COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 2, pp. 69-82, 2010
© 2010 International Press
001

METHODS FOR ALLOCATING AMBIGUOUS SHORT-READS*

MARGARET TAUB†, DORON LIPSON‡, AND TERENCE P. SPEED†§

Abstract. With the rise in prominence of biological research using new short-read DNA sequencing technologies comes the need for new techniques for aligning and assigning these reads to their genomic location of origin. Until now, methods for allocating reads which align with equal or similar fidelity to multiple genomic locations have not been model-based, and have tended to ignore potentially informative data. Here, we demonstrate that existing methods for assigning ambiguous reads can produce biased results. We then present new methods for allocating ambiguous reads to the genome, developed within a framework of statistical modeling, which show promise in alleviating these biases, both in simulated and real data.

1. Introduction.

1.1. Background. In recent months, there has been an explosion in the use of new DNA sequencing technologies which produce relatively short reads (<100 base-pairs (bp)) [1, 2, 3, 4]. For many biological applications of this technology such as transcriptome analysis or detection of epigenetic features, these reads need to be aligned to a reference genome so that their genomic location of origin can be determined. Alignment fidelity can be affected by many factors associated with the read itself, including read length, error rate, and error type (e.g. substitution, insertion, deletion); as well as factors inherent to the genome being aligned against, such as levels of sequence homology (due to gene families, for example), presence of repetitive regions and errors in the reference sequence.

This report presents a new method for allocating what have been referred to in the literature as *multireads* [5], by which we mean sequenced reads which can be mapped with equal or close to equal fidelity to multiple locations in the genome. While some previous methods have been designed to make use of such reads in downstream analysis (see Section 1.2), we present here some steps toward developing a more unified statistical framework for allocating these reads. Throughout, we assume a simplified context where the genome has been divided in some way into what we will call a set of *transcripts*, which can refer to any set of non-overlapping genomic regions, such as exons, genes or transcripts in the absence of alternative splicing. Extension of our methods to more complex situations should be possible with some modifications. We also note that by using the term *allocate* we do not mean to imply that the true origin of each read is inferred. Instead we allow each ambiguous read to contribute to the estimated abundance of a set of potential transcripts of origin, as explained below.

*Dedicated to Michael Waterman on the occasion of his 67th birthday.

†Department of Statistics, University of California, Berkeley, CA 94720.
‡Helicos BioSciences Corporation, Cambridge, MA 02139.
§The Walter and Eliza Hall Institute of Medical Research, Victoria 3050, Australia.

An important place where particular care is needed in allocating ambiguous reads arises when two or more transcripts have very similar sequence compositions but quite different abundances in the sample of interest. In this case, the low-abundance transcript is likely to be over-estimated by existing allocation methods. Since one potential advantage of digital technologies like the current short-read sequencing techniques is to more accurately detect low-expressing genes, such biases should be avoided if possible. Our main goal here is to develop a method which adequately corrects for this phenomenon.

1.2. Previous work. The issue of assigning genetic origin to ambiguous short sequences of DNA is not unique to new short-read sequencing technologies. In fact, technologies such as serial analysis of gene expression (SAGE) [6], cap analysis of gene expression (CAGE)[7], massively parallel signature sequencing (MPSS)[8] and polony multiplex analysis of gene expression (PMAGE)[9] are all similar in flavor to current short-read sequencing technologies, and in general have an even higher prevalence of ambiguous reads, due to their shorter sequence length (17-20nt) [4]. A method of proportional reassignment (described further below) which has since been applied to short-read sequencing data [4] was developed for working with CAGE data [10]. Otherwise, ambiguous read allocation methods developed for these previous technologies do not seem to have been applied to more recent short-read sequencing studies.

Regarding more recent high-throughput sequencing studies, most initial studies employed a method of discarding any reads that mapped equally well to multiple locations, thus counting only reads that had a unique maximally-scoring alignment. This was in part due to the defaults of the alignment software which were used, such as Illumina's ELAND aligner, which automatically discards all reads which have more than a unique best match [1]. We will refer to this method as the *unique* assignment method (see Equation 1). As an alternative, some researchers have used a method of proportional assignment to map reads with multiple exact hits [10, 4, 5]. First, they align their reads to the reference genome, using parameters which return an exact best match, if one exists, or a list of equally good matches, if no unique best is available. They then map all reads with exactly one best hit to the transcriptome, giving an initial estimate of transcript abundance. Finally, they allocate reads that hit multiple places equally well in proportion to the current estimates of transcript prevalence for those locations. We will refer to this as the *proportional* assignment method (see Equation 2). We will compare our methods to both of these.

To our knowledge, there is one additional method of resolving ambiguous reads, which is employed by the read mapping software MAQ [11]. MAQ randomly selects one location from the set of best assignments, and assigns the read to that location. In our tests, this method performs similarly to unique and proportional assignment

(results not shown).

2. Model.

2.1. Overview. In developing our method, we noted that there were two pieces of information available to us that previous methods were, for the most part, ignoring. First, some prior or current estimate of transcript abundance (which is used by the proportional assignment method), and second, some sort of alignment quality score. As an illustration, given a read which maps to a low-abundance transcript with a perfect score, or to a high-abundance transcript with a slightly lower score, current methods would always assign the read to the low-abundance transcript. We argue that this may not always be desirable, as we will illustrate below.

2.2. A first pass. The idea of combining transcript abundance along with some kind of *weight* corresponding to alignment strength inspired our first iterative read-assignment method. While this method is not based on any model, its promising performance encouraged us to further development, as described below.

We assume we have a set of R reads, which we want to map to a set of T transcripts. We assume that our reads have been aligned to the reference set of transcripts, and that for each read-transcript pair, we have a score for that alignment, $S(r,t)$. The score should reflect the quality of the alignment, defined in some suitable way, depending on the technology that has generated the reads.

Our goal is to estimate $p(t)$, the fraction of reads in our sample that come from transcript t. In this context, we can make previous methods more explicit. The unique assignment method would estimate $p(t)$ by taking

$$\hat{p}_u(t) = \frac{1}{|R_u|} \sum_{r \in R_u} 1(S(r,t) > S(r,t'), \forall t' \neq t), \tag{1}$$

where R_u is the set of reads with a unique maximum score and where $1(x) = 1$ when x holds, and 0 otherwise. The proportional counting method would calculate $\hat{p}_u(t)$ as above, and then compute

$$\hat{p}_p(t) = \frac{1}{R} \sum_r \frac{\hat{p}_u(t) 1(S(r,t) \geq S(r,t'), \forall t' \neq t)}{\sum_{t''} \hat{p}_u(t'') 1(S(r,t'') \geq S(r,t'), \forall t' \neq t'')}. \tag{2}$$

Our method will be similar in flavor to the proportional counting method, but will incorporate the magnitude of the score $S(r,t)$. Rather than just weighting our reads by the score, we wanted our method to depend on how extreme one score was, relative to the score a random read generated from a particular transcript would have. To quantify this, we introduce a cumulative distribution function $F(x) = \text{pr}(S(r,t) \leq x)$, which gives the probability of observing a score less than or equal to x when read r originates from transcript t. We iterate by initializing $\hat{p}_c(t)$ by allocating reads equally

among transcripts. We then update $\hat{p}_c(t)$ by taking

$$\hat{p}_c^{new}(t) = \frac{1}{R}\sum_r \frac{\hat{p}_c(t)F(S(r,t))}{\sum_{t'}\hat{p}_c(t')F(S(r,t'))}.$$

We iterate this process until convergence. Since this process is based on a cumulative distribution function, we will refer to it as the *cumulative* assignment method. While it is not a complete probabilistic representation of our estimation context, it is intuitive and performs well.

2.3. An explicit probabilistic model. To construct our model, we introduce an unobserved variable $a(r,t)$ which is equal to 1 if read r originated from transcript t and 0 otherwise. The model parameter is the vector $\mathbf{p} = (p(t)), 0 \le p(t) \le 1, t = 1,\ldots,T, \sum_t p(t) = 1$, and we suppose that the vectors $\mathbf{a}(r) = (a(r,1),\ldots,a(r,T)), r = 1,\ldots,R$ are independent and identically distributed (i.i.d.) with distributions

$$\mathrm{pr}(a(r,t) = 1, a(r,t') = 0, \forall t' \neq t) = p(t),$$

or equivalently, as a multinomial,

$$\mathrm{pr}(\mathbf{a}(r)) = \prod_t p(t)^{a(r,t)}.$$

Our objective is to estimate $\mathbf{p}$, the proportion of reads originating from each transcript t. Note that this objective does not require unambiguously associating each read with a specific transcript.

Further, we introduce two density functions, $f_0(\cdot)$ and $f_1(\cdot)$, where $f_0(S(r,t))$ gives the probability of observing the score $S(r,t)$ when $a(r,t) = 0$, and $f_1(\cdot)$ is defined similarly for $a(r,t) = 1$. We suppose that the observed alignment scores are conditionally mutually independent random variables given the $\mathbf{a}(r)$ vectors, with conditional densities

$$\mathrm{pr}(S(r,t) = x|\mathbf{a}(r)) = f_{a(r,t)}(x).$$

We will work within an expectation-maximization (EM) framework for estimating $\mathbf{p}$. In this case, our full data are given by

$$(\mathbf{a},\mathbf{S}) = (a(r,t), S(r,t) : r = 1,\ldots,R; t = 1,\ldots,T)$$

and given values for $\mathbf{p}$, we have the following joint distribution, and hence likelihood, when viewed as a function of $\mathbf{p}$,

$$\mathrm{pr}(\mathbf{a},\mathbf{S};\mathbf{p}) = \mathrm{pr}(\mathbf{a};\mathbf{p})\mathrm{pr}(\mathbf{S}|\mathbf{a};\mathbf{p}).$$

Treating the $a(r,t)$ as known, this gives a full data log-likelihood function of

$$l(\mathbf{p}) = \sum_{r,t}\{a(r,t)\log(p(t)) + \log(f_{a(r,t)}(S(r,t)))\},$$

which, to complete the M-step, we want to maximize subject to the constraint $\sum_t p(t) = 1$. This gives

$$p^*(t) = \frac{\sum_r a(r,t)}{\sum_{r,t} a(r,t)}.$$

For the E-step, we want the conditional expected value of the full data log-likelihood given the observed data, which reduces to finding $\mathrm{E}(\mathbf{a}(r)|S(r,1), S(r,2), \ldots, S(r,T))$ which we evaluate using Bayes' Theorem, for $t = 1, \ldots, T$, as

$$\begin{aligned}
a^*(r,t) &= \mathrm{pr}(a(r,t) = 1, a(r,t') = 0, \forall t \neq t' | S(r,1), S(r,2), \ldots, S(r,T)) \\
&= \frac{\mathrm{pr}(S(r,1), S(r,2), \ldots, S(r,T) | a(r,t) = 1, a(r,t') = 0, \forall t' \neq t) p(t)}{\mathrm{pr}(S(r,1), S(r,2), \ldots, S(r,T))} \\
&= p(t) f_1(S(r,t)) \prod_{t' \neq t} f_0(S(r,t')) \Bigg/ \sum_{t'} \left(p(t') f_1(S(r,t')) \prod_{t'' \neq t'} f_0(S(r,t'')) \right) \\
&= p(t) \frac{f_1(S(r,t))}{f_0(S(r,t))} \Bigg/ \sum_{t'} p(t') \frac{f_1(S(r,t'))}{f_0(S(r,t'))}.
\end{aligned}$$

This outlines the EM framework which we can use to obtain our estimate of $\mathbf{p}$. We will refer to this method as the *EM* assignment method.

2.4. Density estimation. One thing we have not yet addressed is the determination of f_0 and f_1, or of our cumulative distribution F. We estimate f_0 by taking the full distribution of scores from all reads in the data set and creating a smoothed density estimate based on these scores. We can estimate f_1 in a similar way, taking the max score for each read and then creating a smoothed density estimate based on these max scores. This non-parametric density estimation is used in our EM assignment method. The cumulative distribution function F used in the cumulative assignment method is obtained by integrating the non-parametrically estimated density f_1 in the appropriate way.

As an extension and possible improvement, we have implemented a semi-parametric estimation method as well. As can be seen from the derivation above, the function of interest in our estimation is actually the ratio $f_1(x)/f_0(x)$. Based on our non-parametrically estimated density f_0, we define f_1 parametrically by

$$f_1(x) = \frac{e^{\alpha x} f_0(x)}{M_0(\alpha)}$$

where $M_0(\alpha)$ is the moment-generating function of f_0, evaluated at α. This functional form arises through a parameterization of the log of the ratio $f_1(x)/f_0(x)$ as a linear function of α. At this point we expect a simple exponential likelihood ratio, analogous to comparing two normals with different means and the same standard deviation (SD), to be adequate as it is simple and robust. We may want to extend this parameterization to include a quadratic term, analogous to comparing two normals with different means and SDs, in the future.

We can then rewrite our full data log-likelihood as

$$\begin{aligned} l(\mathbf{p}) &= \sum_{r,t}\{a(r,t)\log(p(t)) + \log(1(a(r,t)=0)f_0(S(r,t)) + 1(a(r,t)=1)f_1(S(r,t)))\} \\ &= \sum_{r,t}\{a(r,t)\log(p(t)) + \log(f_0(S(r,t))(1(a(r,t)=0) + 1(a(r,t)=1)\frac{f_1(S(r,t))}{f_0(S(r,t))}))\} \\ &= \sum_{r,t}\{a(r,t)\log(p(t)) + \log(f_0(S(r,t))) + 1(a(r,t)=1)(\alpha S(r,t) - \log(M_0(\alpha)))\}. \end{aligned}$$

Here, we want to maximize with respect to α, which tells us we want to find α^* which is the solution to

$$\frac{\sum_{r,t} a^*(r,t)S(r,t)}{\sum_{r,t} a^*(r,t)} = \frac{M_0'(\alpha)}{M_0(\alpha)}.$$

In this way, we can iteratively estimate the function f_1/f_0. We will refer to the assignment method employing this density estimation process as the *EM-Alpha* assignment method.

3. Results.

3.1. Simulation data set. We simuluated a set of 359,058 reads of length 15-40 bp from 5750 verified open reading frames (ORFs) in *S. cerevisiae* [12], with an error model based on performance metrics for an early prototype version of Helicos' HeliScope technology which, unlike Illumina's Genome Analyzer platform, has a relatively high presence of indels (insertion or deletion errors) compared to substitution errors. ORF lengths varied from 50-15,000 bp and the transcript abundance profile was set according to a yeast transcription profile previously measured with microarrays [13]. These reads were mapped against the yeast genome using a Smith-Waterman based aligner [2], which returns not only the alignment location, but also a score for the correspondence between the read and that location, based on a penalty model for each type of error. All transcripts which matched with a score of 3.5 or greater (out of a maximum of 5) were recorded, along with the scores, for each read. In our simulated data set, the number of transcripts mapped to by each read ranges from 1 to 2527, with 75% mapping to 25 transcripts or fewer. For our analysis, we excluded all reads that mapped to ≥ 100 transcripts, leaving a set of 306,223 reads. Of these, 75% map to 6 transcripts or fewer, with a mean of 9.5 hits per read.

For our simulated data set, we are fortunate enough to know what the true counts are for the genes we are considering, which makes it relatively easy to compare among different assignment methods. Since the total number of reads successfully assigned varies between the methods we convert our estimated counts to transcripts-per-million (tpm). In Figure 1, we show a series of MA plots comparing the true counts to those predicted by four methods: unique, proportional, cumulative and EM assignment. In

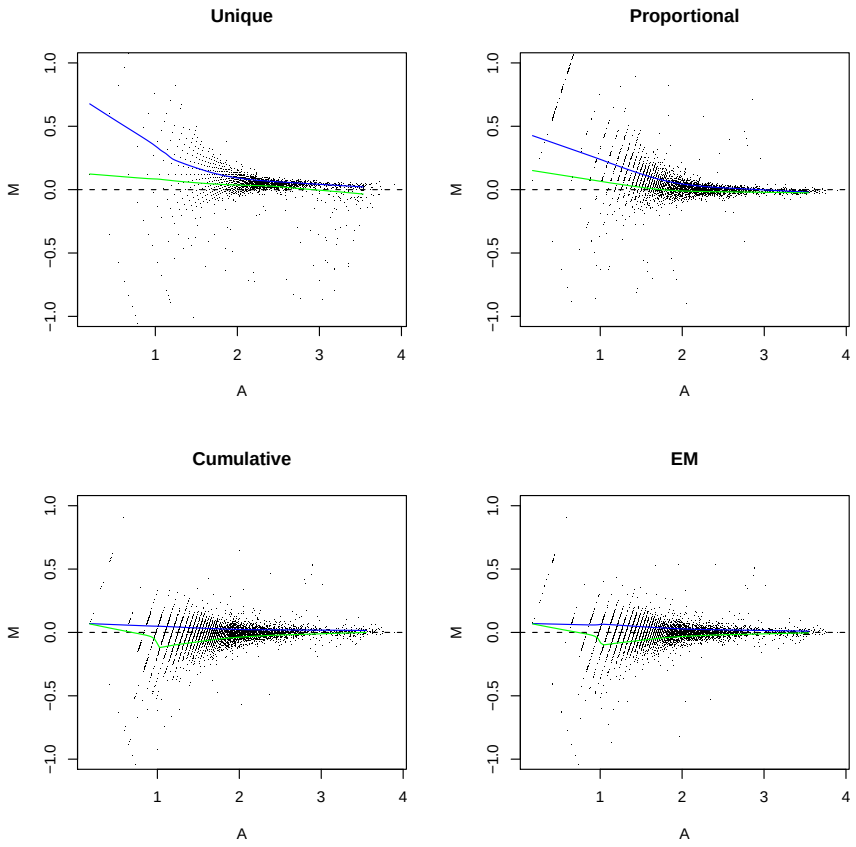

FIG. 1. *Comparison of mapping methods using MA plots of estimated counts against true counts, on a* $\log_{10}(tpm)$ *scale. Plots show mean* $\log_{10}$ *tpm on the* x*-axis and difference in* $\log_{10}$ *tpm on the* y*-axis, comparing each method to the truth. The dashed line indicates a perfect fit. The upper and lower solid lines indicate smoothed window-estimated* 75^{th} *and* 25^{th} *percentiles, respectively.*

these and all plots, 0.5 has been added to all counts before conversion to tpm, to avoid taking a log of 0, but to still allow for the display of predicted values for absent transcripts. These plots show the mean $\log_{10}$ tpm on the x-axis (A values), and the difference in $\log_{10}$ tpm on the y-axis (M values), comparing predicted results to the truth. The upper solid line shows a smoothed version of the 75^{th} percentile of M values, calculated over windows of A values of size 50 with an overlap of 45. The lower solid line is determined similarly for the 25^{th} percentile. The dashed line shows a perfect fit. The first two methods clearly show over-estimation of the low-abundance genes, while our methods improve on this considerably. However, very little difference is evident comparing cumulative and EM assignment. Figure 2 shows box plots of the difference between predicted and true $\log_{10}$ tpm values. While the proportional assignment method has a similar spread to the cumulative and EM assignment results,

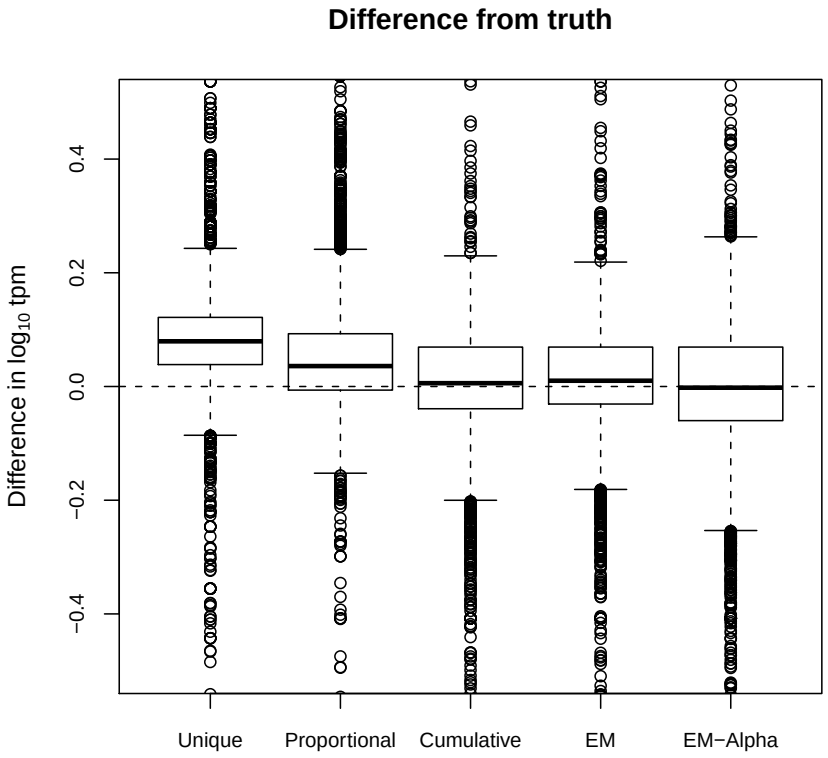

FIG. 2. *Comparison of mapping methods using box plots of differences between estimated and true* $\log_{10}$ *tpm values.*

the cumulative and EM boxes are more centered around zero, demonstrating less bias. Also included on this plot is a box for the EM-Alpha method, fit using semi-parametric estimation for the ratio f_1/f_0. We can see that for this data set, there is no improvement in fit with this further estimation step.

3.2. Yeast data set. In addition to our simulated data set, we have a set of reads generated from an mRNA sample from *S. cerevisiae*, using a prototype of Helicos' HeliScope single-molecule sequencer. Read lengths vary from 20-70 bp. In this case, the reference set is the complete set of 6,719 verified, uncharacterized and dubious ORFs from the SGD repository [12]. These reads have been aligned to the yeast genome as above.

Since in this case we do not know the true abundance of the different genes in the sample, we will make our method comparison by using the fact that longer reads can in general be mapped with greater specificity than shorter reads. We will take two

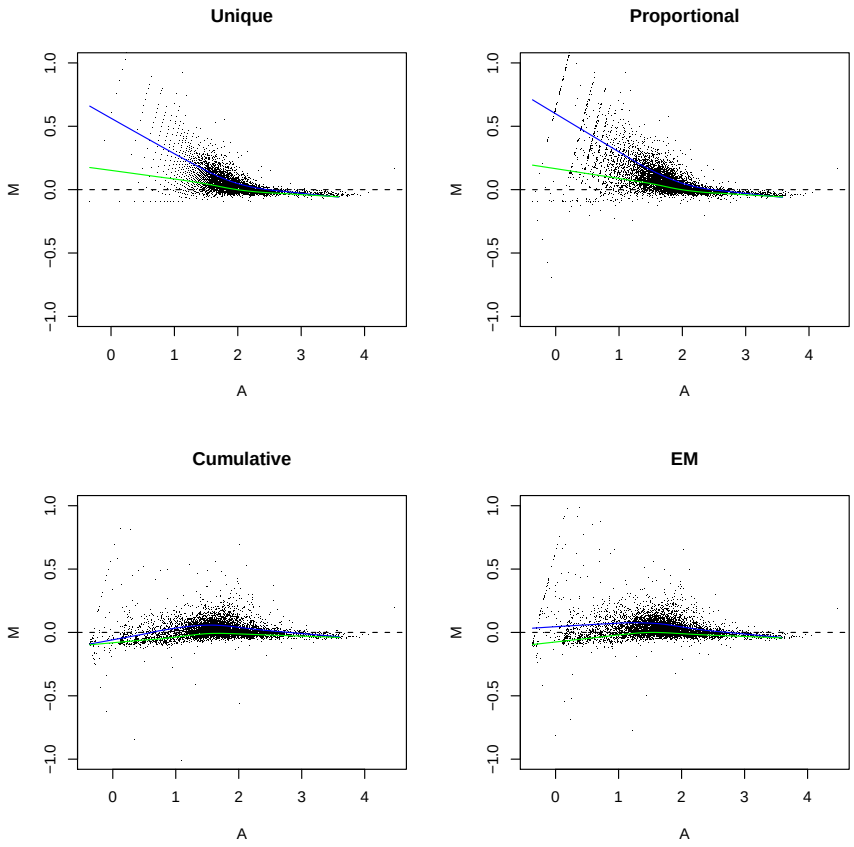

FIG. 3. *Comparison of mapping methods using MA plots of estimated counts from L20 data (using reads of length 20 or more) against estimated counts from L25 data (using reads of length 25 or more), on a* $\log_{10}(tpm)$ *scale. Plots show mean* $\log_{10}$ *tpm on the x-axis and difference in* $\log_{10}$ *tpm on the y-axis, comparing results from the two data sets. The dashed line indicates a perfect fit. The upper and lower solid lines indicate smoothed window-estimated* 75^{th} *and* 25^{th} *percentiles, respectively.*

different subsets of our data set: one with reads of length 20 or greater (L20 data), and one with reads of length 25 or greater (L25 data), and compare the estimated gene abundances from these two methods. Preferably, the gene abundances would be the same for the two data subsets, however, since we expect there to be more ambiguous reads from the set which allows for shorter reads, we know that the assignment method may affect the quality of replication.

For the data subset with reads of length 20 or higher, we have a total of 1,291,777 reads, with 75% mapping to 8 genes or fewer, and a mean of 10 hits per read. For the data subset with reads of length 25 or higher, we have a total of 1,039,137 reads, with 75% mapping to 2 genes or fewer, and a mean of 3.6 hits per read.

Figures 3 and 4 show the results of this comparison. We can see that again,

the simple assignment and proportional assignment methods tend to overcount low abundance genes, while the cumulative and EM methods correct this problem.

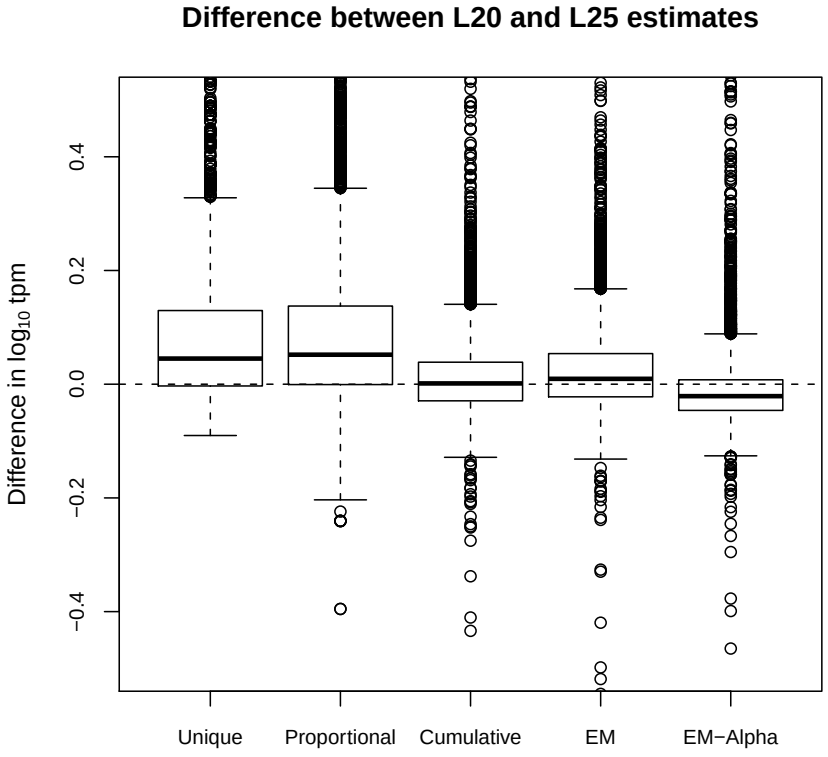

FIG. 4. *Comparison of mapping methods using box plots of differences between* $\log_{10}$ *tpm values estimated from L20 data (reads of length 20 or more) and those estimated from L25 data (reads of length 25 or more).*

In this case, it seems that the EM method with semi-parametric density estimation, EM-Alpha, may be performing better, as shown in the last box of Figure 4. A more specific comparison of the two EM implementations is shown in Figure 5. It is interesting to note that for this data set, as seen in Figures 4 and 5, the semi-parametric method looks about as good as the non-parametric method, with a slightly smaller interquartile range, but with some bias.

4. Discussion. First, it is clear from our analysis that there is room for improvement over existing methods of mapping ambiguous reads, particularly in the context we are considering. Both the unique and the proportional assignment methods overcount low-abundance transcripts, while our cumulative and EM assignment methods correct for this. Even though the majority of our reads had a unique maximum score

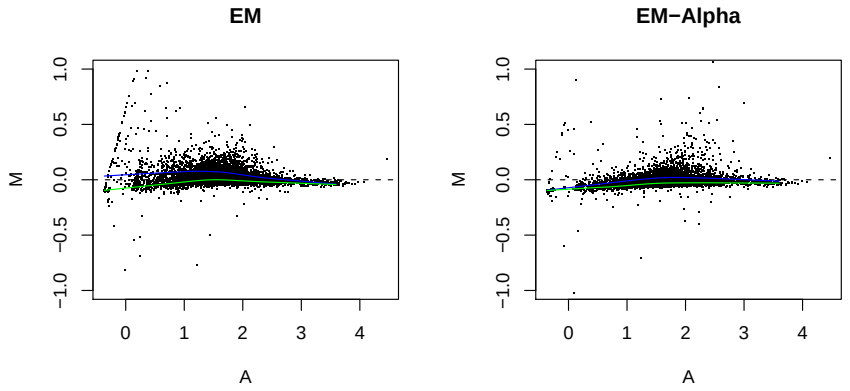

FIG. 5. *Comparison of EM and EM-Alpha mapping methods using MA plots of estimated counts from L20 data against estimated counts from L25 data, on a* $\log_{10}(tpm)$ *scale. Plots show mean* $\log_{10}$ *tpm on the x-axis and difference in* $\log_{10}$ *tpm on the y-axis, comparing results from the two data sets. The dashed line indicates a perfect fit. The upper and lower solid lines indicate smoothed window-estimated* 75^{th} *and* 25^{th} *percentiles, respectively.*

(89% in the simulated data, 93% in the L20 data set, and 94% in the L25 data set), we found that using only these reads did not give good estimates of transcript abundance, particularly for the less frequent transcripts. It seems reasonable to suppose that this effect could easily carry over to other contexts, with different score distributions within a read. An area for future work will be an application of these methods to other types of short-read data, including data sets with higher rates of substitution errors, where there is potentially a quite different distribution in scores between reads and transcripts.

In terms of development of our methodology, it is interesting to note that for both data sets, our initial iterative method based on the cumulative distribution function performs as well as our model-based method with non-parametric density estimation, and better than our model-based method with semi-parametric density estimation, at least in the simulated data example. While this may at first seem surprising, upon further examination, it becomes clear that there is a strong similarity between the cumulative distribution F and the ratio f_1/f_0, based on non-parametric estimation. These distributions are shown in Figure 6 for the simulated data. Also included is the semi-parametrically estimated ratio f_1/f_0. Figure 7 shows the same sets of plots for the two additional data sets. In this case, it is less clear why we are getting the results we have seen.

Since the semi-parametric density estimation requires approximating a moment-generating function, our current implementation may be hindered by numerical instabilities. There is clearly room for future work in improving this estimation step, with

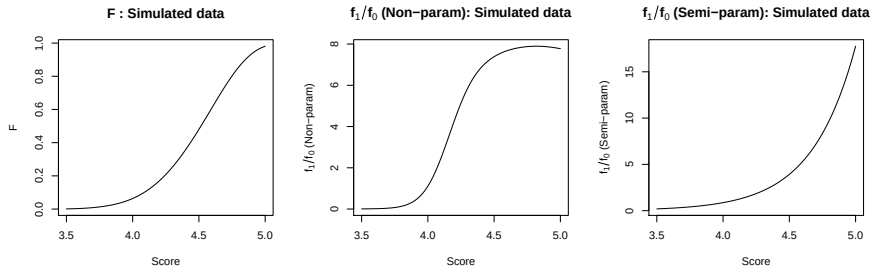

FIG. 6. *Comparison of F, the ratio f_1/f_0 estimated non-parametrically, and the ratio f_1/f_0 estimated semi-parametrically, for the simulated data set.*

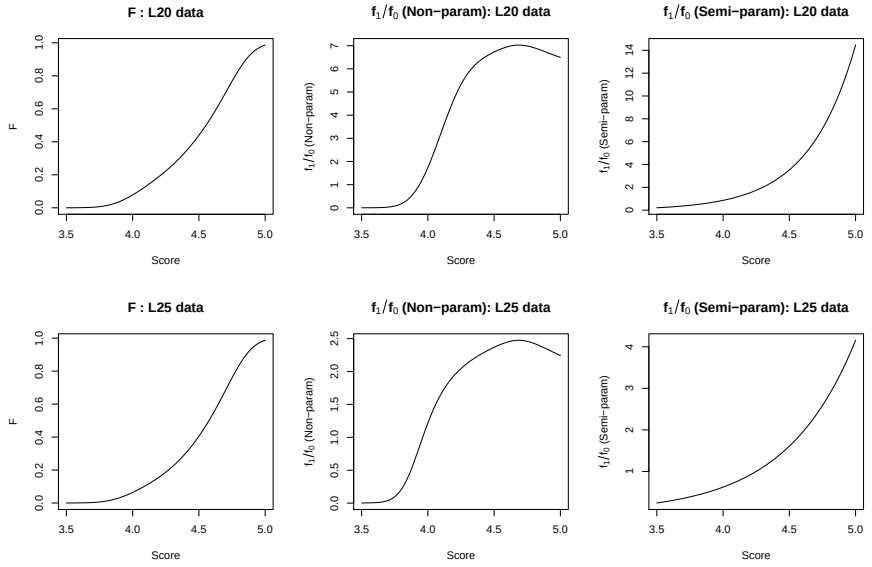

FIG. 7. *Comparison of F, the ratio f_1/f_0 estimated non-parametrically, and the ratio f_1/f_0 estimated semi-parametrically, for the non-simulated data sets. L20 data includes all reads of length 20 or more, and L25 data includes all reads of length 25 or more.*

the hope that further improvements of our overall method may be possible once this estimation is improved.

Regardless, there are clear advantages to incorporating alignment scores into an allocation method, as shown by our strong attenuation of the biases present in the existing methods used for allocation. While these results are to some extent dependent on characteristics of the particular context, for example, sequencing methodology and read length, we remain confident that our method can be applied beneficially in other contexts as well.

REFERENCES

[1] D. R. Bentley et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 456(2008), pp. 53–59.

[2] T. D. Harris et al., *Single-molecule DNA sequencing of a viral genome.* Science, 320(2008), pp. 106–109.

[3] M. Margulies et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 437(2005), pp. 376–380.

[4] N. Cloonan et al., *Stem cell transcriptome profiling via massive-scale mRNA sequencing.* Nat Methods, 5(2008), pp. 613–619.

[5] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, *Mapping and quantifying mammalian transcriptomes by RNA-seq.* Nat Methods, 5(2008), pp. 621–628.

[6] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, *Serial analysis of gene expression.* Science, 270(1995), pp. 484–487.

[7] T. Shiraki et al., *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.* Proc Natl Acad Sci U S A, 100(2003), pp. 15776–15781.

[8] J. Reinartz et al., *Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms.* Brief Funct Genomic Proteomic, 1(2002), pp. 95–104.

[9] J. B. Kim et al., *Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy.* Science, 316(2007), pp. 1481–1484.

[10] G. J. Faulkner et al., *A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE.* Genomics, 91(2008), pp. 281–288.

[11] H. Li, J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res 18(2008), pp. 1851–1858.

[12] D. G. Fisk et al., *Saccharomyces cerevisiae S288C genome annotation: a working hypothesis.* Yeast 23(2006), pp. 857–865.

[13] F. C. Holstege et al., *Dissecting the regulatory circuitry of a eukaryotic genome.* Cell, 95(1998), pp. 717–728.