Annotation-Free Estimates of Gene-Expression from mRNA-Seq

Elizabeth Purdom

Department of Statistics
University of California

September 5, 2012

Abstract

Motivation: mRNA-Seq experiments provide an impressive array of information about the transcriptome of a sample. Yet in organisms that undergo alternative splicing, correctly estimating the standard measures of gene expression can be a complex problem because of complications caused by alternative splicing. The simple estimate based on the number of fragments aligning to a gene has the potential to be biased. Many methods now exist that estimate individual isoform estimates, which can then be combined to give accurate gene expression estimates. However, isoform estimates require either knowledge of the transcriptome or the ability to accurately predict it. Yet many mRNA-Seq experiments are run on organisms with no known genome, much less a transcriptome. In addition, these methods are computationally intensive and usually require access to the raw reads, making them difficult to use for researchers who want to analyze large numbers of samples.

Results: We examine estimates based on summaries that are easy to obtain and analyze, specifically methods based on counting the number of sequenced fragments that overlap exons. We compare these methods to isoform-based gene estimates. We show that in simulated data our gene estimation methods based on exon counts give reasonable gene estimates in the presence of moderate alternative splicing. We compare all of these methods on two mRNA-Seq datasets and observe little difference between any of the methods. In which case, simple count-based methods can be sufficient and allow the experimenter to make use of statistical techniques that appropriately account for the biological variation between samples.

Contact: epurdom@stat.berkeley.edu

1 Introduction

mRNA-Seq data is rapidly becoming the platform of choice for high throughput mRNA studies because of its additional sensitivity. Sequencing data also allows in a single experiment the examination of many questions that would have previously required specialized microarrays, such alternative splicing, mutation analysis, or allele specific expression. However, gene expression estimates often remain of primary interest. When the sequenced organism displays no alternative splicing, estimates of gene expression estimates are straightforward and are based on the number of sequenced fragments that are observed to have come from the gene. Using these counts, statistical approaches of varying levels of complexity can be used to test for differential gene expression (Robinson and Smyth, 2007; Anders and Huber, 2010; Zhou et al., 2011). Issues regarding normalization and bias-correction have arisen that are specific to the sequencing technology and are continuing to be addressed (Dohm et al., 2008; Oshlack and Wakefield, 2009; Bullard et al., 2010; Hansen et al., 2010; Li et al., 2010), but in general there is a fairly straightforward path toward reproducing the types of analyses that are traditional in microarray studies.

However, the presence of alternative splicing makes even a gene-centric analysis of mRNA-Seq data conceptually more difficult. In this case, the biological definition of gene-expression would usually be understood to be the total amount of expression across all isoforms of the gene. The count of the number of fragments mapping to the gene is expected to be an underestimate of gene expression when alternative splicing is present in the gene, and the level of this bias depends on the alternative splicing behavior of the gene.

Another route to obtaining gene expression estimates is to estimate the expression of the individual isoforms and to estimate the gene-expression estimates from the isoform estimates. Isoform expression levels cannot be reliably estimated from simple counts of fragments overlapping the isoform, and statistical methods have been proposed to estimate isoform levels that account for the probability that a read originating from the different isoforms (Denoeud et al., 2008; Jiang and Wong, 2009; Trapnell et al., 2010; Richard et al., 2010; Salzman et al., 2010). These methods rely on a model for how reads are distributed across a given isoform, from a simplistic model of uniform distribution of reads across an isoform to more complicated models and corrections (Li et al., 2010). These are conceptually straightforward estimates that are consistent if the model used in estimation is correct and if the set of isoforms is known and is identifiable (Hiller et al., 2009). However these methods require knowledge of the set of possible isoforms for a gene based on previous annotation; alternatively, the set of isoforms must be estimated from the data (Trapnell et al., 2010; Guttman et al., 2010).

From a practical perspective, another limitation of isoform deconvolution methods is that they require access to the raw alignments and are fairly computationally intensive. With large sequencing projects, hundreds of such samples must be processed, and acquiring the raw data can be onerous, particularly since there are important privacy concerns regarding access to sequencing data. We consider here simpler estimates of gene expression, based on simple count summaries, and evaluate whether they are viable alternatives to the isoform-based methods. The most obvious such method is the total number of sequenced fragments within the boundaries of a gene, mentioned above.

We also propose and evaluate methods of estimation of gene expression based on more detailed exon-level measurements of expression. The estimates we propose based on exon summaries are compelling because they are easy to store and exchange but provide more detailed information about the internal behavior of the gene, for example in determining uniformity of coverage or other quality control problems. The only knowledge of the splicing behavior that is needed is the location of the exon boundaries. These boundaries can be based on previous annotation, on junctions and exons discovered de novo (Bona et al., 2008; Trapnell et al., 2009; Bryant et al., 2010), or a combination of the two. Compared to isoform prediction, prediction of exon boundaries is a much easier task. Because the data unit is at the exon level, rather than individual reads, the computational complexity of these estimation methods is trivial compared with that of estimating individual isoforms.

Similar to microarray methods, our methods for summarizing exon-counts into a single gene estimate rely on robust methods of estimation; the robust methods can hopefully account for both unknown alternative splicing as well as possible errors in our modeling of the data. However, to combine these exon summaries into an estimate of gene expression requires normalizing the different length exons to the correct scale, and we also derive more sensitive techniques for normalizing exon counts.

We show that that for moderate amounts of alternative splicing in the sample, the gene estimates we investigate are accurate and similar to isoform-based estimates, which are more computationally expensive and require more knowledge of the existing gene structures. In many settings, such count-based estimates may even be preferred, such as when the data comes from organisms for which there is no existing annotation on the set of possible isoforms. Even when there is comparatively complete transcriptome information, the low computational overhead can make count-based gene expression estimates useful for initial quality control checks of the data or for when there are large numbers of samples. Count-based summaries, including our exon-based summaries, can also be easily incorporated into existing statistical frameworks for differential expression analysis (Robinson and Smyth, 2007; Anders and Huber, 2010), which have been shown to be important in accounting for the biological variability between samples.

2 System and Methods

Gene expression refers to the total amount of mRNA transcripts originating from a gene in a sample library, where the total amount means the total over all transcripts in a gene. We let μ be the expected number of transcripts from a gene if sampling full mRNA transcripts from the library at the given sequencing depth[1] (and thus μ will depend on the depth at which the sequence is sampled; μ can be normalized by dividing by the sequencing depth or with more accurate normalization methods, see Bullard et al. (2010)). We will refer to the quantity μ as the expression level of the gene. We note that in fact the library is fragmented before sequencing, so that the expected number of *fragments* (or equivalently read counts) that actually comes from the gene in an mRNA experiment is not equivalent to μ; rather we define μ as the rate at which transcripts from the gene would have been sampled from the original mRNA pool before fragmentation.

Similarly, let λ_t be the expression level of an isoform t associated with the gene. Then the gene expression level μ is given by sum of all expression over all transcripts, $\mu = \sum_t \lambda_t$. Isoform deconvolution techniques described in the introduction explicitly try to estimate the λ_t and then estimate μ by summing the λ_t.

[1] Up to an unidentifiable constant that is the same across samples

The simplest alternative gene estimate is a count of the number of fragments mapping within the gene boundaries, divided by the total length of the gene. If there is only a single isoform from a gene, then this is the same as the isoform deconvolution estimate, assuming a simple model of fragment distribution (see Supplementary Text). In what follows, we call this estimate the 'total' gene estimate. The total gene estimate is proportional to the RPKM estimate (Mortazavi et al., 2008), though RPKM estimates also divide by the total sequencing depth so as to normalize across samples.

The total gene estimate will underestimate the gene expression if there is alternative splicing since there will be less fragments than expected coming from those regions that are spliced out; similarly, if there are regions of poor quality, such as regions of poor mappability, these can negatively influence the gene estimate. Exon counts, based on a smaller portion of the gene, have the potential to pinpoint the location of such irregularities and in doing so could be excluded from the estimate of gene expression. We propose robust methods to seamlessly downweight exons that are poor predictors of gene expression (whether due to quality problems or alternative splicing). The basic premise of estimating gene expression from exon-level expression is that many of the exons will be *constitutive* (contained in *all* the transcripts expressed by the gene), and thus will be good estimates of μ. Clearly, alternatively spliced exons, which are not contained in all of the expressed isoforms, will not follow this pattern – they will will be expressed at a level less than μ depending on both the set of isoforms in which they are contained and the expression level of those isoforms – therefore exon-based methods will handle only limited amounts of alternative splicing. Importantly, to be constitutive for these purposes, it is only necessary that the exon be contained in all of the isoforms actually expressed in this sample, not all possible isoforms that the gene can produce.

In what follows, we denote our exon-level counts for an exon τ, as Y_τ.

2.1 Normalizing Exon Counts to a Standard Expression Level

Our premise for exon-based estimates implies that we can convert our exon-level counts Y_τ into reasonable estimates of the exon's expression, defined as the expression of the transcripts that include the exon τ: $\mu_\tau = \sum_{t:\tau\in t} \lambda_t$ (where in the case of constitutive exons, $\mu_\tau = \mu$, the gene expression for the sample).

It is clear that many factors contribute to different exons having a different level of expected overlapping fragments. The most obvious is due to the different length of exons. Because of fragmentation of the isoforms before sequencing, longer regions are expected to represented at higher rates than shorter regions, and therefore the exon levels must be normalized. This has been long recognized, and measures such as RPKM (Mortazavi et al., 2008) divide by the region length for this reason. More precisely, we would like there to exist a fragmentation normalization constant c_τ so that $E(Y_\tau)$ could be written as $c_\tau \mu_\tau$, in which case Y_τ could be corrected by c_τ to obtain an underlying expression of that exon, such as is done with RPKM, where $c_\tau \propto |\tau|$.

In fact, the expected number of fragments overlapping an exon depends not just on the length of the exon. Rather, the expected overlap depends on 1) the total number of possible locations for a fragment that would result in the fragment overlapping the exon and 2) the varying probability of fragments in these locations to get sequenced. The total number of possible locations is roughly proportional to $|\tau|$, as we discuss further below, which is why generally the length of the exon is roughly the right normalizing factor to bring exon-level expression to the correct value. But an exon's position within the transcript – namely its distance to the termini of the transcript – will influence the possible number of locations that overlap an exon. This is because there are fewer possible locations for a fragment that result in the sequenced portion of the fragment overlapping an exon at the terminus of the transcript, as compared to an exon in the middle of the transcript (this will be true for both paired and single-end sequencing).

There has similarly been a great deal of evidence that different locations have varying probability of being sequenced or mapped (Dohm et al., 2008; Hansen et al., 2010). We will for convenience assume fragments are equally likely to be sequenced from any starting position in a transcript, but we will return to this assumption in the discussion.

Under this uniform assumption, we can make more precise what is a reasonable definition of the fragmentation normalization constant. We assume that the lengths of fragments that are sequenced follow a known distribution H and that F_t is the expected number of fragments contributed by one transcript of isoform t after fragmentation. Then if each isoform t in the gene is expressed at level λ_t, the expected number of fragments overlapping an exon can be written explicitly (see Supplementary Text),

$$E(Y_\tau) = \sum_{t:\tau\in t} \lambda_t F_t E_H\left(\frac{K_\tau^t(\ell)}{L_t - \ell + 1}\right) = \sum_{t:\tau\in t} \lambda_t a_\tau^{(t)} \tag{1}$$

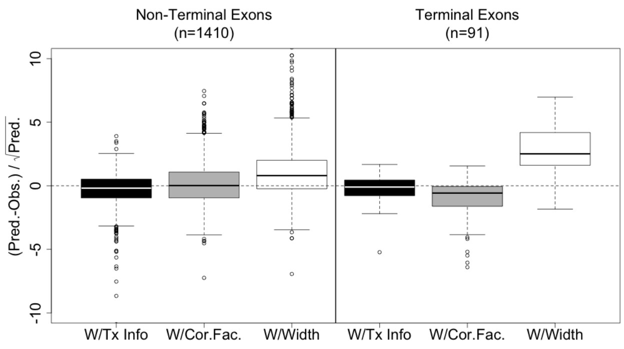


Figure 1: Boxplots comparing the predicted versus the observed number of exon counts for simulated data. Three methods for predicting the expected value are compared: the simple length correction (white), the exon correction factor c_τ described in the text (grey), and the true factor $a_\tau^{(t)}$ based on knowing the complete information for each transcript (black). Exons are split into whether they are at the terminus of the gene or not. For all three methods the expected value is calculated by $\sum_{t:\text{exon}\in t} a_\tau^{(t)}\ \lambda_t$ (where $a_\tau^{(t)}$ is a constant value for all transcripts for the first two methods) and λ_t is the true expression value of the isoform used in simulating the data.

where the sum is taken over all transcripts that contain the exon τ. $K_\tau^t(\ell)$ is the number of distinct starting locations in transcript t such that a fragment of length ℓ overlaps exon τ. $K_\tau^t(\ell)$, as discussed above, is isoform-specific and depends on the location of the exon within the isoforms t, which can vary for different isoforms. We show how to calculate $K_\tau^t(\ell)$ explicitly for paired-end or single-end reads of length r in the Supplementary Text.

This equation shows that even under this simple model $a_\tau^{(t)}$ is isoform-specific and a function of the number of possible locations for a fragment of a possible length, averaged over the distribution of fragment lengths expected. Therefore, there is not a single fragmentation normalization constant c_τ for an exon, i.e. $E(Y_\tau) \neq c_\tau \mu_\tau$.

However, under two plausible scenarios, the value of $K_\tau^t(\ell)$ is independent of isoform: 1) when the exon is far from the either terminus of all the isoforms that contain τ or 2) when the exon is the terminus in all the isoforms containing τ. In these cases, and therefore $a_\tau^{(t)}$ can be reasonably approximated to be the same for all isoforms and gives a fragmentation normalization constant c_τ for the exon. Therefore, in our analysis, we make a simplifying assumption that exons are in one of these two situations and hopefully come close to our goal of correctly normalizing exon counts for most exons. Specifically, we classify each exon as either being the terminus of the *gene* or not and calculate the corresponding c_τ according to one of the two cases described above.

The resulting c_τ is not proportional to the length of the exon, but varies for different length exons, see Supplemental Figure S1. For paired-end data c_τ is roughly equal to $2|\tau|$ for smaller internal exons, $|\tau| + constant$ for longer internal exons, and $|\tau|$ terminal exons; for single-end data, c_τ is $|\tau|$ for internal exons, $|\tau|/2$ for shorter terminal exons, and $|\tau|/2 +$ constant for longer terminal exons (see Supplementary Text for precise equations).

In Supplemental Figure S2 we give a comparison of the values of $K_\tau^t(\ell)$, $E_H(K_\tau^t(\ell))$, and c_τ for a simple example of two isoforms. Our example illustrates that for some of the exons, there can be a large difference between a simple length correction based on τ and our approximation c_τ, particularly for terminal exons. Furthermore, we can see that our simple assumption that allows us to calculate c_τ without isoform knowledge will not be a good approximation for some exons, particularly those that are the terminus of some transcripts but are not the terminus of the gene.

Using simulated data, we compare the exact calculation of $E_H(K_\tau^t(\ell))$ with our c_τ based on the approximations described above. Figure 1 demonstrates how well our approximations of $K_\tau^t(\ell)$ perform compared to the observed exon-overlap counts for simulated data. We can see that by using the more careful prediction of c_τ we do better than the simple length correction for predicting the observed data. We can make the same comparison with real data (Figure 2 and S3), where we use only exons annotated to be constitutive as a proxy for the truth (see Section 2.3 for a description of the data); in this case, the difference in the fit of the estimates only noticeably differs for terminal exons, where it is unclear whether the fit is substantially improved.

The difference is most striking for the terminus exons – where there is the largest discrepancy from the standard length-based correction. They will obviously form a small percentage of exons in any gene. Therefore we do not see a large difference in our gene-estimates and the length-based correction is probably generally sufficient. However,

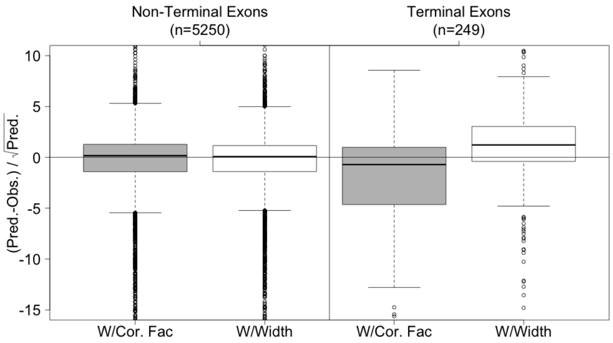


Figure 2: Boxplots comparing the predicted versus the observed number of exon counts for an AML tumor. The expected value was calculated by multiplying the estimated gene expression value by either the correction factor c_τ (grey) or the length correction (white). Only exons annotated as constitutive are considered. The gene estimates are from the symmetric robust method.

we note that there are many situations where exon-level expression are used, such as visualization of the data or for evaluating the uniformity of coverage, in which case more explicit calculations such as we present here are quite useful.

2.2 Methods for estimating gene-expression

The simplest estimate of gene expression is the "total" estimate described earlier. An alternative and equally simple estimate is to take the average of exon counts appropriately normalized, Y_τ/c_τ. If $E(Y_\tau) = c_\tau\mu_\tau$ and $\mu_\tau = \mu$ (i.e. no alternative splicing), the average is also an unbiased estimate of μ. It is generally less efficient, but will often be more robust to alternative splicing or other irregularities than the sum because each exon contributes equally to the estimate of μ regardless of its size or level of expression. Indeed the mean generally performs similarly to our other robust methods described below. In contrast, the total fragment estimate can be dominated by longer exons which have more counts, and will therefore be sensitive to deviations from μ in these exons, for example if the exon is alternatively spliced.

In analogy with microarray methods for estimating gene expression (Irizarry et al., 2003), we can alternatively create a linear model for the log expression data, with the $Y_{i\tau}$ as our observed data, and robustly fit the parameters. Because the $Y_{i\tau}$ are counts, an appropriate way to do this is to create a generalized linear model (GLM), which for Poisson data relies on the model

$$E(Y_{i\tau}) = \theta_{i\tau}(\boldsymbol{\mu}) = \exp(\log c_\tau + \mathbf{x}_{i\tau}^T\boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is our vector of gene expression values. In the case of estimating the gene expression, our linear model is just the simple estimate of the mean μ_i for each observation (so that $\mathbf{x}_{i\tau} = \mathbf{e}_i \in \mathbb{R}^n$ is just a vector with one in the ith position and zero otherwise), but could be expanded to estimate differential expression between groups, for example.

If we assume the $Y_{i\tau}$ are Poisson and find the maximum likelihood estimate of $\boldsymbol{\mu}$, then our estimate is roughly that of the sum estimate, again giving higher priority to those $Y_{i\tau}$ with higher counts. We propose to use a robust estimate to down-weight those $Y_{i\tau}$ whose expected value poorly corresponds to the model, whether due to alternative splicing or other factors. We follow the work of Cantoni and Ronchetti (2001) to create a robust alternative for fitting gene estimates.

If we assume a distribution for the counts where the mean and variance are equal, as with the Poisson, the standard GLM gives an estimating equation of $\boldsymbol{\mu}$ as the solution to

$$\sum_{i,\tau} \psi(r_{i\tau}(\boldsymbol{\mu}))\frac{\nabla\theta_{i\tau}(\boldsymbol{\mu})}{\theta_{i\tau}(\boldsymbol{\mu})^{1/2}} - a(\boldsymbol{\mu}) = 0,$$

where $a(\boldsymbol{\mu})$ is a correction factor for consistency and ψ is a function that weights observations based on their standardized residual from their expected value. For the standard GLM, $\psi(r) = r$, which can allow large outliers to highly influence the estimate, but the more robust version of Cantoni and Ronchetti (2001) downweights the impact of outliers by using the Huber function to cap the penalty that is paid for large deviations in either direction. In the setting of gene estimation, our deviations might be expected to be largely due to some exons being alternatively spliced, in which case the $Y_{i\tau}$ would consistently underestimate μ_i, and thus we also consider an asymmetric penalty function that only caps the penalty for large *negative* deviations.

In what follows, we assume that our exons follow a distribution *within* a sample that has equal mean and variance, in which case the standardized residual is given by,

$$r_{i\tau} = \frac{Y_{i\tau} - \theta_{i\tau}(\boldsymbol{\mu})}{\theta_{i\tau}(\boldsymbol{\mu})^{1/2}}.$$

We note that this addresses only the variations among exons in the same sample, and does not address the variation across samples, which is likely to not have equal mean and variance. The GLM framework could be extended across samples and allow for sample variability by changing the score function to allow for an overdispersion parameter per sample, resulting in a model with equal mean and variance within the sample, but overdispersion across samples.

Note that we do not include an exon effect e_τ in our predictor vector $\mathbf{x}_{i\tau}$. This means that the system of equations can be solved separately for each observation i. There are many reasons to believe that some effects per exon would be shared across samples, such as GC biases or mappability. However, in order to have a identifiable model, an exon effect will require a constraint as to the relationship of the exon effects. The standard choice in constraints is that the exon effects sum to 0 which effectively shifts the average exon effect to the estimate of μ_i. This interferes with our biological interpretation of μ_i as the *total* amount of expression across isoforms, and instead μ_i becomes the average expression across exons. This is also a conceptual problem with evaluating differences in exon effects from a GLM to detect differential alternative splicing between groups: because the exon effect is defined relative to a μ_i that is not total expression, differential alternative splicing of a single exon can be either diffused across the gene and/or be identified to the wrong exon.

2.3 Data Processing

To create simulated sequencing data, we randomly selected gene models from Ensembl annotation for a total of 70 genes and 415 transcripts (see Table S1 in the Supplementary Text). We note that our distribution of genes over-represents genes with large numbers of transcripts compared to the Ensembl annotation where smaller numbers of transcripts per gene are more common. For each gene, we generated true expression rates for the transcript and based on this parameter generated a random number of fragments per transcript using either a Poisson model or a Negative binomial model. The fragments were assigned a length from a truncated normal distribution. The position of the fragment within the transcript was either uniform across the transcript or proportional to the GC content of fragment. The resulting 75bp mate pairs of reads were then aligned to determine all transcripts for which they were compatible and also which exons the reads overlapped (see Supplementary Text for further details).

For real mRNA-Sequencing data, we used two sets of data: samples from commercial-grade Brain and UHR mRNA used in Bullard et al. (2010) and samples from AML tumors sequenced as part of The Cancer Genome Atlas (TCGA) project. For each data sample, we aligned the reads to the GTF annotation file from Ensembl 57 using TopHat version 1.3.0 (Trapnell et al., 2009) and created exon overlap counts for each using internal programs. The MAQC data are single-end, 35bp reads from Illumina's Genome Analyzer II, while the AML tumors are paired-end, 50bp reads.

Using the transcript annotation, we defined "exons" to be a contiguous region of the genome where all of the bases in the exon are always either present or absent in the same set of transcripts. We note that there are different possible ways to count the number of fragments that contribute to the counts per exon. We choose to count any fragment whose sequenced portion overlaps the exon (see Supplementary Text S3); in this way there is no difficulty in counting fragments that overlap multiple exon regions.

For the AML samples, we also aligned the reads directly to the transcripts provided by Ensembl using Bowtie (Langmead et al., 2009) and created genes estimates via isoform deconvolution. In order to make the gene estimation methods comparable, we only considered fragments that were successfully aligned by both alignment strategies. We also only considered genes that contained no multiply-mapped fragments in either mapping strategy; genes were also filtered to only protein-coding genes with non-zero overlap counts. This resulted in 20M aligned fragments and 2,000

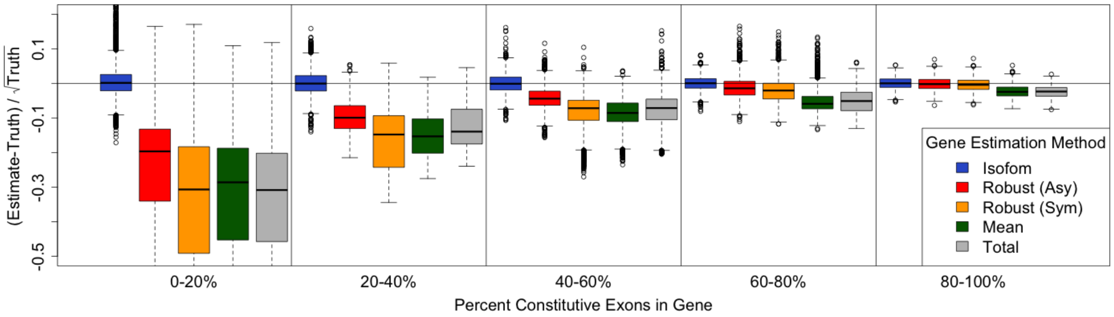


Figure 3: Boxplots of the accuracy of the estimates of gene expression in simulated data. The 70 simulated genes are divided into five categories based on the proportion of exons in the gene that are constitutive in the gene and within those categories, the individual simulations are pooled together across genes. Accuracy is measured with χ^2-residuals: the difference between the estimate and the truth, divided by the square-root of the truth.

genes, depending on the sample considered. When the analysis required normalization, we used the Upper-Quartile method of Bullard et al. (2010). See Supplementary Text S2 for more details regarding the data processing.

3 Implementation

3.1 Simulations

We first evaluate our gene-estimation method using simulated data on randomly selected gene models described in Section 2.3. Using this data, we estimate gene expression by 1) estimating λ_{it} using our implementation of a Poisson deconvolution, 2) the robust GLM(s) described above, and 3) the total gene count (see Supplementary Text for a full description). In the simplest case where the reads are precisely Poisson and all of the annotation is known exactly, we see in Figure 3 that the robust methods give reliable estimates of gene expression when there is moderate alternative splicing (60% or more of the exons in a gene are constitutive) and out-performs the total counts estimate, which is biased downward. The asymmetric penalty function performs slightly better in this context than the other robust alternatives, though as we will see this doesn't necessarily hold with real data that have additional sources of variation. With extensive alternative splicing in a gene there is not enough constitutive signal to recover the gene expression estimates. Thus, all the methods become increasingly downward biased as the percentage of constitutive exons decreases, except for the isoform method which uses the precise transcript information. Indeed, if it is known which exons are constitutive and the robust methods use only those exons as inputs, the robust methods continue to give reliable gene estimates for genes with large alternative splicing, though with increased variance since less of the data is used (Supplemental Figure S4).

When the underlying distribution of fragment counts varies from the assumed Poisson model – for example with the reads being distributed as a Negative Binomial – the variability of all of the estimates increases (Supplemental Figure S5). In this case, the improvement in accuracy observed for the isoform deconvolution is somewhat diminished relative to the variability of its estimates. Similarly, when the position of the fragment depends on the fragment's GC content, the robust estimates perform closer to that of the isoform deconvolution for moderate splicing.

In the simulations described above, the gene expression estimates given by the isoform deconvolution assume the annotation is accurately known. If some isoforms are missing or are misspecified in the annotation, the isoforms can be poorly estimated and sometimes the resulting gene expression estimates are also affected. In our simple simulation, we see that when an important isoform is missing from the annotation (e.g. an isoform representing 10% or more of the total gene expression), isoform deconvolution can also result in severely downward-biased estimates of gene expression. In our simulations, however, missing isoforms tend to influence the gene expression most strongly when there is also a low percentage of constitutive exons in the gene. Supplemental Figure S6 demonstrates this phenomena for three specific genes, but the pattern was seen generally across all of the genes. While this is a simplified setting, it suggests that in the the setting where the robust exon-level methods are working well, gene expression estimates from isoform deconvolution are also robust to mild problems in the annotation.

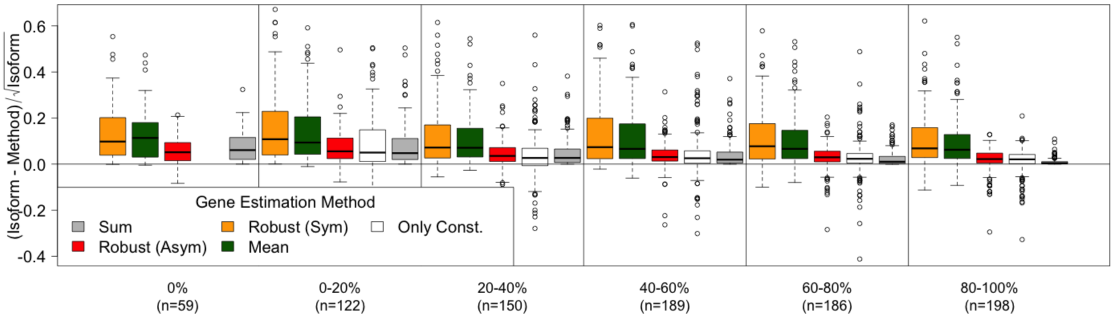


Figure 4: Boxplots of the difference in the isoform gene estimates from the count-based gene estimates after UQ normalization of the estimates. Genes are grouped based on the proportions of exons in the gene that are annotated to be constitutive. For comparison an exon-based estimate (asymmetric Robust GLM) using only annotated constitutive exons as input is shown in white.

3.2 Real Data

We apply the competing gene estimation methods on two sets of real data: 1) samples from commercial-grade Brain and UHR mRNA used in Bullard et al. (2010) with corresponding PCR data available from the MicroArray Quality Control (MAQC) study (Shi et al., 2006) and 2) samples from AML tumors sequenced as part of The Cancer Genome Atlas project.

The differences between the count-based methods is generally small, as is their difference from the isoform deconvolution in the AML sample, see Supplemental Figure S9. In particular, the total fragment gene estimate is generally no different from the isoform deconvolution estimate, perhaps suggesting that there are low levels of alternative splicing in the AML samples we examine. In those situations where there is a difference, the total fragment estimate under-estimates the gene expression as expected. The other exon-based count methods also under-estimate gene expression, as compared to the isoform-deconvolution, and generally do not match the isoform-based estimates as closely as the total fragment estimate.

In practice, it is not the absolute measurement of gene expression that is most important, since experiments comparing multiple samples must be normalized to a common level (Mortazavi et al., 2008; Bullard et al., 2010; Robinson and Oshlack, 2010). After normalizing the gene estimates, we see that as in our simulations, moderate levels of annotated splicing result in similar bulk gene expression estimates, with the symmetric robust GLM being perhaps slightly better than the other exon-based methods. With large amounts of alternative splicing, all of the methods tend to under-estimate the gene expression, though when only annotated constitutive exons are used as input this problem is ameliorated for the symmetric robust GLM. The more noticable difference in the methods occurs in the variability of their difference from the isoform-deconvolution estimates. The symmetric robust GLM, while in average tracking the isoform estimates, also shows the greatest variability.

We can also consider how well the gene estimates predicted the observed exon counts by calculating the expected value of the exon counts using the estimated gene expression given by a method. However, in order to convert the gene expression into the scale of exon counts, we must make use of a fragmentation normalization constant as well. After doing so, we see evidence of the effect of the choice of the fragment normalization constant (Supplemental Figures S7 and S8). Surprisingly, the isoform-deconvolution has the poorest fit to the exon counts for the AML data for both $c_{\mathcal{T}}$ and the length normalizations of the exon, with the isoform predictions overestimating the expected level of exon counts even for genes with small levels of alternative splicing. If we trust the isoform estimates, this indicates that both normalization constants are too large and predict too many overlapping reads. In both the AML and MAQC data, the predicted exon counts from the symmetric robust method best fits the observed exon counts across a range of alternative splicing levels for either choice of our fragmentation normalization constant (if the same normalization constant was used in estimating the gene effects).

In evaluating differential expression between the samples, neither the AML tumors or the MAQC data showed any substantial differences between the methods (Supplemental Figures S11,S12, S13), nor did they show any difference in correlating to the gold-standard PCR data. This is not surprising, since already we saw that normalizing the estimates made the estimates of gene expression basically equivalent between methods. Furthermore, any shared biases or shifts in gene estimation would cancel out for differential expression; only differential alternative splicing between samples would have an effect. In short, for the most common application of gene expression estimates – comparison of expression analysis between samples – the differences between the methods is irrelevant for the bulk of genes we examined, assuming they are correctly normalized. Even using the total fragment count does not show a measurable difference in estimates of differential expression for this data for most genes.

We emphasize that since we categorize genes only by with levels of *all possible* alternative splicing known in our annotation, the percentage of actually occurring alternative splicing in our sample is likely to be much less than that indicated by the annotation. Indeed, many of the genes may show little or no alternative splicing in our samples, so the annotated level of alternative splicing are just indications of the true alternative splicing.

We also note that using only those reads from constitutive regions generally results in a large loss in data. This loss can result in significant reduction in power for GLM-based methods such as Robinson and Smyth (2007); Bullard et al. (2010); Anders and Huber (2010). In the samples we investigated, we saw little improvement in restricting the analysis to constitutive measurements.

4 Discussion

Exon expression levels are an intuitive way to summarize the large amount of data produced by mRNA sequencing experiments, but for many experimenters, the question of interest remains gene expression estimates. We give methods for creating gene expression estimates from exon expression summaries, and show that for moderate amounts of alternative splicing in the sample, the estimates will be generally reliable. Indeed for most real data sets, we expect that there will be little difference in the estimates for most purposes. We also give a conceptual basis for calculating the correct normalization of the exon counts to account for the actual number of fragments that can be in the exon, a consideration particularly important for exons at the terminus of transcripts. The impact on our gene estimation is not large, since genes consist mostly of non-terminal exons, but the principles of normalization can be important for other aspects such visualization and quality-control of mRNA-Seq, which is often done using exonic regions of the genome.

We demonstrated our ideas with some simplified assumptions, but as frequently alluded to, more complicated models and assumptions can be incorporated which would only affect the calculation of the normalization constant per exon, c_τ. An obvious generalization can come from data-derived distribution of the fragment lengths. Other biases, such as GC-bias and mappability could be incorporated as well but to be exact will generally require transcript information since they require information about the entire fragment. Alternatively, these types of bias could be fit into the larger GLM and the appropriate adjustments made via the GLM. Regardless, the calculation of c_τ does not require processing all of the raw reads, but rather parameters of the exon τ, such as position within the gene or transcript.

While our estimates of gene expression are reasonable for moderate amounts of alternative splicing, they do break down in simulations when there are not enough constitutive exons to give a reliable signal of the overall gene expression. In contrast, explicit deconvolution of the isoform estimates can often give good estimates of gene expression regardless of the level of alternative splicing, particularly when there is reasonably complete annotation of the genes; indeed our data examples are all with human samples with a relative wealth of knowledge. And while the differences are striking in simulation, in our examinations of real data the differences are small, probably due to lower levels of alternative splicing, and generally don't effect our estimates of differential expression. Furthermore, many experimenters are using mRNA-sequencing for organisms where there is no annotation about the possible isoforms.

Another interest in having gene expression estimates is to be able to do quick, initial quality control checks of the data as it is being produced. Count-based gene expression estimates can be quickly computed from exon summaries, giving a fast avenue to performing quality control, compared to isoform deconvolution. Similarly, as large sequencing data is produced (e.g. The Cancer Genome Atlas project) it becomes infeasible for interested researchers to download and process all of the raw sequencing data; privacy concerns also make this a complicated endeavor. Smaller summaries, such as exon summaries, are much much more portable, and will allow researchers to refit and compare the effects of different models for large numbers of samples without needing the sensitive raw data files on which isoform deconvolution relies.

Furthermore, previous work has shown the importance of taking into account the biological variability between samples, using models of variation across the samples, such as the negative binomial (Robinson and Smyth, 2007; Anders and Huber, 2010), but such methods start with counts as measures of expression, making the results from isoform deconvolution inappropriate for usage directly. Our gene expression estimates are based on standard GLM models, and thus can be adapted to the differential expression work that has been created to estimate dispersion parameters for count data from mRNA-Seq.

5 Acknowledgements

We wish to thank Marla Johnson for helpful code used in the creation of the exon summaries and Terry Speed for his valuable comments on earlier versions of this manuscript.

Funding: This research was partially funded by NIH grant U24 CA143799 and NSF grant DMS-1026441.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*.

Bona, F. D., et al. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**(16), i174.

Bryant, D., et al. (2010). Supersplat–spliced RNA-seq alignment. *Bioinformatics*, **26**(12), 1500.

Bullard, J. H., et al. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**, 94.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**(455), 1022–1030.

Denoeud, F., et al. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol*, **9**(12), R175.

Dohm, J. C., et al. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Research*, **36**(16), e105.

Guttman, M., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, **28**(5), 503–10.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, **38**(12), e131.

Hiller, D., et al. (2009). Identifiability of isoform deconvolution from junction arrays and rna-seq. *Bioinformatics*.

Irizarry, R. A., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–64.

Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-seq. *Bioinformatics*, **25**(8), 1026–32.

Langmead, B., et al. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, **10**(3), R25.

Li, J., Jiang, H., and Wong, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol*, **11**(5), R50.

Mailman, M. D., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, **39**(10), 1181–1186.

Mortazavi, A., et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**(7), 621.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, **4**, 14.

Richard, H., et al. (2010). Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Research*, **38**(10), e112.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3), R25–R25.

Robinson, M. and Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881.

Salzman, J., Jiang, H., and Wong, W. H. (2010). Statistical modeling of RNA-SEQ data. Technical Report BIO-252, Division of Biostatistics, Stanford University, Palo Alto.

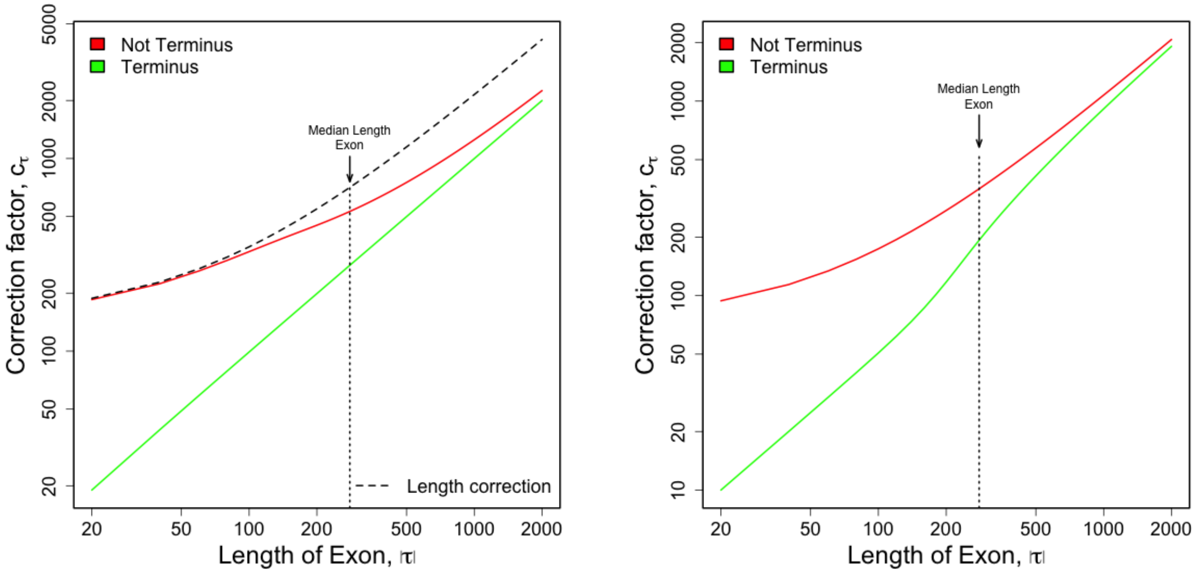
Shi, L., et al. (2006). The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**(9), 1151.

Trapnell, C., Pachter, L., and Salzberg, S. (2009). Tophat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**(9), 1105.

Trapnell, C., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511.

Zhou, Y. H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics (Oxford, England)*, **27**(19), 2672–2678.

Supplemental Figures



(a) Paired-End

(b) Single-End

Figure S1: Plot of c_τ as a function of length of τ for a paired-end and b single-end data. The correction for internal exons is given in red, that of terminal exons in green. For the paired-end data, the length-based correction of $2(|\tau|+r)$ is also plotted (dashed black line); for single-end reads, the length correction is exactly equal to c_τ for internal exons. The read length for both is assumed to be 75bp and the fragment length distribution is as in the simulation: a $N(250, 50^2)$ distribution truncated to be within the interval $[75, 500]$.

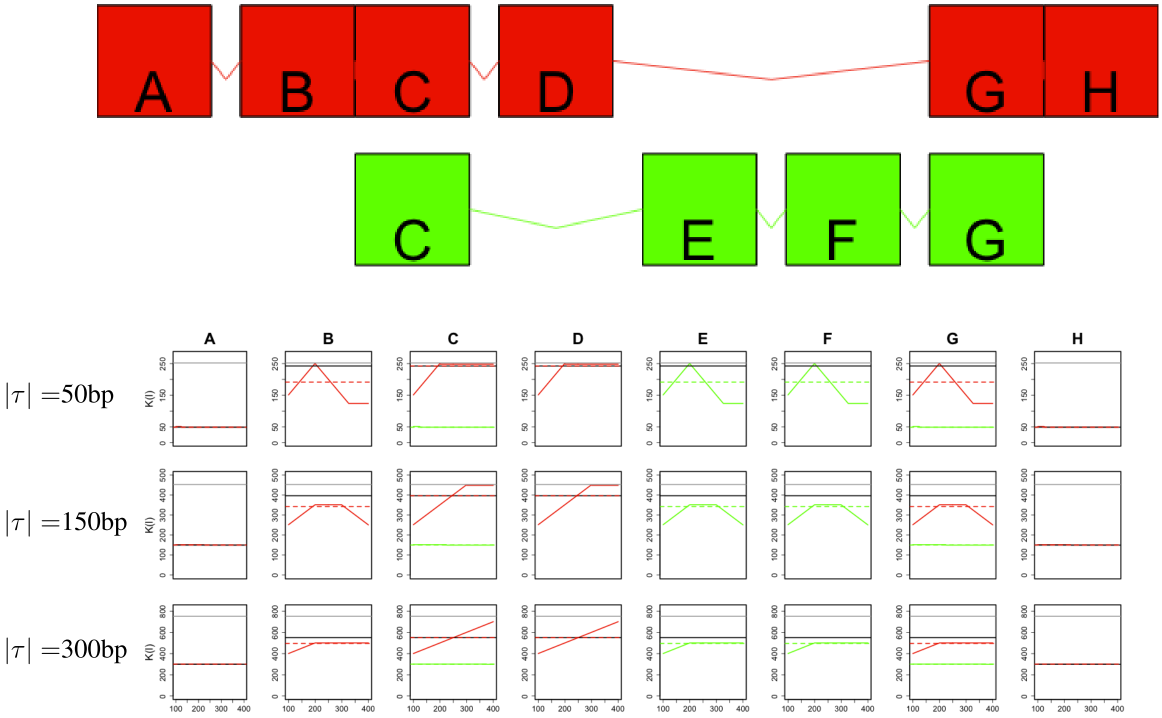


Figure S2: Simple example of the number of fragments overlapping an exon, as a function of the length of the fragment and the transcript. : the simple gene model with two transcripts and exons labelled A-H. plots of $K_t^\tau(\ell)$, the number of overlapping fragments, against the length of the fragment ℓ, assuming 75bp paired-end reads. Each column of plots corresponds to the K_t^τ for the exon indicated, with the red and green line corresponding to the red and green transcript, respectively. Each row corresponds to changing the length of the exon, $|\tau|$ to a length equal to 50, 150, or 300 base pairs. Because $K_t^\tau(\ell)$ is a function of the distance of the exon to the end of the transcripts, these exon-length values correspond only to the exon length of the exon considered in the plot; the distance of the exon to the end of the transcript is kept to be the same across all rows and is equal to assuming the exons separating the exon in question from the terminus have length 250bp.

The red and green horizontal lines show the value $E_H(K_t^\tau(Z))$ for each transcript (red or green) – the number of fragments expected to overlap the exon when averaged across the expected distribution H of fragment lengths (in this case a normal distribution with mean 250 and standard deviation 50, truncated to be between 100 and 500 bp). The horizontal black line shows the approximation c_τ described in the text, where each exon is either the terminus of the gene or not; the dark grey line shows the length correction equal to $2(|\tau| + 75)$, and thus the same for all of the exons.

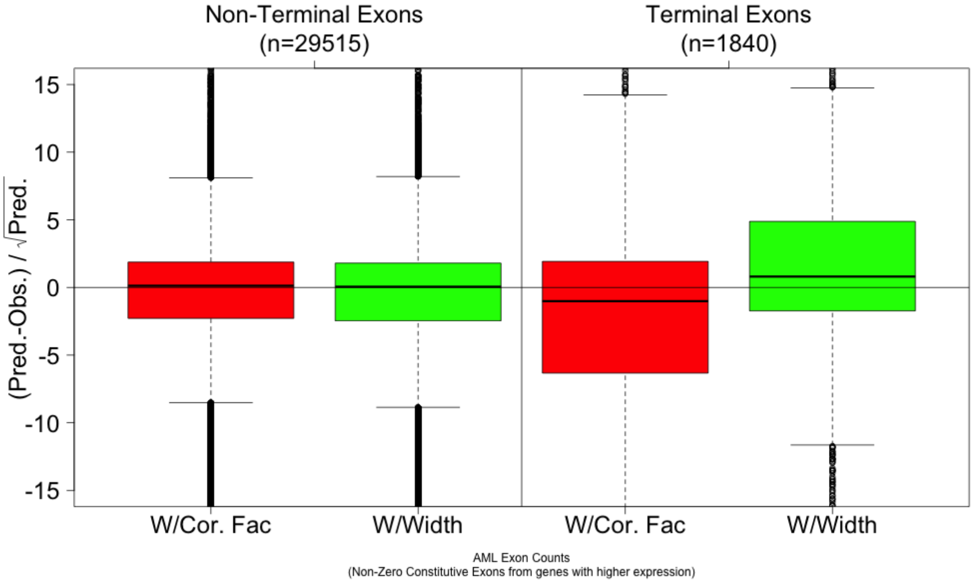


Figure S3: Boxplots comparing the predicted versus the observed number of exon counts for the UHR sample from the MAQC data. See Figure 2 for details. Note that the MAQC data is single-end reads, and therefore $c_\mathcal{T}$ and the length correction are *exactly* the same for the non-terminal genes.

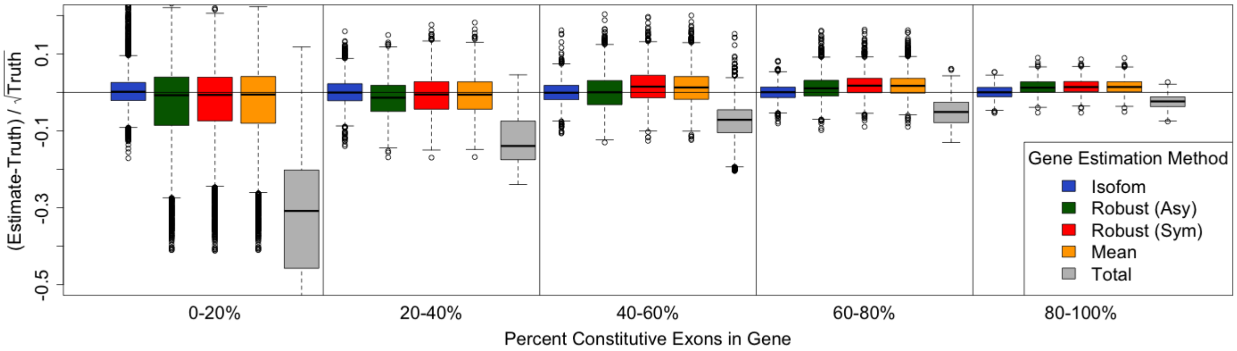


Figure S4: Boxplot of the accuracy of the estimates of gene expression in simulated data using robust methods with only known constitutive exons compared to that of the total count method and the method using isoform deconvolution. See Figure 3 for more details.

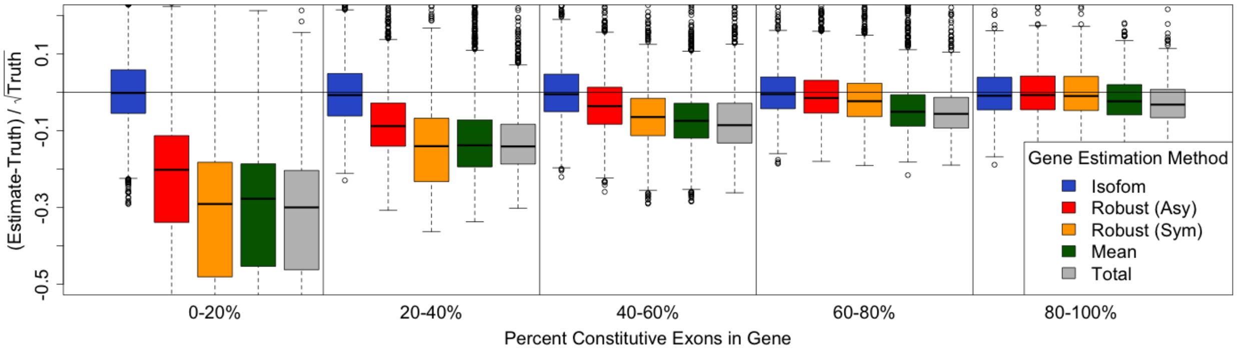


Figure S5: Boxplot of the accuracy of the estimates of gene expression when reads are distributed according to a Negative Binomial distribution in simulated data. See the legend for Figure 3 for more details.

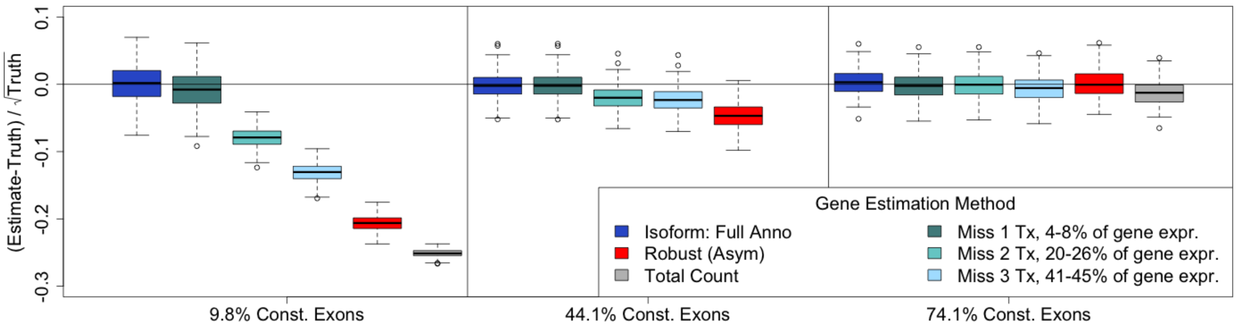


Figure S6: Boxplot of the accuracy of the estimates of gene expression in simulated data using isoform deconvolution when not all isoforms expressed are included in the annotation. Also shown are estimates based on the asymmetric version of the robust GLM method and the total count method. Each panel are simulations from a single gene; the three genes each have five isoforms and differ mainly in the percentage of constitutive exons the gene. For each gene, the x lowest expressed genes where omitted from the annotation, with x ranging from 1-3 isoforms. This resulted in three estimates of gene expression based on annotation that missed 4-8%,20-26%, and 41-45% of the total gene expression, respectively (the range in percentages is due to the fact that the five isoforms were randomly assigned expression values that differed in the three genes and dropping one isoform resulted in slightly different loss in percentages in the different genes).

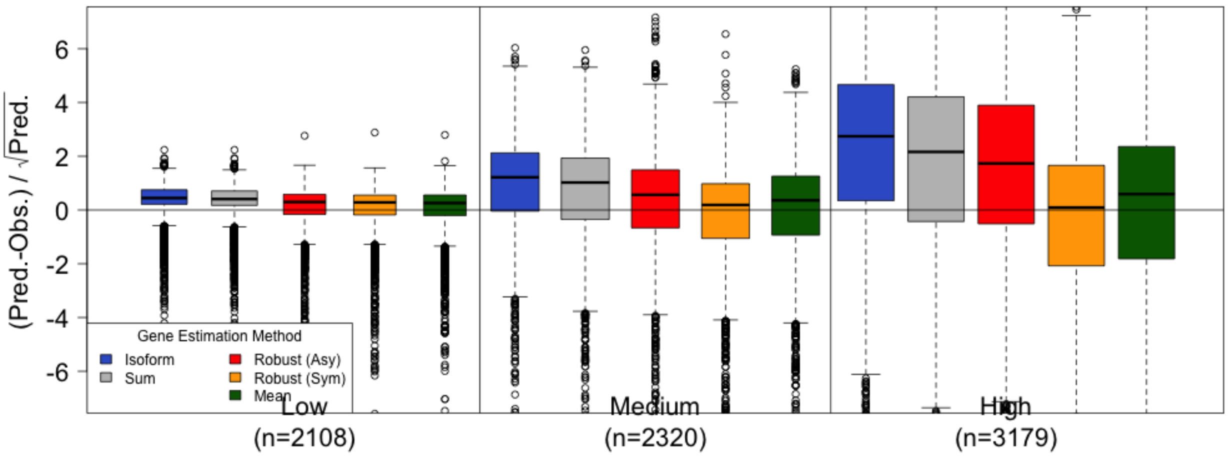


Figure S7: Boxplots of observed exon counts versus that predicted for AML tumor. Predicted levels for the exons are calculated as in Figure **??**. Data is from a single AML tumor sample. Only exons annotated to be constitutive are considered. Exons are grouped based on the estimated gene expression levels of its gene. See Supplemental Figure S8 for the similar plot for the MAQC data

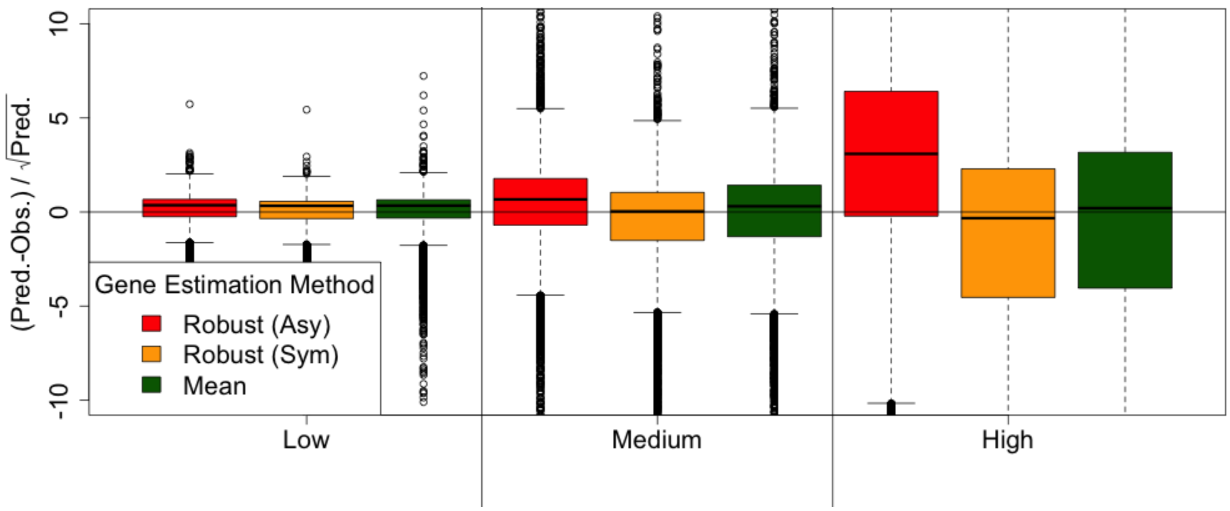


Figure S8: Boxplots of observed exon counts versus that predicted for UHR sample from the MAQC experiment. See figure S7 for details.

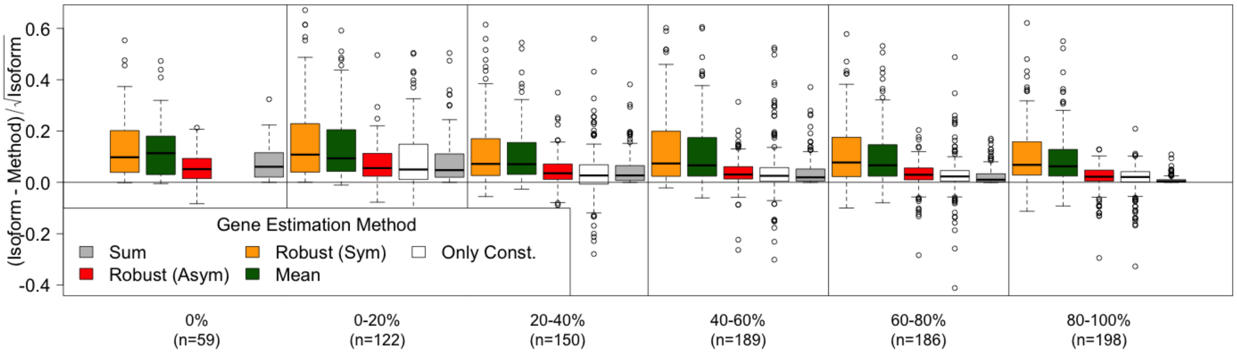


Figure S9: Boxplots of the difference in gene estimates of the exon-based methods of gene estimation, as compared to the gene estimates provided by isoform deconvolution for a single AML tumor. Genes are grouped based on the proportions of exons in the gene that are annotated to be constitutive.

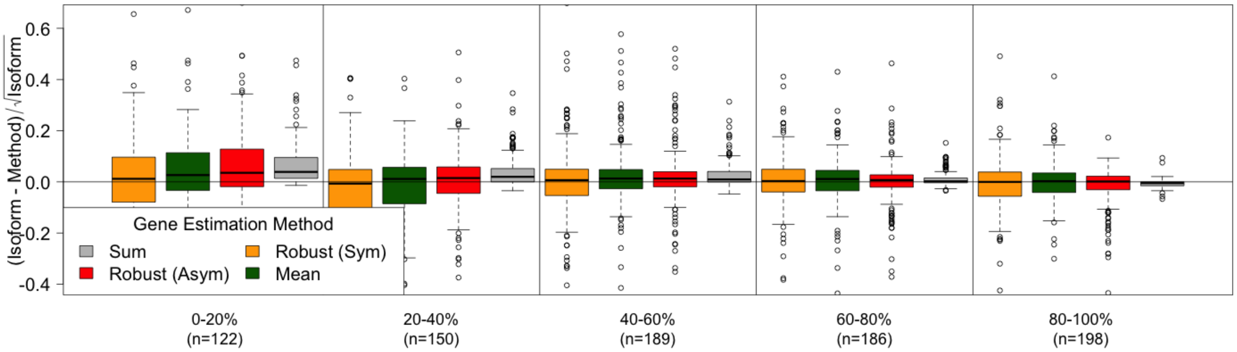


Figure S10: Boxplots of the difference in gene estimates of the exon-based methods of gene estimation, as compared to the gene estimates provided by isoform deconvolution. Shown are the results only using the constitutive exons in the gene estimation for a single AML tumor. See figure S9 for details.

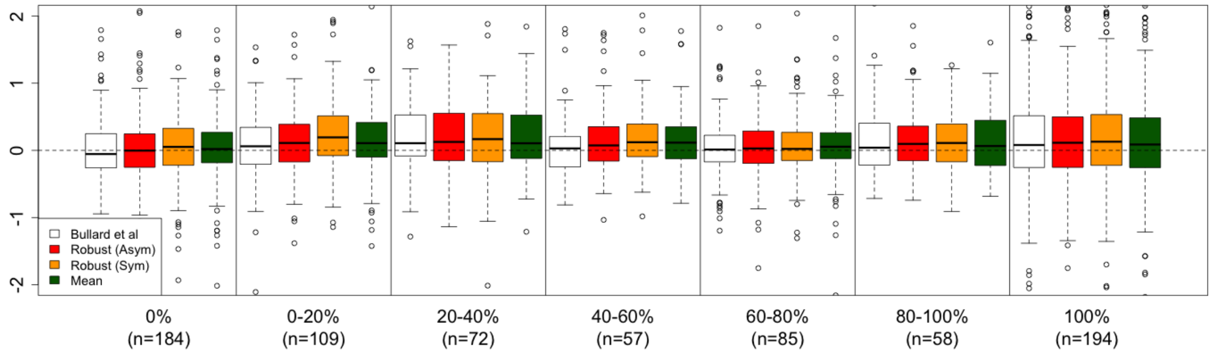


Figure S11: Boxplot comparing the difference in the estimate of differential expression between that of the PCR data and the gene expression technique. Genes are grouped by the percentage of exons annotated to be constitutive in the gene.

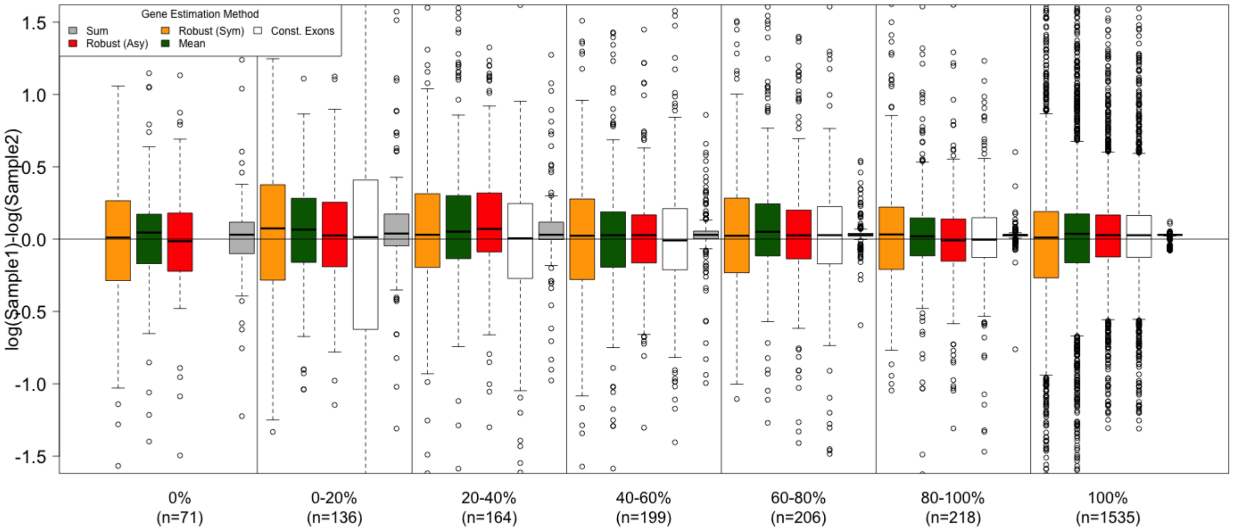


Figure S12: Boxplots of estimates of differential gene expression for two AML samples for different methods of gene estimation.

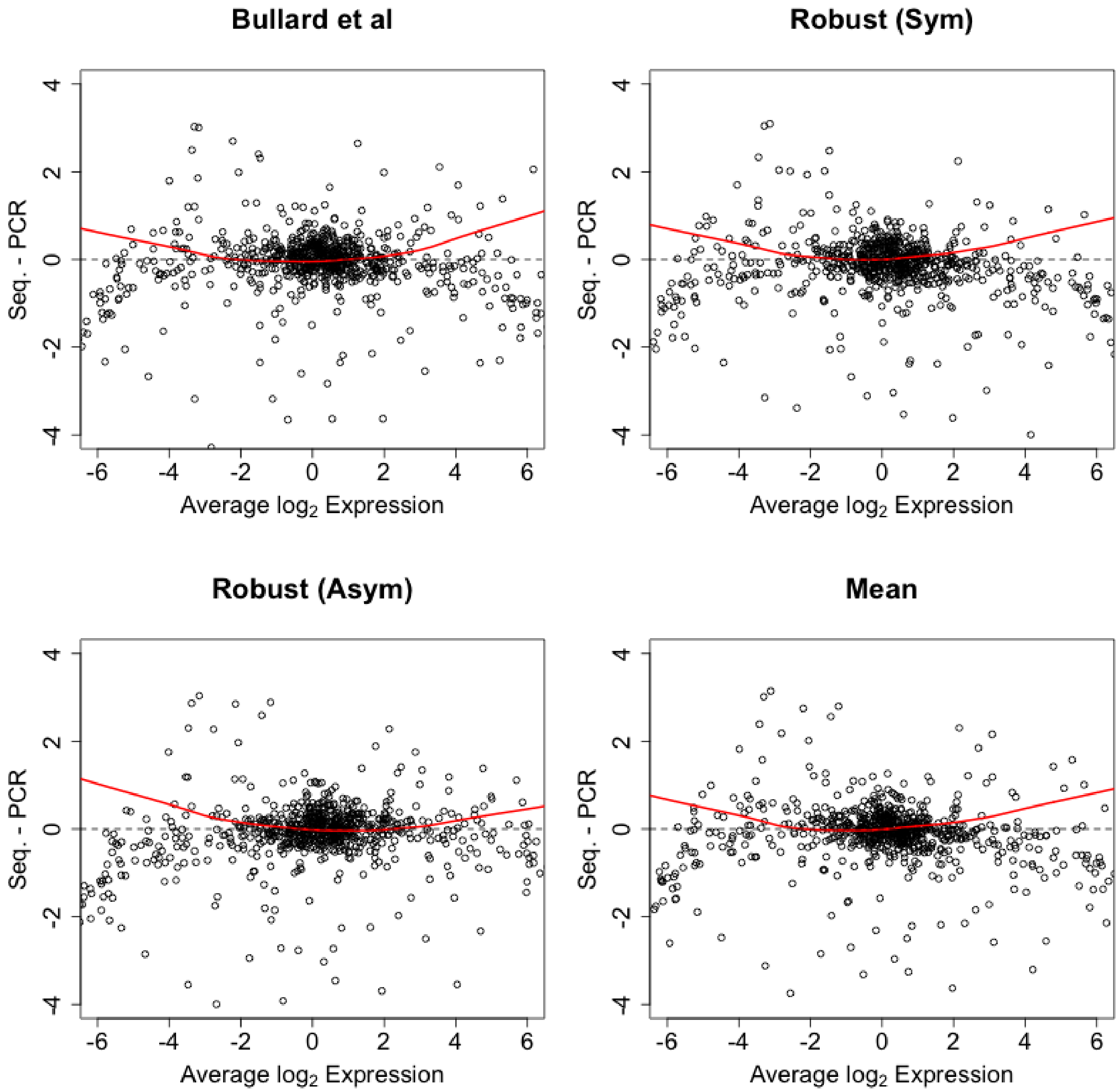


Figure S13: Scatter plots comparing the estimates of differential expression based on the PCR data and different gene expression estimates. The y-axis is the log-difference between the estimates (PCR-Sequencing) and the x-axis is the average log expression. Genes that are absent in either brain or UHR in the original paper Bullard et al. (2010) are not shown.

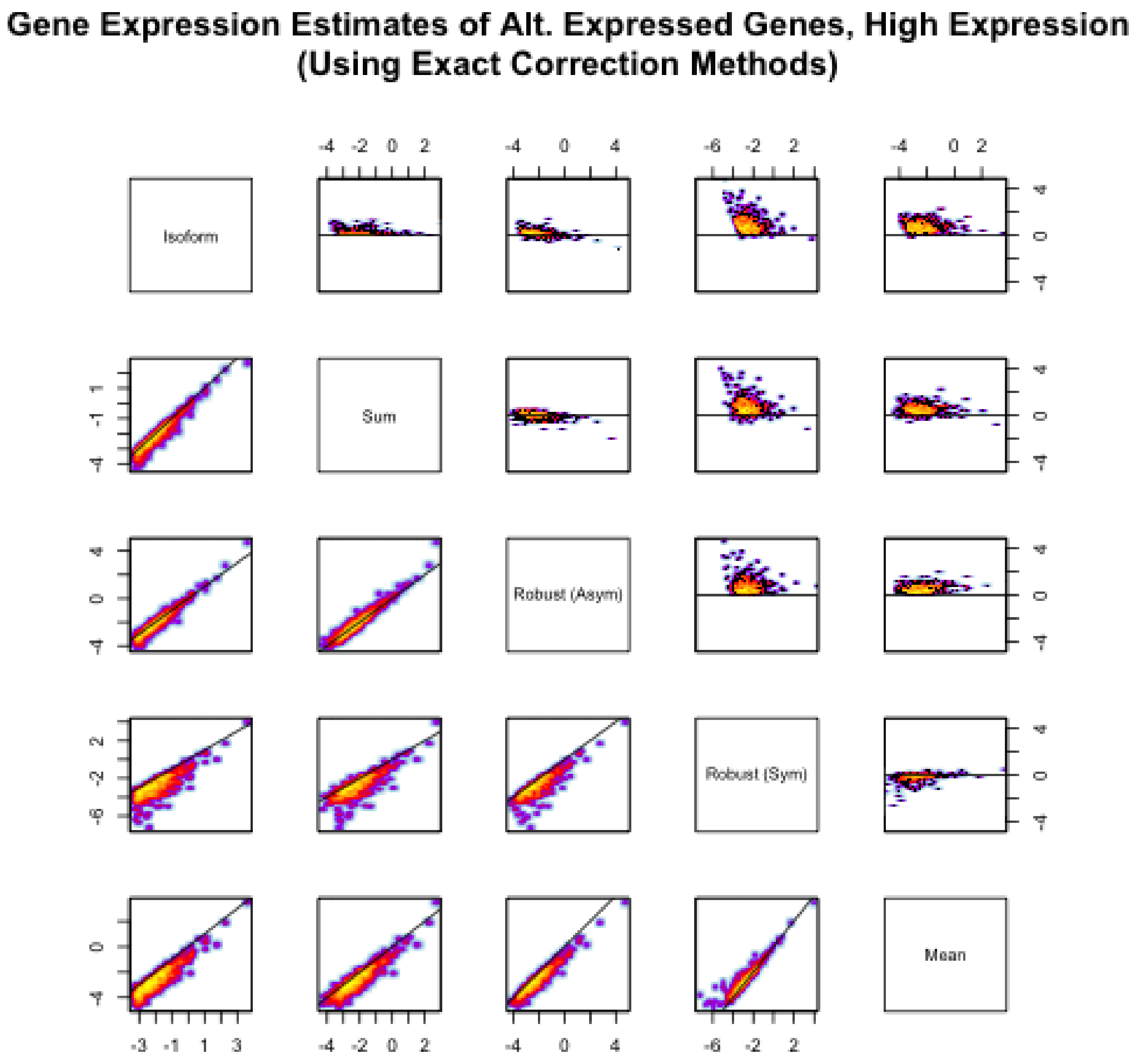
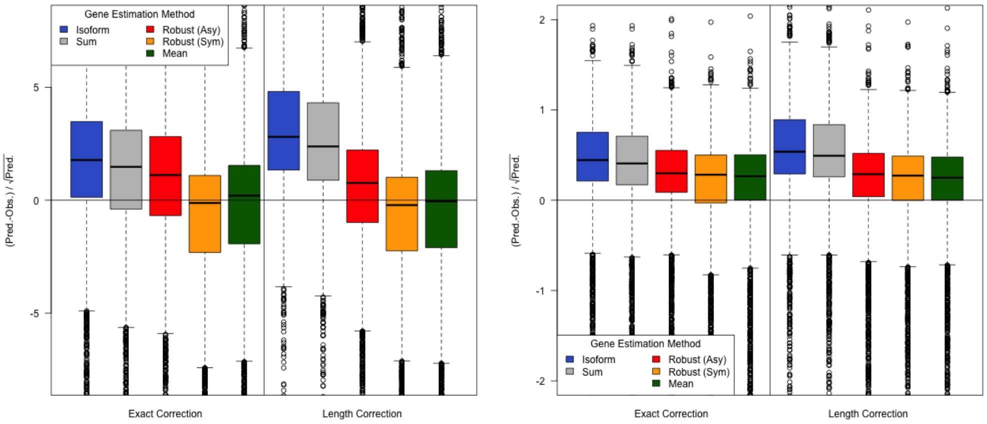


Figure S14: Scatter plots of the estimated gene levels for different methods applied to a single AML tumor. In the lower left plots are scatterplots of the log_2 of gene estimates; in the upper right plots are shown MA plots for comparing the two methods (i.e. the difference in log expression vs the average log expression). Plotted are genes which are annotated to be alternatively spliced (have at least 1 exon that is not contained in all transcripts). Shown here are methods using the c_τ correction. Some extreme genes that differ between the isoform estimates and the exon-based estimates are not shown in this plot, see Supplementary text for more details.



(a) High/Medium Expressed Genes

(b) Low Expressed Genes

Figure S15: Boxplots of observed exon counts versus that predicted for a single AML tumor, comparing the two exon fragmentation normalization methods. See legend for Figure S7 for details.

Supplementary Text

S1 Description of Simulation

We randomly selected gene models from Ensembl 58 annotation so as to have 10 genes each with 2-5 transcripts per gene, 5 genes with each of 6-10 transcripts per gene, and 5 genes with 15 transcripts per gene, with specified numbers of transcripts per gene, resulting in 70 genes and 415 transcripts (Table S1). There were no overlapping transcripts in this simulation that were not annotated to be from the same gene. For the 415 transcripts, we created true expression values λ_t for each transcript by randomly simulating values from an exponential distribution with rate $1/50$. To then create the true expression value for the sequenced data after considering fragmentation, we rescaled the λ_t values by the length of the transcript divided by the median transcript length, $L_t/median_t(L_t)$. With these expression values, we then simulated the number of fragments per transcript, X_t, from either a Poisson or a Negative Binomial with mean equal to the rescaled λ_t expression values. For the Negative Binomial distribution, the variance expressed as a function of the mean μ of each gene was $\mu + \mu^2/10$. For each transcript, X_t fragment lengths were simulated from a truncated normal distribution, with the original normal distribution having mean 250bp and standard deviation 50bp, truncated to require the fragment lengths to be 100-500bp. Then X_t starting positions within the transcript (one for each fragment length) were simulated either uniformly from the range $[1, L_t - \ell + 1]$, where ℓ is the fragment length, or from a distribution with the probability of selection proportion to the GC-content of the entire fragment. The sequences for these fragments were then obtained and aligned to the 415 transcript sequences using Bowtie (Langmead et al., 2009), allowing 0 mismatches and reporting all alignments.

# Transcripts per Gene	2	3	4	5	6	7	8	9	10	15
# Genes selected	10	10	10	10	5	5	5	5	5	5

Table S1: Number of genes randomly selected, broken down by number of transcripts per gene.

The alignments were processed to calculate the exons that each fragment overlapped for our robust GLM calculations. For the isoform deconvolution estimates, we calculated for each fragment the length and start position $(l_t(k), b_t(k))$ for each transcript for which the fragment aligned. Using these values, we calculated the p_{tk} values and found the maximum likelihood estimates per sample for the true expression values λ_t.

S2 Processing of the sequencing data

We processed three AML samples: TCGA-AB-2821, TCGA-AB-2843, and TCGA-AB-2888. For the AML samples, we converted the bam files retreived from dbGap (Mailman et al., 2007) into fastq files using the SamToFastq program provided with Picard-Tools (`http://picard.sourceforge.net/`). For the MAQC dataset, we downloaded the raw reads from the SRA Archive.

For each data sample, we aligned the reads with TopHat version 1.3.0 (Trapnell et al., 2009) with the 'no-novel-junctions' options activated (–no-novel-juncs) and instead provided the GTF annotation file from Ensembl 57. For the paired-end data we set the insert size of 100bp with standard deviation 50bp; 65-75% of reads aligned as proper pairs for most of our samples with these settings. We then calculated, per exon, the number of fragments with a unique mapping to the genome that overlapped the exon; for paired-end data only properly paired fragments were considered.

For the MAQC data set, fragments were required to overlap the exon by 5bp. For the AML data, we also aligned the paired data using Bowtie (Langmead et al., 2009) to the transcriptome fasta file provided by Ensembl 57. For the Bowtie alignment to the transcriptome, we set the maximum insert size to 500 (-X), and reported all alignments (-a). For the AML data, only reads that both aligned to the genome using TopHat and that aligned to the transcriptome using Bowtie were further considered (see Table S2). Since these reads already met the more strenuous requirement of mapping in both alignments, the exon overlap summaries for AML were calculated with slightly different parameters and only a 1bp overlap was required.

Samples	TCGA-AB-2821	TCGA-AB-2843	TCGA-AB-2888
All Mate Pairs (Passing QC)	62.2M	50.1M	63.0M
Mapped with Bowtie	25.3M	17.9M	30.8M
Mapped with Tophat	44.5M	24.2M	46.5M
Mate Pairs after filtering	18.5M	12.0M	23.4M

Table S2: Counts of number of fragments at different stages for TCGA data

Number of	Genes	Transcripts	Exons	Single Transcript Genes	Single Exon Genes
	21,268	100,286	387,556	6,240	594

Table S3: Basic information regarding annotation after creating exon definitions for Ensembl 57

S3 Definition of Exons and Exon Overlap

Our definition of exon varies slightly from many databases. We define exons as regions of the transcriptome that are contained in all of the same set of isoforms. Specifically, we assume that the genome is broken into regions τ of contiguous bases, all of which are annotated to be in the same transcripts. We assume that the regions τ are non-overlapping across the genome.

Exon boundaries were calculated using our exon definition defined above based on the transcript annotation provided by Ensembl. We also divided UTR regions from coding regions, so that for each transcript each "exon" in the transcript is either entirely UTR or entirely coding (however, a single exon so defined could be UTR in one transcript and coding in another transcript). See Tables S3 and S4

In this work, we use as our basic unit of exon summary the number of fragments whose sequenced parts overlap the exon. Let Y_τ be the number of fragments that overlap the exon τ, where fragments may be included in the count of both Y_{τ_1} and Y_{τ_2} if the sequenced part of the fragment overlaps both τ_1 and τ_2. For example, with paired-end reads, this may happen either because one mate pair overlaps both τ_1 and τ_2 (such as exon-exon junctions) or because one mate overlaps τ_1 and the other mate pair overlaps τ_2.

S4 Calculation of gene-estimates

For each gene, we considered five methods of gene estimation: isoform deconvolution, total count, and three exon-based summaries. The isoform deconvolution estimates were based on finding the maximum likelihood estimates based on the model described in Section S5. The total fragment count of a gene consisted of all fragments within the boundaries of the gene. Gene length was calculated as the total number of coding nucleotides in the gene, and then the total count was normalized by dividing by gene length so as to give the estimate of gene expression.

The exon-based robust GLM estimates were based on solving the score functions given in the main text. The

	Exon Width (bp)	Gene Length (bp)	# Transcripts Per Gene (for multi-Tr. Genes)	# Exons Per Gene (for multi-Tr. Genes)	% Constitutive Exons Per Gene (for multi-Tr. Genes)
Min	10	30	2	2	0.0%
1st Quartile	57	1,718	3	12	0.0%
Median	106	3,021	4	19	18.1%
Mean	206.6	3,765	6.3	23.11	29.2%
3rd Quartile	174	5,011	8	30	54.6%
Max	17040	115,900	68	374	100%

Table S4: Five-number summaries of statistics regarding annotation after creating exon definitions for Ensembl 57

symmetric robust method used the penalty function given by the Huber penalty function,

$$\psi_{sym}(r) = \begin{cases} r & |r| \leq q \\ q\ sign(r) & |r| > q \end{cases},$$

where q was a fixed constant set to 1.345. The asymmetric penalty function is given by,

$$\psi_{asym}(r) = \begin{cases} r & r \geq -q \\ -q & r < -q \end{cases},$$

with $q = 1.345$ as well. Another exon-based estimate was the simple mean of Y_τ / c_τ for all exons in the gene.

For comparison, we also calculated all of these exon-based estimates using a 'length' normalization instead of c_τ. In the case of paired-end reads, we took this to be $2(|\tau| + r - o)$ and for single end reads $|\tau| + r - o$, where r is the read-length and o the amount of overlap required of the exon. The paired-end reads are multiplied by a factor of two to account for the two sequenced mate ends that can overlap them.

We also calculated all of these exon-based estimates using only those exons annotated to be constitutive within the gene so as to give our constitutive estimates.

For the MAQC data, only exons with at least 10bp were considered in creating the exon-based gene estimates. For AML, we considered only gene estimates for the 2,786 genes that 1) had no overlapping reads in any of the three samples that also overlapped another gene 2) all exons in the gene were greater than 20bp, 3) was annotated as 'protein-coding' by Ensembl.

S5 Normalization of Exon-level Counts

We give here the derivation of our the fragmentation normalization constant c_τ. We consider only our definition of exon-overlap counts described above, but a similar derivation could be considered for other definitions by recalculating $K^t_\tau(\ell)$

We assume for simplicity that every possible fragment has a single possible alignment to the genome. We note that this assumption allows the same fragment to have multiple alignments to the transcriptome (i.e. be compatible with multiple different transcripts), but a single genomic alignment. We index the set of all possible genomic alignment of all fragments by the index k.

Let p_{tk} be the probability that a fragment from transcript $t = 1, \dots, T$ corresponds to a specific genomic alignment indexed by k,

$$p_{tk} = P(\text{fragment has genomic alignment } k | \text{fragment comes from } t).$$

We assume $p_{tk} = 0$ if k is not possible in transcript t. p_{tk} might also be 0 if a certain sequence is masked in the mapping due to repeat regions. Let F_t be the number of fragments we expect a single transcript t to contribute to the pool of mRNA after fragmentation, which will usually be estimated as $\approx L_t$, the length of the transcript t.

To be precise about which parameters are shared across samples and which are sample-specific, we will consider the case with multiple samples. Let Y_{ik} be the number of fragments in sample i which correspond to the genomic alignment k (generally zero or one, particularly if paired-end data). Then if each isoform t has expression level λ_{it}, the a simple Poisson model for the distribution of Y_{ik} is given by

$$Y_{ik} \sim \text{Poisson}(\sum_t \lambda_{it} F_t p_{tk}),$$

λ_{it} can be further parameterized into a portion representing the sequencing depth of the sample, $\lambda_{it} = \theta_i \eta_{it}$ (e.g. for RPKM, $\theta_i = N_i$, the total number of mapped reads in a sample i). We can more simply write the expression in vector notation

$$Y_{ik} \sim \text{Poisson}(\boldsymbol{\lambda}_i^T \mathbf{a}_k).$$

where the vector $\mathbf{a}_k \in \mathbb{R}^T$ consists of elements $a_k^{(t)} = F_t p_{tk}$, all of which are assumed known. This is essentially the same model as Salzman et al. (2010), but Salzman et al. instead have a constant a_{itk} which takes the place of $\theta_i F_t p_{tk}$ and they assume $\theta_i = N_i$; we instead allow for additional choices of normalization following Bullard et al. (2010).

Note that if $T = 1$, there is no alternative splicing. In this case, the maximum likelihood estimates of $\lambda_{it} = \mu_i$ is $\sum_k Y_{ik}/F_t$ (if there is uniform distribution of the reads across the transcript), which is the total fragment count estimate with $F_t = L_t$.

Then the exon overlap counts $Y_{i\tau}$ are given by

$$Y_{i\tau} = \sum_{k \in \tau} Y_{ik},$$

where $\{k : k \in \tau\}$ will be short-hand for the set of genomic alignments that overlap the exon τ in the required way. Because of the additivity of the Poisson distribution and the independence of the counts Y_{ik}, then

$$Y_{i\tau} \sim \text{Poisson}(\boldsymbol{\lambda}_i^T \mathbf{a}_\tau),$$

where $\mathbf{a}_\tau = \sum_{k \in \tau} \mathbf{a}_k$.

In order for there to be a single fragmentation normalization constant, c_τ, for an exon we need that $a_\tau^{(t)} = c_\tau$ be the same across all isoforms. We show that for an assumption of uniform distribution of fragments across the transcript,

$$a_\tau^{(t)} = E_H\left(\frac{F_t K_\tau^t(Z)}{L_t - Z + 1}\right),$$

where $K_\tau^t(Z)$ is proportional to the number of unique locations k of length Z that can overlap the exon in transcript t and H is the distribution of fragment lengths. As discussed in the main text, $K_\tau^t(Z)$ generally varies for different isoforms. However, in certain situations, $K_\tau^t(\ell)$ is independent of the isoform. For such exons, we can then make the assumption that $\frac{F_t}{E_H(L_t - \ell + 1)} \approx 1$ and define,

$$c_\tau = E_H(K_\tau^t(\ell)).$$

The assumption that $\frac{F_t}{E_H(L_t - \ell + 1)} \approx 1$ is common, since most isoform-deconvolution methods ignore such 'boundary effects' in their models for p_{tk} (e.g. Salzman et al. (2010) and Trapnell et al. (2010)).

The value $K_\tau^t(\ell)$ can be explicitly calculated and is a function of ℓ and the distance of the region τ from the 3' and 5' end of the transcript. In particular there are two cases when it is clear that $K_\tau^t(\ell)$ does not depend on the isoform for either paired-end or single-end sequencing: when the exon is in the middle of the isoform and when the exon is at the terminus of the isoform. For the exon-methods described in the main text, we classify each exon as one of these two classes of exons. Specifically, if the exon is at the end of the *gene* it will therefore be at the end of every isoform that contains it; otherwise, we classify the exon as in the middle of a transcript, which may be inaccurate.

In this work we make the assumption that the fragment lengths follows a known normal distribution truncated to be within a known range of fragment length values ($[50, 500]$), and calculate the corresponding expectations explicitly. Namely if Z follows a normal distribution truncated between z_1, z_2 then for $a \in [z_1, z_2]$ we have,

$$E(Z|Z > a) = \frac{1}{\Phi(z_2) - \Phi(z_1)}\{\mu + \sigma\frac{\varphi(\frac{a-\mu}{\sigma}) - \varphi(\frac{z_2-\mu}{\sigma})}{\Phi(\frac{z_2-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\}$$

and

$$E(Z|Z < a) = \frac{1}{\Phi(z_2) - \Phi(z_1)}\{\mu + \sigma\frac{\varphi(\frac{z_1-\mu}{\sigma}) - \varphi(\frac{a-\mu}{\sigma})}{\Phi(\frac{a-\mu}{\sigma}) - \Phi(\frac{z_1-\mu}{\sigma})}\}$$

For data-derived distributions H, we can calculate $E_H(K_\tau^t(\ell))$ computationally.

We now give a derivation of c_τ. This requires requires a model for the probability p_{tk}, which is dependent on the type of sequencing (e.g. paired-end vs. single-end) and also the assumptions about the distribution of fragments within an isoform. We give a derivation for paired-end and single-end reads and assume a simple model of distribution, but we note that the same calculations can be done, either explicitly or computationally, for more complicated distributions. Similarly, different definitions of exon overlap could be used and the calculations adjusted. Importantly, all of these calculations can be done independently of the data of raw sequence reads and only require the exon boundaries.

S5.1 Paired-end Reads

For the case of paired-end reads, if a fragment with genomic alignment k is known to come from transcript t, then we know the full length of the fragment, given by $l_t(k)$, as well as its starting position of the fragment relative to the 5' end of the isoform, given by $b_t(k)$. These are deterministic functions, given an isoform t and a genomic alignment k.

Then we can write the probability p_{tk} as

$$\begin{aligned} p_{tk} =& P(\text{fragment starts at position } b_t(k) |\text{ length of fragment} = l_t(k) \text{ \& fragment from isoform } t) \\ & \times P(\text{length of fragment} = l_t(k) | \text{fragment comes from isoform } t). \end{aligned}$$

We can write this more concisely as

$$p_{tk} = f_t(b_t(k)|l_t(k))h_t(l_t(k)),$$

where $h_t(\cdot)$ is the probability distribution of fragment lengths coming from transcript t and $f_t(b|\ell)$ is the probability of a fragment sequenced from transcript t with length ℓ starting at position b. We index h by t for generality, for example for those cases where the transcript is short, and thus there is a limit as to how big the fragment can be, but generally in this manuscript we assume that h is the same across different transcripts. Salzman et al. (2010) discuss explicit models for $h(\cdot)$ under assumptions regarding the experimental protocol, but we assume that h is known or can be accurately estimated from the data.

The distribution of starting positions given by f_t could encompass a large number of possible biases, such as GC-bias or bias from the 3' to 5' end. The simplest model, which we use in calculating our exon correction in the manuscript, is that the starting position of a fragment is uniform across the transcript,

$$f_t(b|\ell) = \begin{cases} \frac{1}{L_t - \ell + 1} & 0 < b \leq L_t - \ell + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then p_{tk} is equal for all reads that have the same length, $l_t(k) = \ell$, assuming they are from transcript t, and for such reads we have

$$a_k^{(t)} = \frac{F_t}{L_t - \ell + 1} h(\ell).$$

Note that Salzman et al. (2010) give their insert-length model as $a_{itk} \propto N_i h(\ell)$, which means that the models are essentially the same if we assume $\frac{F_t}{L_t - \ell + 1}$ is constant across transcripts.

Let $K_\tau^t(\ell)$ be the number of possible fragments that overlap τ and have length ℓ in isoform t – in other words the cardinality of the set $\{k : k \in \tau, l_t(k) = \ell\}$. Then if we assume the uniform distribution of fragments across the transcript we have

$$\begin{aligned} a_\tau^{(t)} =& \sum_{k \in \tau} \frac{F_t}{L_t - \ell_t(k) + 1} h_t(\ell_t(k)) \\ =& \int_{\mathbb{Z}_+} \sum_{k \in \tau, l_t(k) = \ell} \frac{F_t}{L_t - \ell + 1} h_t(\ell) d\ell \\ =& \int_{\mathbb{Z}_+} \frac{F_t K_\tau^t(\ell)}{L_t - \ell + 1} h_t(\ell) d\ell \\ =& E_H(\frac{F_t K_\tau^t(Z)}{L_t - Z + 1}) \end{aligned}$$

where $E_H(\cdot)$ is the expected value taken over the fragment length distribution with density $h(\cdot)$, and the approximation is given by the delta method.

If instead we assumed a non-uniform distribution, then the probabilities would not be the same for all reads with the same length ℓ, but would instead also depend on other features of the fragment k. For example, if the distribution depended on GC-bias, instead of calculating the expected number of reads with a particular length that overlapped τ, the number of reads with a particular length and GC-content might need to be calculated for each exon τ.

S5.1.1 Calculating $K^t_\tau(\ell)$

We derive $K^t_\tau(\ell)$ for the definition of exon overlap used in the main text. Without loss of generality, we assume that the first mate is on the 5' end of the transcript and the second is on the 3' end of the transcript. We assume that within isoform t, the exon τ refers to an interval of bases $[s, e]$, where s and e refer to positions relative to the start position of the transcript. Let $d^5_t \geq 0$ be the number of base-pairs separating the exon τ from the 5' end of the isoform and $d^3_t \geq 0$ be the number of base-pairs from the 3' end of the isoform. The read length is given by r, and o is the amount of overlap with the exon required to count a fragment as overlapping an exon. Then $x_5 = min(d^5_t, r - o)$ and $x_3 = min(d^3_t, r - o)$ are the number of base pairs that the fragment can extend beyond the exon region on the 5' and 3' end of the exon, respectively, still overlap the exon, *and* be contained within the isoform.

Recall that $b_t(k)$ refers to the start position for the fragment indexed by k relative to the start position of the transcript; for simplicity, we will drop the subscript t. We calculate $K^t_\tau(\ell)$ by noting that a fragment indexed by k that comes from transcript t can overlap the exon τ by satisfying one of three conditions:

S1. the first mate overlaps the exon and not the second

S2. the second mate overlaps the exon and not the first

S3. both the first and the second mates overlap exon.

We abbreviate the corresponding sets of fragments as S_1, S_2 and S_3, so that $K^t_\tau(\ell) = |S_1| + |S_2| + |S_3|$.

We can write these as conditions on the set of possible start positions of the fragments, which means that the set of fragments can be written in the form of $\{k : b(k) \in [x, y]\}$, for some coordinates x, y, plus the implied condition $x \leq y$ (which is dependent not on k, but on ℓ and the properties of the transcript and exon). Then these sets can be stated as,

S1. $b(k) \in [max(s - x_5, e - \ell + (r - o) + 1), min(e - o + 1, e - (\ell - d^3_t) + 1)]$

S2. $b(k) \in [max(s - d^5_t, s - \ell + x_5 + r), min(s - (r - o) - 1, e - \ell - x_3 + 1)]$

S3. $b(k) \in [s - x_5, e - \ell + x_3 + 1]$

Calculating the cardinality of these sets results in a straightforward set of conditional functions, based on the relationship of ℓ to the transcript and the exon.

Define the following parameters:

$$\begin{aligned}
\gamma_1^{max} &= max(d^3_t + o, |\tau| + (r - o) + x_5) \\
\gamma_1^{min} &= min(d^3_t + o, |\tau| + (r - o) + x_5) \\
\gamma_2^{max} &= max(d^5_t + o, |\tau| + (r - o) + x_3) \\
\gamma_2^{min} &= min(d^5_t + o, |\tau| + (r - o) + x_3)
\end{aligned}$$

If $\gamma_1^{min} \leq r$ or $\gamma_1^{max} \geq |\tau| + x_5 + d^3_t$, then $|S_1| = 0$. Otherwise,

$$|S_1| = \begin{cases} |\tau| - (\ell - d^3_t) + x_5 + 1 & |\tau| + x_5 + d^3_t \geq \ell \geq \gamma_1^{max} \\ \gamma_1^{min} - r & \gamma_1^{min} \leq \ell < \gamma_1^{max} \\ \ell - r + 1 & r \leq \ell < \gamma_1^{min} \\ 0 & \text{otherwise} \end{cases}$$

The calculation of $|S_2|$ is of course symmetric, with d^5_t replacing d^3_t, x_3 replacing x_5,

$$|S_2| = \begin{cases} |\tau| - (\ell - d^5_t) + x_3 + 1 & |\tau| + x_3 + d^5_t \geq \ell \geq \gamma_2^{max} \\ \gamma_2^{min} - r & \gamma_2^{min} \leq \ell < \gamma_2^{max}, \\ \ell - r + 1 & r \leq \ell < \gamma_2^{min} \\ 0 & \text{otherwise} \end{cases}$$

For the final set of alignments we have that,

$$|S_3| = \begin{cases} |\tau| - \ell + x_5 + x_3 + 1, & r \leq \ell \leq |\tau| + x_3 + x_5 \\ 0 & otherwise \end{cases}$$

S5.1.2 Special Cases of $K^t_\tau(\ell)$ independent of isoform

As described above, there are two cases when it is clear that $K^t_\tau(\ell)$ does not depend on the isoform: when the exon is in the middle of the transcript and when the exon is at the terminus of the isoform.

- *Exons in Middle of Transcript* If we assume that $\ell \leq d^3_t \leq d^5_t$, i.e. that the region is far from the ends of the transcripts, then we also reasonably assume that $x_3 = x_5 = (r - o)$ and we have,

$$K^t_\tau(\ell) = \begin{cases} 2(|\tau| + r - 2o) & |\tau| + 2(r-o) \leq \ell \\ |\tau| + \ell - 2o + 3 & r \leq \ell \leq |\tau| + 2(r-o) \end{cases} \tag{S-1}$$

(S-2)

- *Exons at Terminus of Transcript* If we assume that $\ell \leq d^5_t$ and $d^3_t = 0$ with $x_5 = r - o$ and $x_3 = 0$ i.e. the exon τ lies at the terminus of the transcript, then we have,

$$K^t_\tau(\ell) = |\tau| - o$$

We note that in the special case of single-exon genes, we cannot assume that $\ell \leq d^5_t$. Rather, $d^3_t = d^5_t = x_3 = x_5 = 0$ and therefore $C_1 = C_2 = 0$, and

$$K^t_\tau(\ell) = C_3 = \begin{cases} |\tau| - \ell + +1, & r \leq \ell \leq |\tau| \\ 0 & otherwise \end{cases}$$

Because in this case there only a single transcript possible, so that $L_t = |\tau|$, we can calculate the exact correction factor,

$$E_H(\frac{F_t K^t_\tau(Z)}{L_t - Z + 1}) = F_t.$$

Therefore, if F_t is the taken as the length of the transcript (i.e. $|\tau|$) then we still have the correction is exactly $|\tau|$. This is the estimate created by the total gene estimates – which corrects the total number of counts in the gene by the total gene length – and is also the same as the isoform deconvolution method.

Thus the number of fragments that can overlap the exon τ that is located at the end of a transcript is just the length of the exon, regardless of the length ℓ of the fragment, but for other exons, the length distribution of fragments relative to the exon length will affect the number of fragments that can overlap it.

S5.2 Single-End Reads

For the case of single-end reads, our index k refers to genomic alignment of the sequenced end of the read and the orientation of the alignment. For a genomic alignment k, $d(k) \in \{-1, +1\}$ indicate whether the read mapped to the forward or reverse strand of the genome.

From this information, assuming we know what transcript it comes from we know the position of one end of the fragment and which end of the fragment to which it refers. We do not, however, know the length of the fragment. Let $m_t(k)$ be the position, relative to the start of isoform t, of the end of the fragment (which may be either the 5' or 3' end depending on the value of $d(k)$). Then we can write the probability p_{tk} as

$$\begin{aligned} p_{tk} =& P(\text{fragment has alignment } k | \text{from isoform } t) \\ =& P(\text{fragment starts at } m_t(k) | \text{ fragment from isoform } t, d(k)) P(d(k) | \text{ fragment from isoform } t) \end{aligned}$$

An obvious assumption is to $P(d(k) = 1) = P(d(k) = -1) = 1/2$ for all transcripts and alignments. If we knew the length of the fragment, we could use the same model as with the paired-end reads. Instead we can condition on the possible values that the fragment might have taken,

$$p_{tk} = \begin{cases} \frac{1}{2} \int_{\mathbb{Z}_+} \frac{1}{L_t - \ell + 1} \mathbb{1}\{1 \leq m_t(k) \leq L_t - \ell + 1\} h(\ell) & d(k) = 1 \\ \frac{1}{2} \int_{\mathbb{Z}_+} \frac{1}{L_t - \ell + 1} \mathbb{1}\{\ell \leq m_t(k) \leq L_t\} h(\ell) & d(k) = -1 \end{cases}$$

If we do not want to assume that we know the distribution of fragment lengths, then we could use a point-mass at a fixed value of ℓ, which would be a fudge factor that only allows an alignment k to have positive probability of being in isoform t only if fragments of length ℓ or less were compatible with the transcript. A point mass at $\ell = 0$ would give the simple model that is often used that does not account for the rest of the fragment and requires only that the sequenced portion of the fragment lie within the transcript (e.g. Salzman et al. (2010)).

Then we have that $Y_{i\tau}$ is the sum of Y_{ik} over all alignments k that overlap τ, of either orientation.

If we again assume the uniform distribution of fragments across the transcript we have,

$$\begin{aligned}
a_\tau^{(t)} =& \sum_{k\in\tau,d(k)=1} F_t p_{tk} + \sum_{k\in\tau,d(k)=-1} F_t p_{tk} \\
=& \frac{1}{2}\int_{\mathbb{Z}_+} h(\ell)\frac{F_t}{L_t-\ell+1}\Big(\sum_{k\in\tau,d(k)=1,l_t(\ell)=\ell} \mathbb{1}\{1 \leq m_t(k) \leq L_t-\ell+1\} + \sum_{k\in\tau,d(k)=-1,l_t(\ell)=\ell} \mathbb{1}\{\ell \leq m_t(k) \leq L_t\}\Big) \\
=& \frac{1}{2}\int_{\mathbb{Z}_+} \frac{F_t(K^t_{\tau+}(\ell)+K^t_{\tau-}(\ell))}{L_t-\ell+1} h_t(\ell)d\ell \\
=& E_H\Big(\frac{F_t K^t_\tau(Z)}{L_t-Z+1}\Big)
\end{aligned}$$

where $K^t_{\tau+}(\ell)$ is the number of fragments of length ℓ that overlap τ and are in the positive orientation, and similarly for $K^t_{\tau-}(\ell)$. For convenience, we then define $K^t_\tau(Z) = (K^t_{\tau+}(\ell) + K^t_{\tau-}(\ell))/2$, to be half of the total number of fragments that could overlap τ, so as to have similar formulas as with paired-end reads.

S5.2.1 Calculating $K^t_\tau(\ell)$

Using the same logic as before, we can explicitly calculate $K^t_{\tau+}(\ell)$ and $K^t_{\tau-}(\ell)$. For the reads with positive orientation, we have that

$$m_t(k) \in [s - x_5, min(e - o + 1, e - \ell + d^3_t + 1)]$$

and for reads with negative orientation we have that,

$$m_t(k) \in [max(s - o + 1, s + \ell - d^5_t - 1), e + x_3].$$

Using this information, we have that

$$K^t_{\tau+}(\ell) = \begin{cases} |\tau| - (\ell - d^3_t) + x_5 + 1 & d^3_t + o \leq \ell \leq |\tau| + x_5 + d^3_t \\ |\tau| + x_5 - o + 1 & r \leq \ell < d^3_t + o \end{cases}$$

$K^t_{\tau-}(\ell)$ is the same, only with the x_5 and d^3_t changed to x_3 and d^5_t, respectively.

Again, we have that the value of $K^t_\tau(\ell)$ is independent of the isoform in the same two instances.

- *Exons in Middle of Transcript* If we assume that $\ell \leq d^3_t \leq d^5_t$, i.e. that the region is far from the ends of the transcripts, then $x_3 = x_5 = (r - o)$ and we have,

  $$K^t_\tau(\ell) = |\tau| + r - 2o + 1$$

  which is basically equivalent to normalizing the exon counts by the length of the exon.

- *Exons at Terminus of Transcript* If we assume that $\ell \leq d^5_t$ and $d^3_t = 0$ with $x_5 = r - o$ and $x_3 = 0$ i.e. the exon τ lies at the terminus of the transcript, then we have,

  $$K^t_\tau(\ell) = \begin{cases} (|\tau| - o + 1)/2 & |\tau| + r - o \leq \ell \\ |\tau| - o + 1 - (\ell - r)/2 & r \leq \ell \leq |\tau| + r - o \end{cases}$$

Thus the opposite appears for paired end reads: the number of fragments that can overlap the exon τ in the middle of a transcript is just the length of the exon, regardless of the length ℓ of the fragment, but for terminal exons (and more generally exons near the end of the transcript), the length distribution of fragments will affect the number of fragments that can overlap it.