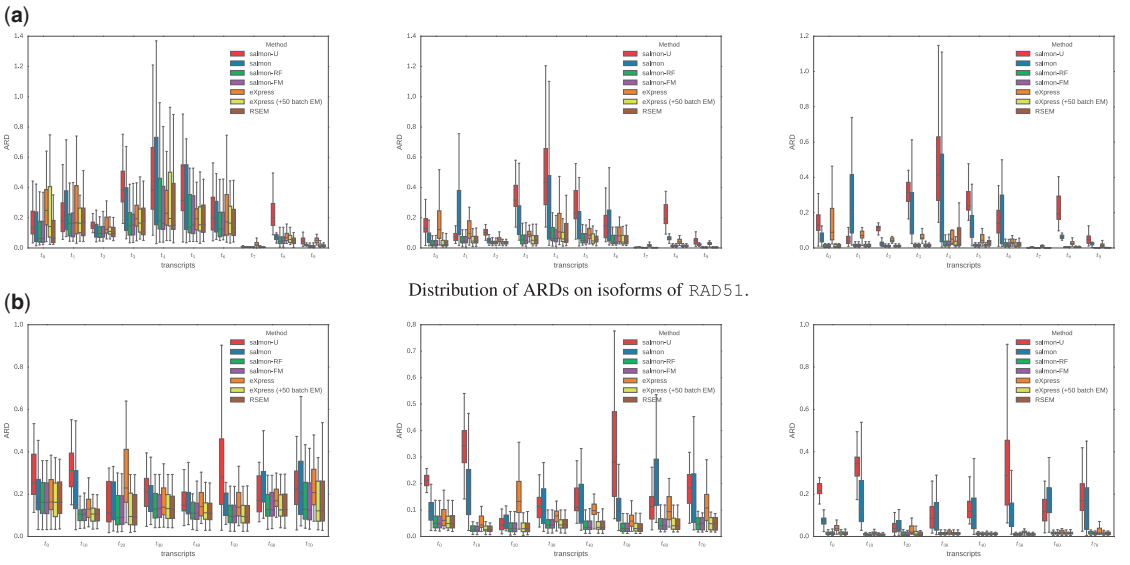

Fig. 4. Applying different methods of transcript abundance estimation in alignment mode on two sets of data in 3 depth of fragment sequencing. Top (**a**) are all isoform transcripts of gene RAD51. The bottom (**b**) is from transcripts of four different paralogs of RAD51, RAD51B, RAD51C, RAD51D. In each row the left most plot refers to experiment with counts of 1X coverage, the middle one to 10X and the most right plot refers to the experiment with fragment counts of 100X coverage

coverage (Fig. 4a, center) and by 100 to derive the counts at 100*X* coverage (Fig. 4a, right).

Given these read counts, the Polyester simulator (Frazee *et al.*, 2015) was then used to simulate five different read sets (replicates) from the same input distribution. This entire procedure was repeated 30 times, setting R's random seed from 1 to 30 in sequence.

Since the reads are simulated, we can assess the deviation of the estimated abundances from the exact abundances for each transcript. We use the absolute relative difference (ARD) of estimated versus true read counts (Equation (7)) as the metric to evaluate the accuracy of different methods for each transcript over replicates, and Figure 4a shows a box plot of the distribution of ARD values over the 30 simulations.

As we expect, *Salmon-U* generally yields the largest ARDs, failing to utilize the information contained in the conditional fragment probabilities. *Salmon* generally performs better, suggesting that, even in this complex scenario, the aggregate weight maintained in the rich equivalence classes helps to recover some (but not all) of the fidelity of the full model. However, *Salmon-RF*, while only slightly increasing the number of equivalence classes considered, produces ARDs very close to those of *RSEM*, *eXpress (+50)* and *Salmon-FM*. This suggests that, even in this adversarial scenario, the *range-based* equivalence classes allow us to recover the inferential accuracy of the *full model*.

To further explore difficult abundance estimation scenarios, we consider the case of the presence of high abundance isoforms from more than one gene in the reference. Therefore, in the second set of experiments we consider four paralogs of RAD51 (RAD51, RAD51B, RAD51C and RAD51D). We extract all transcripts corresponding to these genes and we run the same simulation as above with respect to all of these transcripts. Evaluation of ARDs for every transcript in all genes is displayed in Figure 4b. The results in this case are similar to what was observed in the single gene experiment. In some cases, like transcript ENST00000553595 from RAD51B (which is displayed as t_{10} in Fig. 4b), both *Salmon-U* and *Salmon* fail to estimate an accurate abundance. In other cases *Salmon* performs better than *Salmon-U*, e.g. transcript ENST00000585947 from RAD51D (displayed as t_{50} in Fig. 4b). For almost every transcript, *Salmon-RF*, *Salmon-FM*, *eXpress (+50)* (*eXpress* under default settings performs a bit worse) and *RSEM* all perform similarly and better than the methods that adopt a purely *compatibility-based* factorization of the likelihood. As this simulation contains a large number of transcripts, we plot, in Figure 4b, the box plots for only every tenth transcript to make the plot more interpretable. The complete plot containing the ARD values for all transcripts of this paralogous family is provided in Supplementary Figure S1. Supplementary Figure S2 shows gene specific performance of methods for all transcripts of RAD51C and RAD51D in this experiment. For transcripts of RAD51C, all factorizations (even the basic rich equivalence classes of *Salmon*) perform relatively well compared to *Salmon-U*. For accurately estimating the abundances of transcripts from RAD51D, however, improved factorizations (*Salmon-RF*) seem to be essential.

We ran quantification tools in non-alignment mode on both RAD51's transcripts and also all transcripts from the same RAD51 's paralog set we consider above (Supplementary Fig. S3). We also performed similar simulations at three different depths from the gene SEZ6 and its paralogs (SEZ6L and SEZ6L2) and followed the same set of steps to compare the performance of different tools in non-alignment mode, the result for this gene are presented in Supplementary Figure S4.

4.2 Transcriptome-wide analysis on synthetic data

To assess the performance of the proposed model on a large dataset of RNA seq reads, we generate synthetic data using *RSEM-sim*, and adopting the procedure used by Bray *et al.* (2016). *RSEM* model parameters were generated by running *RSEM* on sample NA12716_7 from the GEUVADIS (Lappalainen *et al.*, 2013) study. Using these model parameters, *RSEM-sim* was then used to generate a sample consisting of 30M 75 bp paired-end RNA-seq reads.

Again, we explore the performance of *RSEM*, *eXpress* (both in default mode and with 50 rounds of batch EM) and four different variants of *Salmon* (*Salmon-U*, *Salmon*, *Salmon-RF* and *Salmon-FM*). We compute the Spearman correlation and MARD metrics of each of these methods compared with the true (i.e. simulated) abundances. As we observe in Table 1, discarding all weight information in equivalence classes (*Salmon-U*) causes a drop in performance compared to the case with a single scalar per equivalence class-transcript pair (*Salmon*). Using the range-factorization proposed in this paper improves both the correlation and MARD measures even further, and brings its accuracy on par with that of *RSEM* and *Salmon-FM*, which adopt no factorization and run an EM algorithm that scales in the number of alignments in each iteration. In the default mode (i.e. using a single online pass), *eXpress* produces a larger MARD and lower correlation than any of the tools that run the batch EM until convergence. With 50 extra batch EM rounds (*eXpress (+50)*), *eXpress* performs more similarly to the other tools. We note that, in this data, the number of equivalence classes produced by the *range-based* factorization is ∼586 000, only ∼150 000 greater than the ∼438 000 *compatibility-based* equivalence classes. Both of these numbers are orders-of-magnitude smaller than the ∼ 100 000 000 distinct alignments for this dataset. The number of equivalence classes for all methods (sequenced fragments for Salmon-FM) is shown in Table 2. This table also reports the number of 'hits'. The number of hits is the sum, over each equivalence class, of the number of transcripts in this equivalence class—i.e. $\sum_{\mathcal{F}^q \in \mathcal{C}} |\Omega(\mathcal{F}^q)|$. This is the total number of items processed during each round of the EM algorithm. This small number of equivalence classes and hits allows the *Salmon-RF* model to run as fast as *Salmon*, which runs considerably faster than *Salmon-FM*, which, in turn, runs considerably faster than *RSEM*. With the exception of *eXpress*, which implements a constant-memory algorithm by design, the memory usage profiles for these different tools track the timing results (as expected). For more details, refer to Supplementary Figures S5 and S6.

Though we observe an improvement for *Salmon-RF* and *Salmon-FM* over *Salmon* and especially *Salmon-U* in this case, we note that it is relatively small in scale. This is because, while the aggressive *compatibility-based* factorizations do give up information, common expression patterns may not be complex or difficult enough to be greatly affected by the lossy factorization of the likelihood. Also, however, these aggregate metrics are computed over the entire transcriptome, and so, difficulties of these factorizations in deconvolving particularly complex scenarios may become lost in the noise of the vast number of good predictions.

To focus on the more difficult cases, we computed our accuracy metrics on a subset of the simulated data. Specifically, retaining the original abundance estimates over the entire transcriptome, we restricted our analysis to those transcripts for which *RSEM* obtained an ARD between 0.25 and 0.75. The motivation for choosing these values is to discard the particularly 'easy' to quantify transcripts (where the *full model* is likely neither necessary nor particularly helpful) as well as the 'hopeless' transcripts (those where the inference exhibits significant error even under the reference implementation of the *full model*). The results of this analysis are shown in Table 3. While the trend is similar to that observed on the full data, the difference between methods (and the impressive performance of *Salmon-RF*) becomes more clear. Specifically, we observe that *Salmon* outperforms *Salmon-U*, but this time the gap between *Salmon* and *Salmon-RF*, *Salmon-FM* and *RSEM* is larger. This is most likely because this particular subset of transcripts presents a more difficult inference challenge, where the conditional probabilities provide useful evidence. In the case of these transcripts, running the EM algorithm until convergence seems particularly important, as we observe that *eXpress* (and even *eXpress (+50)*) trail the other methods, especially in terms of the MARD. This makes it evident that further refinement of the abundance estimates (i.e. more rounds of the EM) over a representation of the data encoding conditional fragment probabilities (as done in *RSEM*, *Salmon-FM* and *Salmon-RF*) is necessary to obtain improved accuracy on these transcripts.

Table 1. Spearman correlation and MARD of quantification results compared to true abundances for synthetic data on all transcripts

	MARD	Spearman
Salmon-U	0.24	0.80
Salmon	0.22	0.81
Salmon-RF	0.21	0.83
Salmon-FM	0.21	0.83
eXpress	0.29	0.78
eXpress (+50)	0.23	0.83
RSEM	0.21	0.82

Table 2. The number of equivalence classes and hits, in the simulated data, under different likelihood factorizations

	Salmon-U	*Salmon*	*Salmon-RF*	*Salmon-FM*
# eq. classes	438 393	438 393	625 638	29 447 710
# hits	5 986 371	5 986 371	8 212 669	103 663 423

Table 3. The performance of different methods when restricted to the subset of transcripts where *RSEM*'s ARD is in $[0.25, 0.75]$

	MARD	Spearman
Salmon-U	0.46	0.56
Salmon	0.43	0.58
Salmon-RF	0.41	0.64
Salmon-FM	0.41	0.65
eXpress	0.53	0.54
eXpress (+50)	0.48	0.59
RSEM	0.41	0.65

We further investigate the performance of tools in non-alignment mode as well. Spearman correlation and MARD of transcript quantification with different tools on RSEM simulated data is presented in Supplementary Tables 1 and 2.

4.3 Transcriptome-wide analysis on experimental data

Finally, we explore the effect of our data-driven factorization method with the different versions of *Salmon* using experimental data from the SEQC(MSEQ-III) consortium (Consortium *et al.*, 2014) (NCBI GEO accession `SRR1215996 - SRR1217002`). Specifically, the library is prepared on Universal Human Reference RNA (UHRR) from Stratagene and ERCC Spike-In controls and consists of ∼11M 100 bp, paired-end reads sequenced on an Illumina HiSeq 2000 platform. The experiment consists of seven replicates with the same flowcell and barcodes but on different lanes.

As described previously in section 4 we compare the performance of *Salmon*, *Salmon-FM*, *Salmon-RF*, *Salmon-U*, *eXpress*, *eXpress (+50)* with *RSEM*. However, unlike in previous sections,

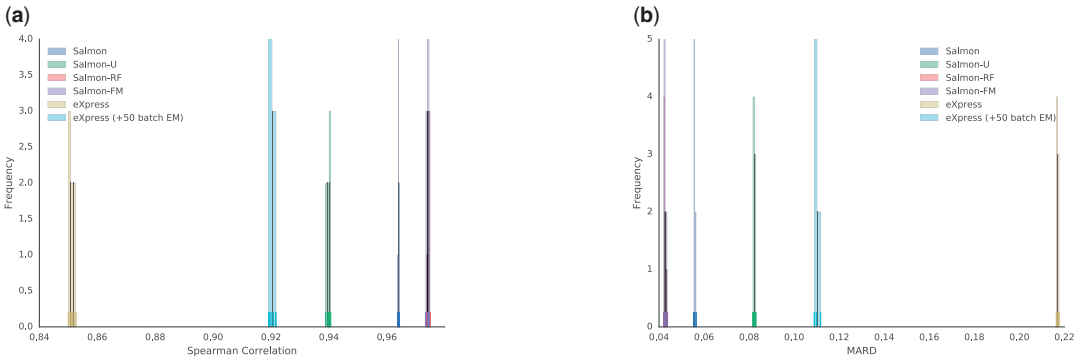

Fig. 5. Comparison of the transcript abundances in different versions of salmon on the experimental data with seven technical replicates and using rsem abundance estimates as the ground truth (a) The Spearman correlation of transcripts abundance estimations with *RSEM* results reveals that *Salmon-FM* is highly correlated with *RSEM*. Very similar correlation with *RSEM* is observed by the proposed data-driven factorization, *Salmon-RF*. *Salmon* displays a lower correlation than *Salmon-RF*, but a higher correlation than *Salmon-U*. The variants of *eXpress* show a lower correlation than *Salmon-U*, with the offline EM iterations increasing *eXpress'* correlation considerably. (b) Comparing the MARD of estimated transcript fragment counts with respect to *RSEM* results shows similar trend to that observed with the Spearman correlations; i.e. *Salmon-FM* has the least error rate using RSEM abundances as the truth while *Salmon-RF* perform equally well. *Salmon* exhibits a lower MARD than *Salmon-U*, which is followed by both variants of *eXpress*

Table 4. The number of equivalence classes and hits, in the experimental data, under different likelihood factorizations

	Salmon-U	*Salmon*	*Salmon-RF*	*Salmon-FM*
# eq. classes	427 611	427 611	624 340	9 077 708
# hits	5 737 414	5 737 414	8 318 638	50 325 595

here, we lack a ground truth. Thus, we measure the accuracy of each method on the estimated number of reads, treating *RSEM*'s estimations of the number of reads for each transcript (which is observed to be among the most accurate on synthetic data in previous sections) as the truth. We perform a comparison across all seven replicates and consider the Spearman correlation and MARD metrics. Since these are technical replicates, we expect the performance over each replicate to be very similar, though we plot the results as a distribution in Figure 5a and b. The results on experimental data follow the same trend as we observed on synthetic data. That is, *Salmon-FM* correlates well with *RSEM* (as expected) because of the availability of full fragment level transcript probabilities. Likewise, we again observe that our proposed data-driven factorization method, *Salmon-RF*, performs essentially the same as the *full model*. Both of these methods agree more closely with *RSEM* than does *Salmon*, and again, *Salmon-U*, ignoring all fragment-level conditional probabilities, is further from *RSEM*'s results. The number of equivalence classes for each factorization are shown in Table 4. We also observe that *eXpress*, in its default mode, performs most differently from *RSEM* of the methods we considered. As expected, running additional rounds of the batch EM (*eXpress (+50)*) increases the similarity of *eXpress* ' estimations with those of *RSEM*; though it is still less similar than the other methods.

5 Conclusion

While compatibility-based equivalence class factorizations (Bray *et al.*, 2016; Nicolae *et al.*, 2011; Patro *et al.*, 2014; Srivastava *et al.*, 2016; Turro *et al.*, 2011) have paved the way in terms of substantially improving the efficiency of the iterative optimization procedures used for transcript-level quantification from RNA-seq data, they nonetheless make sacrifices in modeling fidelity to achieve this. While these methods generally perform adequately in terms of transcriptome-wide assessments, there are still important situations in which their compatibility-centric factorization of the underlying likelihood function discards information that can be important for accurate abundance estimates. *Salmon* (Patro *et al.*, 2017) uses a dual-phase inference algorithm that allows it to recover some of the information discarded by other approaches. It improves upon the approximate factorization of the full likelihood function by incorporating a notion of *rich* equivalence classes of fragments. In some, but not all cases, this improved factorization is sufficient to recover the lost accuracy of the full model.

In this paper, we have introduced a data-driven factorization of the likelihood function that makes use of *Salmon* 's dual-phase inference algorithm (*Salmon-RF*). We have shown that this improved factorization is able to match the accuracy of the *full model* while still maintaining a reduced representation that is orders of magnitude smaller than the total number of fragment alignments.

We believe that this data-driven factorization represents the right tradeoff between efficiency and accuracy. Specifically, it demonstrates an almost indistinguishable sacrifice in efficiency beyond the factorization already employed by *Salmon* (which, itself, is similar in size to those employed by *mmseq*, *Sailfish* and *kallisto*), while producing no perceptible loss in accuracy compared to the full per-fragment likelihood function used by *RSEM* and similar methods.

In this paper, we have focused on the effect that the adopted factorization of the likelihood function can have on the ability of a method to accurately estimate transcript abundance. However, we note that there still remain small but interesting differences between methods that employ alignment and those that rely on mapping (i.e. quasi-mapping or pseudoalignment). Fully exploring the nature of these differences, and how they interact with the factorizations proposed herein, is an interesting direction for future work.

Finally, while we have investigated the effect different factorizations have on maximum likelihood estimates, fully exploring the effect they have in estimating the variance of these estimates (e.g. via bootstrapping) or even in estimating the full posterior distribution of abundances (e.g. via Gibbs sampling) is another interesting direction for future work.

Funding

We gratefully acknowledge support from National Science Foundation grant BBSRC-NSF/BIO-1564917.

Conflict of Interest: none declared.

References

Bray,N.L. *et al*. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*., **34**, 525–527.

Consortium,S.-I. *et al*. (2014) A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol*., **32**, 903–914.

Frazee,A.C. *et al*. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

Glaus,P. *et al*. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.

Hensman,J. *et al*. (2015) Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, **31**, 3881–3889.

Lappalainen,T. *et al*. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 1.

Li,B. *et al*. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Nariai,N. *et al*. (2013) TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. *Bioinformatics*, btt381.

Nariai,N. *et al*. (2014) TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq readsonline. *BMC Genomics*, **15**, S5.

Nicolae,M. *et al*. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol*., **6**, 9.

Patro,R. *et al*. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol*., **32**, 462–464.

Patro,R. *et al*. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.

Roberts,A. and Pachter,L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, **10**, 71–73.

Salzman,J. *et al*. (2011) Statistical modeling of RNA-Seq data. *Stat. Sci. Rev. J. Inst. Math. Stat*., **26**, 62–83.

Srivastava,A. *et al*. (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, **32**, i192–i200.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.

Turro,E. *et al*. (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*., **12**, 1.

Yates,A. *et al*. (2015) Ensembl 2016. *Nucleic Acids Res*., gkv1157.