Year 2012 *Paper* 97

Quantifying alternative splicing from paired-end RNA-sequencing data

David Rossell[*] Camille Stephan-Otto Attolini[†]

Manuel Kroiss[‡] Almond Stöcker[**]

[*] Institute for Research in Biomedicine of Barcelona, david.rossell@irbbarcelona.org

[†] Institute for Research in Biomedicine of Barcelona, camille.stephan@irbbarcelona.org

[‡] Ludwig Maximilians Universitat, Munchen, kroissm@in.tum.de

[**] Ludwig Maximilians Universitat, Munchen, al.st@web.de

Quantifying alternative splicing from paired-end RNA-sequencing data

David Rossell, Camille Stephan-Otto Attolini, Manuel Kroiss, and Almond Stöcker

Abstract

RNA-sequencing has revolutionized biomedical research and, in particular, our ability to study gene alternative splicing. The problem has important implications for human health, as alternative splicing is involved in malfunctions at the cellular level and multiple diseases. However, the high-dimensional nature of the data and the existence of experimental biases pose serious data analysis challenges. We find that the standard data summaries used to study alternative splicing are severely limited, as they ignore a substantial amount of valuable information. Current data analysis methods are based on such summaries and are hence suboptimal. Further, they have limited flexibility in accounting for technical biases. We propose novel data summaries and a Bayesian modeling framework that overcome these limitations and determine biases in a non-parametric, data-dependent manner. These summaries adapt naturally to the rapid improvements in sequencing technology. We provide efficient point estimates and uncertainty assessments. The approach allows to study alternative splicing patterns for individual samples and can also be the basis for downstream differential expression analysis. We found an over 5 fold improvement in estimation mean square error compared to a popular approach in simulations, and substantially higher correlations between replicates in experimental data. Our findings indicate the need for modifying the routine summarization and analysis of alternative splicing RNA-seq studies. We provide a software implementation in the R package casper.

Quantifying alternative splicing from paired-end RNA-sequencing data

David Rossell[0]
Institute for Research in Biomedicine of Barcelona (Spain)
Camille Stephan-Otto Attolini[0]
Institute for Research in Biomedicine of Barcelona (Spain)
Manuel Kroiss
LMU Munich and TU Munich (Germany)
Almond Stöcker
LMU Munich (Germany)

Abstract

RNA-sequencing has revolutionized biomedical research and, in particular, our ability to study gene alternative splicing. The problem has important implications for human health, as alternative splicing is involved in malfunctions at the cellular level and multiple diseases. However, the high-dimensional nature of the data and the existence of experimental biases pose serious data analysis challenges. We find that the standard data summaries used to study alternative splicing are severely limited, as they ignore a substantial amount of valuable information. Current data analysis methods are based on such summaries and are hence sub-optimal. Further, they have limited flexibility in accounting for technical biases. We propose novel data summaries and a Bayesian modeling framework that overcome these limitations and determine biases in a non-parametric, data-dependent manner. These summaries adapt naturally to the rapid improvements in sequencing technology. We provide efficient point estimates and uncertainty assessments. The approach allows to study alternative splicing patterns for individual samples and can also be the basis for downstream differential expression analysis. We found an over 5 fold improvement in estimation mean square error compared to a popular approach in simulations, and substantially higher correlations between replicates

[0] Both authors contributed equally to this work

in experimental data. Our findings indicate the need for modifying the routine summarization and analysis of alternative splicing RNA-seq studies. We provide a software implementation in the R package `casper` [1].

1 Introduction

RNA-sequencing (RNA-seq) produces an overwhelming amount of genomic data in a single experiment, providing an unprecedented resolution to address biological problems. We focus on gene expression experiments where the goal is to study alternative splicing (AS), which we briefly introduce. AS is an important biological process by which cells are able to express several variants, also known as isoforms, of a single gene. Each splicing variant gives rise to a different protein with a unique structure that can perform different functions and respond to internal and environmental needs. AS is believed to contribute to the complexity of higher organisms, and is in fact particularly common in humans (Blencowe, 2006). Additionally, it is known to be involved in multiple diseases such as cancer and malfunctions at the cellular level. Despite its importance, due to limitations in earlier technologies most gene expression studies have ignored AS and focused on overall gene expression.

Consider the hypothetical example of a gene with three splice variants shown in Figure 1. The gene is encoded in the DNA in three exons, shown as boxes in Figure 1. When the gene is transcribed as messenger RNA (mRNA), it can give rise to three isoforms. Variant 1 is formed by all three exons, whereas variant 2 skips the second exon and variant 3 the third exon. Usually, multiple variants are expressed simultaneously at any given time. In our example, variant 1 makes up for 60% of the overall expression of the gene, variant 2 for 30% and variant 3 for 10%. In practice, these proportions are unknown and our goal is to estimate them as accurately as possible.

We focus on paired-end RNA-seq experiments, as they are the current standard and provide higher resolution for measuring isoform expression than competing technologies, *e.g.* microarrays (Pepke et al., 2009). RNA-seq sequences tens or even hundreds of million mRNA fragments, which can then be aligned to a reference genome using a variety of software, *e.g.* TopHat (Trapnell et al., 2009), SOAP (Li et al., 2009) or BWA (Li and Durbin, 2009). Throughout, we assume that the software can handle gapped alignments (we used TopHat in all our examples). Early RNA-seq studies used single-end sequencing, where only the left or right end of a fragment is sequenced. In

[1] `https://sites.google.com/site/rosselldavid/software`

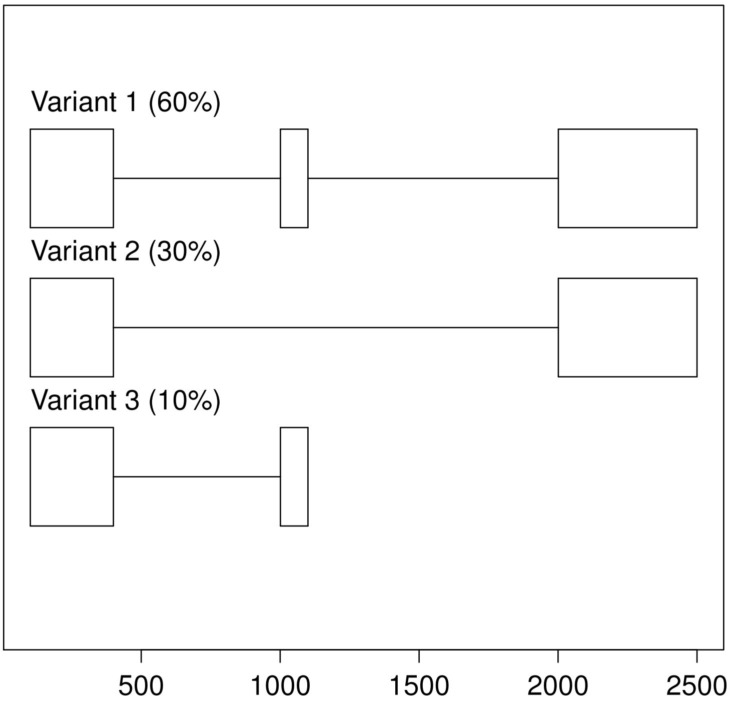

Figure 1: Three splice variants for a hypothetical gene and their relative abundances. Exon1 is located at positions 101-400. Exon 2 at 1001- 1100. Exon 3 at 2001-2500.

contrast, paired-end RNA-seq sequences both fragment ends. Table 1 shows three hypothetical sequenced fragments corresponding to the gene in Figure 1. 75 base pairs (bp) were sequenced from each end. For instance, both ends of fragment 1 align to exon 1. As the three variants contain exon 1, in principle this fragment could have been generated by any variant. For fragment 2 the left read aligned to exons 1 and 2 (*i.e.* it spanned the junction between both exons), and the right read to exon 3. Hence, fragment 2 can only have been generated from variant 1. Finally, fragment 3 visits exons 1 and 2 and hence it could have been generated either by variants 1 or 3. The example is simply meant to provide some intuition. In practice most genes are substantially longer and have more complicated splicing patterns. Precise probability calculations are required to ensure that the conclusions are sound.

Ideally, one would want to sequence the whole variant, so that each fragment can be uniquely assigned to a variant. Unfortunately, current technologies sequence hundreds of base pairs, which is orders of magnitude shorter than typical variant lengths. Current statistical approaches are based on

	Chromosome	Left read	Right read	Exon path
Fragment 1	chr1	110-185	200-274	{1} , {1}
Fragment 2	chr1	361-400; 1001-1035	2011-2085	{1,2} , {3}
Fragment 3	chr1	301-375	1021-1095	{1} , {2}
...				

Table 1: Three paired-end RNA-seq fragments. Aligned chromosome and base pairs are indicated for both ends, allowing for gapped alignments. The exon path indicates the sequence of exons visited by each end. A typical experiment contains tens of millions of fragments.

the observation that, while most sequenced fragments cannot be uniquely assigned to a variant, it is possible to make probability statements. For instance, fragment 3 in Table 1 may have originated either from variants 1 or 3, but the probability that each variant generates such a fragment is different. As we shall see below, this observation prompts a direct use of Bayes theorem.

In principle, one could formulate a probability model that uses the full data, *i.e.* the exact base pairs covered by each fragment such as provided in Table 1. However, the massive number of sequences renders this approach computationally prohibitive even when only considering a few genes. Several authors proposed summarizing the data by counting the number of fragments either covering each exon or each exon junction (*e.g.* Xing et al. (2006), Mortazavi et al. (2008), Jiang and Wong (2009)). In fact, large-scale genomic databases report precisely these summaries, *e.g.* The Cancer Genome Atlas project [2]. One can then pose a probability model that uses count data from a few categories as raw data, which greatly simplifies computation. While useful, this approach is seriously limited to considering pairwise junctions, which discards relevant information. For instance, suppose that a fragment visits exons 1, 2 and 3. Simply adding 1 to the count of fragments spanning exons 1-2 and 2-3 ignores the joint information that a single fragment visited 3 exons and decreases the confidence when inferring the variant that generated the fragment. Our results suggest that ignoring this information can result in a serious loss of precision. It is not uncommon that a fragment spans more than 2 exons. Holt and Jones (2008) found a substantial proportion of fragments bridging several exons in paired-end RNA-seq experiments. In the 2009 RGASP experimental data set (Section 4) 38.0% and 40.9% fragments spanned $\geq$3 exons in replicate 1 and 2, respectively. In

[2] `http://cancergenome.nih.gov`

the 2012 ENCODE data set we found 64.7% and 65.2% in each replicate. The 2012 data had substantially longer reads and fragments, which illustrates the rapid advancements in technology. As sequencing evolves these percentages are expected to increase further.

We propose novel data summaries that preserve most information relevant to alternative splicing, while maintaining the computational burden at a manageable level. We record the sequence of exons visited by each fragment end, which we refer to as *exon path*, and then count the number of fragments following each exon path. The left end of Fragment 2 in Table 1 visits exons 1 and 2 and the right end exon 3, which we denote as $\{1,2\},\{3\}$. Notice that a fragment following the path $\{1\},\{2,3\}$ visits the same exons, so one could be tempted to simply record $\{1,2,3\}$ in both cases. However, the probability of observing $\{1,2\},\{3\}$ for a given variant differs from $\{1\},\{2,3\}$, and hence combining the two paths would result in a potential loss of information. Table 1 contains hypothetical exon path counts for our example gene. We use these counts as the basic input for our probability model.

Exon path	Count
{1} , {1}	2824
{2} , {2}	105
{3} , {3}	5042
{1} , {2}	27
{1} , {1,2}	423
{1} , {3}	127
{2,3} , {3}	394
{1,2} , {3}	2
{1} , {2,3}	13

Table 2: Exon path counts for hypothetical gene

Paired-end RNA-seq is critical for AS studies. Intuitively, compared to single-end sequencing it increases the probability of observing fragments that connect exon junctions. Lacroix et al. (2008) showed that, although neither protocol guarantees the existence of a unique solution, in practice paired end (but not single end) can provide asymptotically correct estimates for 99.7% of the human genes. In contrast, for single-end data the figure is 1.14%. Unfortunately, much of the current methodology has been designed with single end data in mind. Xing et al. (2006) formulate the problem as that of traversing a directed acyclic graph and formulate a latent variable based approach to estimate splice variant expression. Jiang and Wong (2009) propose a similar approach within the Bayesian framework. Both approaches were designed for single-end RNA-seq data. Guttman et al. (2010) and Ameur et al. (2010) proposed strategies to detect splicing junctions, and Katz et al. (2010) and Wu et al. (2011a) introduced models to estimate the percentage of isoforms skipping individual exons. However, these approaches do not estimate expression at the variant level.

Several authors propose strategies that use paired-ends. Mortazavi et al. (2008) and Trapnell et al. (2009) model the number of fragments spanning exon junctions. Salzman et al. (2011) extend the model to consider counts for individual exons as well as exon junctions. These approaches focus on pairwise exon connections, ignoring valuable higher-order information, and do not fully incorporate important technical biases. First, the sample preparation protocols usually induce an enrichment towards the 3' end of the transcript, *i.e.* fragments are not uniformly distributed along the gene. Wu et al. (2011b) extend the approaches above by relaxing the uniformity assumption. Further, the fragment length distribution plays an important role in the probability calculations and needs to be estimated accurately. While the approaches above acknowledge this issue, they do not estimate this distribution from the data. Rather, strong parametric assumptions are made based on reports provided by sequencing facilities. Our examples illustrate that these reports can be inaccurate, and that the distribution can exhibit multi-modalities that are not captured by the chosen parametric forms.

In summary, we propose a flexible framework to estimate alternative splicing from RNA-seq studies, by using novel data summaries and accounting for experimental biases. In Section 2 we formulate a probability model that goes beyond pairwise connections by considering exon paths. We model the read start and fragment size distributions non-parametrically, in order to accommodate arbitrary shapes. Section 3 discusses model fitting and provides algorithms to obtain point estimates, asymptotic credibility intervals and posterior samples. We show some results in Section 4 and provide concluding remarks in Section 5.

2 Probability model

We formulate the model at the gene level and perform inference separately for each gene. In some cases, exons from different genes overlap with each other. When this occurs we group the overlapping genes and consider all their isoforms simultaneously. It is also possible that two variants share only a part of an exon. We sub-divide such exons into the shared part and the part that is specific to each variant. For simplicity, from here on we refer to gene groups simply as genes and to sub-divided exons simply as exons.

Consider a gene with E exons starting at base pairs $s_1, \ldots, s_E$ and ending at $e_1, \ldots, e_E$. Denote the set of splicing variants under consideration by $\boldsymbol{\nu}$ (assumed to be known) and its cardinality by $|\nu|$. Each variant is characterized by an increasing sequence of natural numbers $i_1, i_2, \ldots$ that indicates the exons contained therein.

2.1 Likelihood and prior

As discussed above, we formulate a model for exon paths. We denote an exon path by $\boldsymbol{\iota} = (\boldsymbol{\iota}_l, \boldsymbol{\iota}_r)$, where $\boldsymbol{\iota}_l = (i_j, \ldots, i_{j+k})$ are the exons visited by the left-end and $\boldsymbol{\iota}_r = (i_{j'}, \ldots, i_{j'+k'})$ those by the right-end. Let $\mathcal{P}^*$ be the set of all possible exon paths and $\mathcal{P}$ be the subset of observed paths, *i.e.* the paths followed by at least 1 sequenced fragment.

The observed data is a realization of the random variable $\mathbf{Y} = (Y_1, \ldots, Y_N)$, where N is the number of paired-end reads and $Y_i \in \{1, \ldots, \mathcal{P}^*\}$ indicates the exon path followed by read pair i. Formally, Y_i arises from a mixture of $|\nu|$ discrete probability distributions, each component corresponding to a different splicing variant. The mixture weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{|\nu|})$ give the proportion of reads generated by each variant, *i.e.* its relative expression. That is,

$$P(Y_i = y_i | \boldsymbol{\pi}, \boldsymbol{\nu}) = \sum_{d=1}^{|\nu|} p_{y_i d} \pi_d,$$

where $p_{kd} = P(Y_i = k | \delta_i = d)$ is the probability of path k under variant d and δ_i is a latent variable indicating the variant that originated Y_i. Let S_i and L_i denote the relative start and length (respectively) of fragment i. The exon path Y_i is completely determined given S_i, L_i and the variant δ_i. Hence,

$$p_{kd} = \int \int I(Y_i = k \mid S_i = s_i, L_i = l_i, \delta_i = d) dP_L(l_i \mid \delta_i) dP_S(s_i \mid \delta_i, L_i), \quad (1)$$

where P_L is the fragment distribution and P_S is the read start distribution conditional on L. As discussed in Section 2.2, by assuming that P_S and P_L are shared across genes it is possible to estimate them with high precision. Hence, for practical purposes we can treat p_{kd} as known and pre-compute them before model fitting. Full derivations for p_{kd} are provided in Appendix A.

Assuming that each fragment is observed independently, the likelihood function can be written as

$$P(\mathbf{Y} | \boldsymbol{\pi}, \boldsymbol{\nu}) = \prod_{i=1}^{N} \sum_{d=1}^{|\nu|} p_{y_i d} \pi_d = \prod_{k=1}^{|\mathcal{P}|} \left(\sum_{d=1}^{|\nu|} p_{kd} \pi_d \right)^{x_k}, \quad (2)$$

where $x_k = \sum_{i=1}^{N} \mathrm{I}(y_i = k)$ is the number of reads following exon path k. (2) is log-concave, which guarantees the existence of a single maximum. Log-concavity is given by (i) the log function being concave and monotone increasing, (ii) $\sum_{d=1}^{|\nu|} p_{kd}\pi_d$ being linear and therefore concave, and (iii) the fact that a composition $g \circ f$ where g is concave and monotone increasing and f is concave is again concave. To see (iii), notice that

$$g \circ f(t\mathbf{z}_1 + (1-t)\mathbf{z}_2) \geq g(tf(\mathbf{z}_1) + (1-t)f(\mathbf{z}_2)) \geq tg \circ f(\mathbf{z}_1) + (1-t)g \circ f(\mathbf{z}_2)$$

where the first inequality is given by g being increasing and f concave, and the second inequality is given by g being concave.

We complete the probability model with a Dirichlet prior on $\boldsymbol{\pi}$:

$$\boldsymbol{\pi}|\boldsymbol{\nu} \sim \mathrm{Dir}(q_1, \ldots, q_{|\nu|}). \tag{3}$$

In Section 4 we assess several choices for q_d. By default we set the fairly uninformative values $q_d = 2$, as these induce negligible bias and stabilize the posterior mode by pooling it away from the boundaries 0 and 1. It is easy to see that (3) is log-concave when $q_d \geq 1$ for all d. Given that (2) is also log-concave, this choice of $\mathbf{q}$ guarantees the posterior to be log-concave, and therefore the uniqueness of the posterior mode.

2.2 Fragment length and read start distribution estimates

Evaluating the exon path probabilities in (1) that appear in the likelihood (2) requires the fragment start distribution P_S and fragment length distribution P_L. We assume a common P_L across all genes and variants, with the restriction that a fragment cannot be longer than the variant that generated it. Denoting by T the length of variant δ_i (in bp), we let $P_L(l \mid \delta) = P_L(l \mid T) = P(L = l)\mathrm{I}(l \leq T)/P(l \leq T)$. That is, the conditional distribution of L given δ is simply a truncated version of the marginal distribution.

Further, we assume a common fragment start distribution relative to the variant length T. Conditional on L and T, P_S is truncated so that the fragment ends before the end of the variant. Specifically,

$$P_S(S \leq s \mid \delta_i, L = l) = P\left(\frac{S}{T} \leq z|T, L = l\right) = \frac{\varphi\left(\min\{z, \frac{T-l+1}{T}\}\right)}{\varphi\left(\frac{T-l+1}{T}\right)}, \tag{4}$$

where $z = s/T$ and $\varphi(z) = P(\frac{S}{T} \leq z)$ is the distribution of the relative read start $\frac{S}{T}$.

To estimate P_L note that the fragment length is unknown for fragments that span multiple exons, but it is known exactly when both ends fall in the same exon. Therefore, we select all such fragments and estimate P_L with the empirical probability mass function of the observed fragment lengths. In order to prevent short exons from inducing a selection bias, we only use exons that are substantially longer than the expected maximum fragment length (by default $> 1000bp$).

Estimating the fragment start distribution P_S is more challenging, as we do not know the variant that generated each fragment and therefore its relative start position cannot be determined. We address this issue by selecting genes that have a single annotated variant, as in principle for these genes all fragments should have been generated by that variant. Of course, the annotated genome does not contain all variants, and therefore a proportion of the selected fragments may not have been generated by the assumed variant. However, the annotations are expected contain most common variants (*i.e.* with highest expression), and hence most of the selected fragments should correspond to the annotated variant. Under this assumption, we can determine the exact start S_i and length L_i for all selected fragments. A difficulty in estimating the read start distribution is that the observed (S_i, L_i) pairs are truncated so that $S_i + L_i < T$, whereas we require the untruncated cumulative distribution function $\rho(\cdot)$ in (4). Fortunately, the truncation point for each (S_i, L_i) is known and therefore one can simply obtain a Kaplan-Meier estimate of $\rho(\cdot)$ (Kaplan and Meier, 1958). We use the function `survfit` in the R `survival` package (Therneau and Lumley, 2011).

3 Model fitting

We provide algorithms to obtain a point estimate for $\boldsymbol{\pi}$, asymptotic credibility intervals and posterior samples.

Following a 0-1 loss, as a point estimate we report the posterior mode, which is obtained by maximizing the product of (2) and (3), subject to the constraint $\sum_{d=1}^{|\nu|} \pi_d = 1$. We note that maximum likelihood estimates are obtained by simply setting $q_d = 1$ in (3). This constrained optimization can be performed with many numerical optimization algorithms. Here we used the EM algorithm (Dempster et al., 1977), as it is computationally efficient even when the number of variants $|\nu|$ is large. For a detailed derivation see Appendix B. As noted above, for $q_d > 1$ the log-posterior is concave and therefore the algorithm converges to the single maximum. The steps required for the algorithm are:

1. Initialize $\pi_d^{(0)} = q_d / \sum_{d=1}^{|\nu|} q_d$.

2. At iteration j, update $\pi_d^{(j+1)} = q_d - 1 + \sum_{k=1}^{|\mathcal{P}|} x_k \frac{p_{kd}\pi_d^{(j)}}{\sum_{i=1}^{|\nu|} p_{ki}\pi_i^{(j)}}$.

Step 2 is repeated until the estimates stabilize. In our examples we required $|\pi_d^{(j+1)} - \pi_d^j| < 10^{-5}$ for all d. Notice that p_{kd} and x_k remain constant through all iterations, and hence they only need to be computed once.

We characterize the posterior uncertainty asymptotically using a normal approximation in the re-parameterized space $\theta_d = \log(\pi_{d+1}/\pi_1)$, $d = 1, \ldots, |\nu| - 1$ and the delta method (Casella and Berger, 2001). Denote by $\boldsymbol{\mu}$ the posterior mode for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{|\nu|-1})$ and by S the Hessian of the log-posterior evaluated at $\boldsymbol{\theta} = \boldsymbol{\mu}$. Further, let $\pi(\boldsymbol{\theta})$ be the inverse transformation and $G(\boldsymbol{\theta})$ the matrix with (d, l) element $G_{dl} = \frac{\partial}{\partial \theta_l} \pi_d(\boldsymbol{\theta})$. Detailed expressions for S, $\pi(\boldsymbol{\theta})$ and $G(\boldsymbol{\theta})$ are provided in Appendix C. The posterior for $\boldsymbol{\theta}$ can be asymptotically approximated by $N(\boldsymbol{\mu}, \Sigma)$, where $\Sigma = S^{-1}$. Hence, the delta method approximates the posterior for $\boldsymbol{\pi}$ with $N\left(\pi(\boldsymbol{\mu}), G(\boldsymbol{\mu})' S G(\boldsymbol{\mu})\right)$.

The asymptotic approximation is also useful for the following independent proposal Metropolis-Hastings scheme. Initialize $\theta^{(0)} \sim T_3(\boldsymbol{\mu}, \Sigma)$ and notice that a prior $P_\pi(\boldsymbol{\pi})$ on $\boldsymbol{\pi}$ induces a prior $P_\theta(\boldsymbol{\theta}) = P_\pi(\boldsymbol{\pi}(\boldsymbol{\theta})) \times |G(\boldsymbol{\theta})|$ on $\boldsymbol{\theta}$, where $G(\boldsymbol{\theta})$ is as above. At iteration j, perform the following steps:

1. Propose $\boldsymbol{\theta}^* \sim T_3(\boldsymbol{\mu}, \Sigma)$ and let $\boldsymbol{\pi}^* = \boldsymbol{\pi}(\boldsymbol{\theta}^*)$.

2. Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\min\{1, \lambda\}$, where

$$\lambda = \frac{P(\mathbf{Y}|\boldsymbol{\pi}^*, \boldsymbol{\nu}) P_\pi(\boldsymbol{\pi}^*) |G(\boldsymbol{\theta}^*)|}{P(\mathbf{Y}|\boldsymbol{\pi}^{(j-1)}, \boldsymbol{\nu}) P_\pi(\boldsymbol{\pi}^{(j-1)}) |G(\boldsymbol{\theta}^{(j-1)})|} \frac{T_3(\boldsymbol{\theta}^{(j-1)}; \boldsymbol{\mu}, \Sigma)}{T_3(\boldsymbol{\theta}^*; \boldsymbol{\mu}, \Sigma)} \tag{5}$$

   Otherwise, set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$.

Posterior samples can be obtained by discarding some burn-in samples and repeating the process until practical convergence is achieved.

4 Results

We assess the performance of our approach in simulations and two experimental data sets. We obtained the two human sample K562 replicates[3] from the RGASP project (`www.gencodegenes.org/rgasp`), and two ENCODE Project Consortium (2004) replicated samples obtained from A549 cells (accession number wgEncodeEH002625[4]). The sequences were aligned with TopHat (Trapnell et al., 2009) to the human genome hg19, using the default parameters and an average insert size of 200bp. We compare our results with Cufflinks (Trapnell et al., 2012).

[3] `ftp.sanger.ac.uk/pub/gencode/rgasp/RGASP1/inputdata/human_fastq`

[4] `genome.ucsc.edu/ENCODE`

4.1 Simulation study

We generated human genome-wide RNA-seq data, setting the simulations to resemble the K562 RGASP data in order to keep them as realistic as possible. We set $\boldsymbol{\pi}$ for each gene with 2 or more variants to its estimated value in the K562 data using our approach and $q_d = 2$, and simulated a number of fragments with expected value equal to the observed number of sequences in the K562 sample. We set the fragment start and length distributions P_S and P_L to their K562 estimates (Figure 3, left). We estimated $\boldsymbol{\pi}$ from the simulated data using our approach with prior parameters $q_d = 1$ and $q_d = 2$ and Cufflinks.

Table 3 reports the absolute and square errors ($|\pi_d - \hat{\pi}_d|$ and $(\pi_d - \hat{\pi}_d)^2$) averaged across all $19,951$ isoforms and 100 simulated datasets. We also report the average squared bias and variance. Compared to Cufflinks, Casper reduces the MAE by around 20% with $q_d = 1$ and over 200% fold with $q_d = 2$. The improvement in MSE is even more pronounced, with an over 5 fold improvement of $q_d = 2$ over Cufflinks. The main advantage for $q_d = 1$ lies in a reduced bias, whereas $q_d = 2$ improves both bias and variance substantially.

Figure 2 (a)-(b) compare the simulation truth *vs.* the Casper ($q_d = 2$) and Cufflinks estimates. We appreciate that setting $q_d = 2$ pushes the posterior mode away from the boundaries 0 and 1, stabilizing the estimates. The average Pearson correlations between the truth and the estimates across the 100 simulations were 0.929, 0.798 and 0.709 for Casper with $q_d = 2$, $q_d = 1$ and Cufflinks, respectively. Figure 2 (c) shows the error for each transcript as a function of the reads per kilobase per million (RPKM), a measure of overall gene expression. It is worth noticing that Casper improves estimates at all RPKM values. Figure 2 (d) assesses the error *vs.* the mean pairwise difference between variants in a gene (number of base pairs not shared). When all variants in a gene share most exons this difference is small, and intuitively it indicates that variants are hard to distinguish. Conversely, a large difference indicates that many exons are specific to a single variant, which facilitates the estimation problem. Casper estimates are more accurate than Cufflinks, the MAE decreasing by as much as 0.1 even for genes with very similar variants.

We also assessed the frequentist coverage probabilities for the 95% credibility intervals given in Section 3, finding that in 94.0% of the cases they contained the true value.

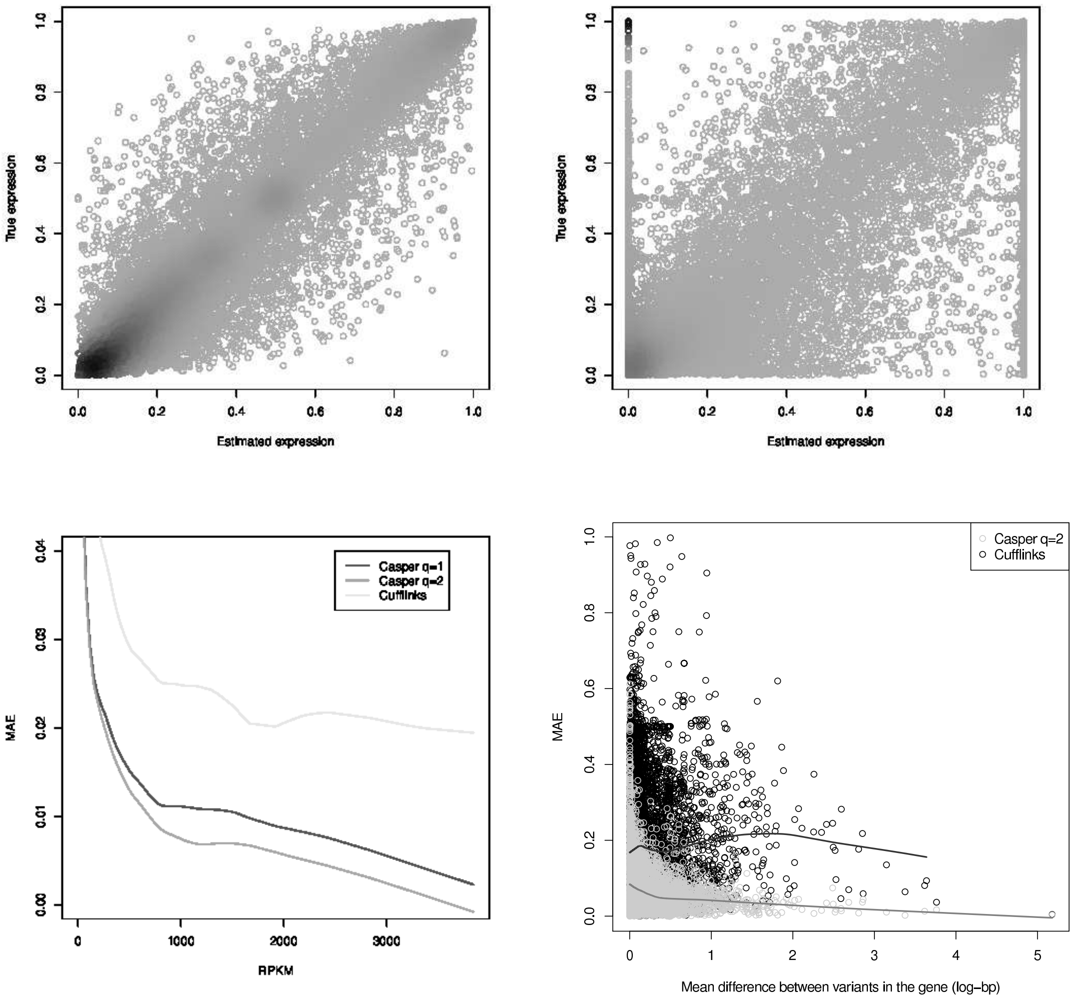

Figure 2: Simulation study. Expression estimates vs. simulation truth for Casper with $q_d = 2$ (a) and Cufflinks (b). Mean absolute error vs. RPKM (c) and the log mean base pair difference between variants in a gene (d).

	MAE	MSE	Bias sqrt	Variance
Casper ($q_d = 1$)	0.126	0.066	0.034	0.032
Casper ($q_d = 2$)	0.070	0.014	0.010	0.005
Cufflinks	0.156	0.076	0.048	0.028

Table 3: Mean absolute and square errors, bias and variance for simulation study.

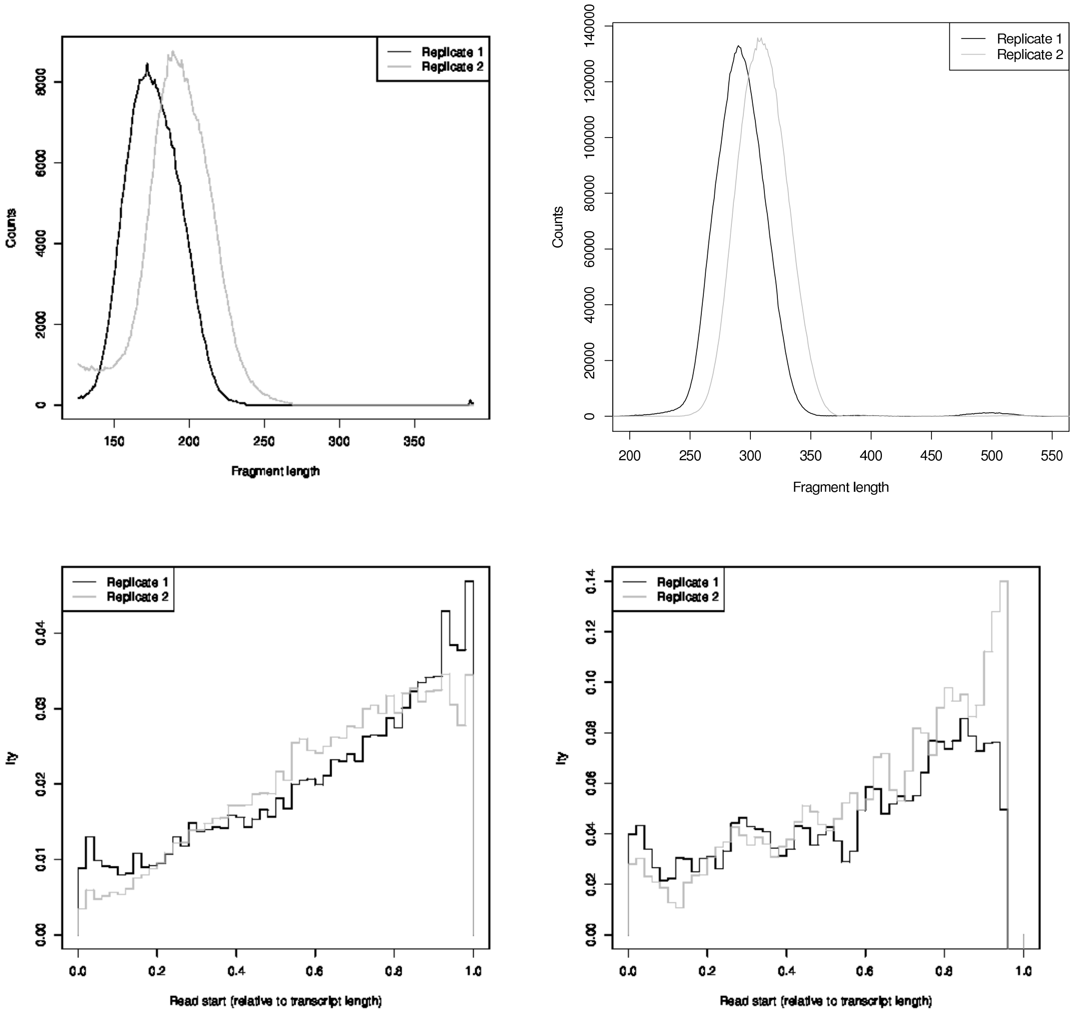

Figure 3: Estimated fragment length (top) and start (bottom) distributions in K562 data (left) and A549 data (right).

4.2 Experimental data from RGASP project

The two K562 replicates were sequenced in 2009 with Solexa sequencing. The read length was 75bp and the mean fragment length indicated in the documentation is 200bp for both replicates. Figure 3 (top, left) shows the estimated fragment length distributions. We observe that the mean length differs significantly from 200bp, and that there are important differences between replicates. Replicate 2 shows a heavy left tail that indicates a subset of fragments substantially shorter than average. This distributional shape cannot be captured with the usual parametric distributions. Figure 3 (left, bottom) shows the relative start distribution. We observe more sequences located near the transcript end in replicate 1, *i.e.* a higher 3' bias. The differences between replicates illustrate the need of flexibly modelling these distributions for each sample separately.

We estimated the expression of all known human splicing variants (genome version hg19) for the two replicated samples separately. Figure 4 (top) compares the estimates obtained in the two samples when using Casper ($q_d = 2$) and Cufflinks. The consistency between samples is higher for Casper, with Pearson correlation 0.945 *versus* a 0.899 for Cufflinks. The mean absolute differences between replicates were 0.061 and 0.076 for Casper and Cufflinks (respectively), a 20% reduction in favor of Casper. These results suggest that Casper can provide advantages even with data from earlier sequencing technologies.

4.3 Experimental data from ENCODE project

The two A549 replicated samples were sequenced in 2012 using Illumina HiSeq 2000. The read length was 101bp and the average fragment length was roughly 300bp (Figure 3, top right). These are substantially longer than the 2009 samples from Section 4.2, and reflect the improvement in sequencing technologies. Similar to Section 4.2, Figure 3 reveals important differences in the fragment length (top, right) and start (bottom, right) distributions between samples.

Figure 4 (bottom) compares the estimates obtained in the two replicates when using Casper ($q_d = 2$) and Cufflinks. The correlation between replicates was 0.964 for Casper and 0.874 for Cufflinks. These findings are consistent with the simulation study in Section 4.1, where the average correlations were 0.929 and 0.709. The mean absolute differences between replicates were 0.047 and 0.076 (respectively), a 38% reduction in favor of Casper. The findings show that the advantage of modeling exon path counts over pairwise exon connections becomes more pronounced as the technology evolves to sequence longer fragments.

5 Discussion

We proposed a model to estimate the expression of a set of known alternatively spliced variants from RNA-seq data. The model improves upon previous proposals by using exon paths, which are more informative than single or pairwise exon counts, and by flexibly estimating the fragment start and length distributions from the observed data. We provided computationally efficient algorithms for obtaining point estimates, asymptotic credibility intervals and posterior samples.

We found that a fairly uninformative prior with $q_d = 2$ delivers more precise estimates than the typical $q_d = 1$, the latter being equivalent with maximum likelihood estimation. The advantages stem from the usual argument in favor of Bayesian shrinkage: $q_d = 2$ pools the estimates away from the boundaries, therefore reducing the instability of the results. Compared to competing approaches, we observed a reduction in MSE by a factor of 5 in simulations and an increase in correlation between experimental replicates from 0.90 to 0.95 in older studies and from 0.87 to 0.96 in recent studies. The mean absolute difference between replicates decreased by up to 38%. In modern studies we found that roughly 2 sequences out of 3 visited > 2 exons. These findings suggest that the current standard of simply reporting pairwise exon junctions adopted by most public databases is far from optimal. Instead, reporting exon paths would allow researchers to estimate isoform expression at a much higher precision. Given that the methodology is implemented in the R package `casper`, we believe that it should be of great value to practitioners.

A Derivation of exon path probabilities

Here we describe how to compute the probability p_{kd} of observing exon path k for any splicing variant d. Equivalently, we denote d by $\boldsymbol{\delta} = (i_1, \ldots, i_{|\delta|})$, where i_j indicates the j^{th} exon within d. Consider variant $\boldsymbol{\delta}$ after splicing, *i.e.* after removing the introns. The new exon start positions are given by $s^*_1 = 1$ and $s^*_{k+1} = s^*_k + e_{i_k} - s_{i_k} + 1$ for $k = 1, \ldots, |\delta| - 1$. The end of exon k is $s^*_{k+1} - 1$. Denote by S be the read start position, L the fragment length, r the read length, and let $T = s^*_{|\delta|} - 1$ be the transcript length of $\boldsymbol{\delta}$.

The goal is to compute $P(\boldsymbol{\iota}_l = (i_j, \ldots, i_{j+k}), \boldsymbol{\iota}_r = (i_{j'}, \ldots, i_{j'+k'}) | \boldsymbol{\delta})$. We note that both $i_j, \ldots, i_{j+k}$ and $i_{j'}, \ldots, i_{j'+k'}$ must be consecutive exons under

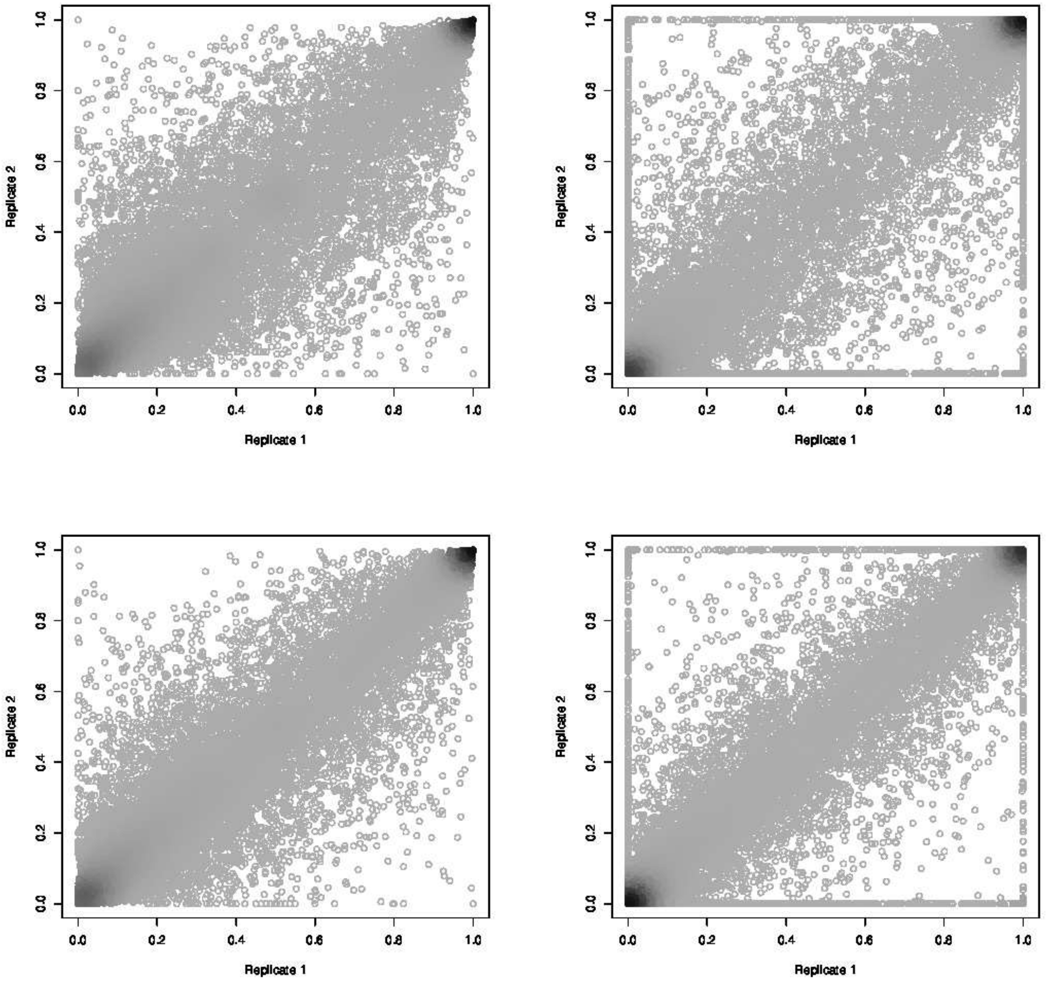

Figure 4: Comparison of the estimated π_d obtained in the two K562 replicates (top) and the ENCODE replicates (bottom). Left: Casper with $q_d = 2$; Right: Cufflinks

variant $\boldsymbol{\delta}$, otherwise the probability of the path is zero. The left read follows the exon path $\boldsymbol{\iota}_l = (i_j, \ldots, i_{j+k})$ if and only if the read:

1. Starts in exon j, *i.e.* $s^*_j \leq S \leq s^*_{j+1} - 1$
2. Ends in exon $j+k$, *i.e.* $s^*_{j+k} \leq S + r - 1 \leq s^*_{j+k+1} - 1$

Similarly, the right read follows $\boldsymbol{\iota}_r = (i_{j'}, \ldots, i_{j'+k'})$ if and only if $s^*_{j'} \leq S + L - r \leq s_{j'+1} - 1$ and $s^*_{j'+k'} \leq S + L - 1 \leq s^*_{j'+k'+1} - 1$. This implies that the desired probability can be written as $P(a_1 \leq S \leq b_1, a_2 \leq S + L \leq b_2 | \boldsymbol{\delta})$, where

$$\begin{aligned} a_1 &= \max\{s^*_j, s^*_{j+k} - r + 1\} \\ b_1 &= \min\{s^*_{j+1} - 1, s^*_{j+k+1} - r\} \\ a_2 &= \max\{s^*_{j'} + r, s^*_{j'+k'} + 1\} \\ b_2 &= \min\{s^*_{j'+1} + r - 1, s^*_{j'+k'+1}\}. \end{aligned} \tag{6}$$

Assuming that the distribution of (S, L) depends on $\boldsymbol{\delta}$ only through its transcript length T, we can write $P(a_1 \leq S \leq b_1, a_2 \leq S + L \leq b_2 | T) =$

$$\sum_l P(a_1 \leq S \leq b_1, a_2 \leq S + L \leq b_2 | T, L = l) P(L = l | T) =$$

$$\sum_l P(\max\{a_1, a_2 - L\} \leq S \leq \min\{b_1, b_2 - L\} | T, L = l) P(L = l | T). \tag{7}$$

In order to evaluate (7) we need to estimate the fragment length distribution $P(L = l|T)$ and the distribution of the read start position S given L. We assume that $P(L|T) = P(L = l)\mathrm{I}(l \leq T)/P(L \leq T)$, *i.e.* the conditional distribution of L given T is simply a truncated version of the marginal distribution. Further, notice that the fragment end must happen before the end of the transcript, *i.e.* $S + L - 1 \leq T$ or equivalently the relative start position is truncated $S/T \leq S_T = (T - L + 1)/T$. The relative start distribution is therefore truncated, *i.e.* $P(\frac{S}{T} \leq z | T, L = l) = \frac{\varphi(\min\{z, S_T\})}{\varphi(S_T)}$, where $\varphi(z) = P(\frac{S}{T} \leq z)$ is the distribution of the relative read start $\frac{S}{T}$.

Under these assumptions, the probability of observing the exon path $\boldsymbol{\iota}_l = (i_j, \ldots, i_{j+k})$, $\boldsymbol{\iota}_r = (i_{j'}, \ldots, i_{j'+k'})$ under variant $\boldsymbol{\delta}$ is equal to

$$\sum_l \left[\frac{\varphi\left(\min\{\frac{b_1}{T}, \frac{b_2 - l}{T}, S_T\}\right) - \varphi\left(\min\left\{\max\{\frac{a_1 - 1}{T}, \frac{a_2 - l - 1}{T}\}, S_T\right\}\right)}{\varphi\left(S_T\right)} \right]_+ P(L = l | T),$$

where a_1, b_1, a_2 and b_2 are given in (6). Given that highly precise estimates of $P(L = l)$ and $\varphi(\cdot)$ are typically available, for computational simplicity we treat them as known and plug them into (8).

B EM algorithm derivation

1. **E-step**

   Let $\delta_i \in \{1, \ldots, |\nu|\}$ be latent variables indicating the variant that reads $i = 1, \ldots, N$ come from. The augmented log-posterior is proportional to to

   $$l_0(\boldsymbol{\pi}|\mathbf{y}, \boldsymbol{\delta}) = \log P(\boldsymbol{\pi}|\boldsymbol{\nu}) + \log P(\mathbf{y}, \boldsymbol{\delta}|\boldsymbol{\pi}) = \sum_{d=1}^{|\nu|}(q_d - 1)\log(\pi_d) + \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} \mathrm{I}(\delta_i = d)\left[\log(p_{y_i d}) + \log(\pi_d)\right]. \quad (8)$$

   Considering δ_i as a random variable, the expected value of (8) given $\mathbf{y}$ and $\boldsymbol{\pi} = \boldsymbol{\pi}^{(j)}$ is equal to

   $$E\left(l_0(\boldsymbol{\pi}'|\mathbf{y}, \boldsymbol{\delta})|\mathbf{y}, \boldsymbol{\pi}^{(j)}\right) = \sum_{d=1}^{|\nu|}(q_d - 1)\log(\pi_d) + \sum_{i=1}^{N}\sum_{d=1}^{|\nu|} P(\delta_i = d|y_i, \boldsymbol{\pi}^{(j)})\left(\log(p_{y_i d}) + \log(\pi'_d)\right) \quad (9)$$

2. **M-step**

   The goal is to maximize (9) with respect to $\boldsymbol{\pi}'$. Let $\gamma_{id} = P(\delta_i = d|y_i, \boldsymbol{\pi}^{(j)})$ and re-parameterize $\pi_{|\nu|} = 1 - \sum_{d=1}^{|\nu|-1} \pi_d$. Setting the partial derivatives with respect to π'_d to zero gives the system of equations

   $$\frac{\pi'_d}{1 - \sum_{d=1}^{|\nu|-1} \pi'_d} = \frac{q_d - 1 + \sum_{i=1}^{N} \gamma_{id}}{q_{|\nu|} - 1 + \sum_{i=1}^{N} \gamma_{i|\nu|}},$$

   which has the trivial solution $\pi'_d \propto q_d - 1 + \sum_{i=1}^{N} \gamma_{id}$. By plugging in $\gamma_{id} = p_{y_i d}\pi_d^{(j)} / \sum_{d=1}^{|\nu|} p_{y_i d}\pi_d^{(j)}$, we obtain

   $$\pi'_d \propto q_d - 1 + \sum_{i=1}^{N} \frac{p_{y_i d}\pi_d^{(j)}}{\sum_{d=1}^{|\nu|} p_{y_i d}\pi_d^{(j)}}.$$

   Finally, since $x_k = \sum_{i=1}^{N} \mathrm{I}(y_i = k)$ we can group all y_i's taking the same value and find the maximum as

   $$\pi'_d \propto q_d - 1 + \sum_{k=1}^{|\mathcal{P}|} x_k \frac{p_{kd}\pi_d^{(j)}}{\sum_{d=1}^{|\nu|} p_{kd}\pi_d^{(j)}}, \quad (10)$$

   re-normalizing $\boldsymbol{\pi}'$ so that $\sum_{d=1}^{|\nu|} \pi'_d = 1$.

C Asymptotic posterior approximation

Here we derive an asymptotic approximation to $P(\boldsymbol{\pi} \mid \boldsymbol{\nu}, \mathbf{Y})$, the posterior distribution of the splicing variants expression $\boldsymbol{\pi}$ conditional on a model $\boldsymbol{\nu}$ and the observed data $\mathbf{Y}$. Given that $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{|\nu|}) \in [0,1]^{|\nu|}$ with $\sum_{i=1}^{|\nu|} \pi_i = 1$, we re-parameterize to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{|\nu|-1}) \in \Re^{|\nu|-1}$, where $\theta_d = \log\left(\frac{\pi_{d+1}}{\pi_1}\right)$ for $d = 1, \ldots, |\nu| - 1$. The goal is to approximate $P(\boldsymbol{\theta} \mid \boldsymbol{\nu}, \mathbf{Y}) \sim N(\boldsymbol{\mu}, \Sigma)$. For notational simplicity, in the remainder of the section we drop the conditioning on $\boldsymbol{\nu}$.

A prior $P_\pi(\boldsymbol{\pi})$ induces a prior $P_\theta(\boldsymbol{\theta}) = P_\pi(\boldsymbol{\pi}(\boldsymbol{\theta})) \times |G(\boldsymbol{\theta})|$ on $\boldsymbol{\theta}$, where $G(\boldsymbol{\theta})$ is the matrix with (d,l) element $G_{dl} = \frac{\partial}{\partial\theta_l}\pi_d(\boldsymbol{\theta})$ and inverse transform $\pi_1(\boldsymbol{\theta}) = \left(1 + \sum_{j=1}^{|\nu|-1} e^{\theta_j}\right)^{-1}$, $\pi_d(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta})\exp\{\theta_{d-1}\}$ for $d > 1$.

Define $f(\boldsymbol{\theta}) = \log\left(P(\mathbf{Y}|\boldsymbol{\theta})\right) + \log\left(P_\theta(\boldsymbol{\theta})\right)$. Up to an additive constant, $f(\boldsymbol{\theta})$ is equal to the target log-posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}$. We center the approximating Normal at the posterior mode, *i.e.* $\boldsymbol{\mu} = \text{argmax}_{\boldsymbol{\theta} \in \Re^{|\nu|-1}} f(\boldsymbol{\theta})$. We set $\Sigma = S^{-1}$ where S is the Hessian of $f(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\mu}$ with (l,m) element $S_{lm} = \frac{\partial^2}{\partial\theta_l\partial\theta_m} f(\boldsymbol{\theta})$. We approximate $\mu_d = \log\left(\frac{\pi^*_{d+1}}{\pi_d}\right)$, where $\boldsymbol{\pi}^*$ is the posterior mode for $\boldsymbol{\pi}$ provided by the EM algorithm.

Under a $\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{q})$ prior, simple algebra gives $\sigma_{lm} = \frac{\partial^2}{\partial\theta_l\partial\theta_m} f(\boldsymbol{\theta}) =$

$$\sum_{k=1}^{|\mathcal{P}|} x_k \frac{\left(\sum_{d=1}^{|\nu|} p_{kd} H_{dlm}\right)\left(\sum_{d=1}^{|\nu|} p_{kd}\pi_d(\boldsymbol{\theta})\right) - \left(\sum_{d=1}^{|\nu|} p_{kd} G_{dl}\right)\left(\sum_{d=1}^{|\nu|} p_{kd} G_{dm}\right)}{\left(\sum_{d=1}^{|\nu|} p_{kd}\pi_d(\boldsymbol{\theta})\right)^2} + \sum_{d=1}^{|\nu|} (q_d - 1)\frac{H_{dlm}\pi_d(\boldsymbol{\theta}) - G_{dl}G_{dm}}{\pi_d(\boldsymbol{\theta})^2}, \quad (11)$$

where $x_k = \sum_{i=1}^N \mathrm{I}(y_i = k)$ is the number of reads following exon path k, $p_{kd} = P(Y_i = k \mid \delta = d)$ is the probability of observing path k under variant d, the gradient for $\pi_d(\boldsymbol{\theta})$ is $G_{dl} = \frac{\partial}{\partial\theta_l}\pi_d(\boldsymbol{\theta})$ as before and the Hessian is $H_{dlm} = \frac{\partial^2}{\partial\theta_l\partial\theta_m}\pi_d(\boldsymbol{\theta})$.

We complete the derivation by providing expressions for G_{dl} and H_{dlm}. Let $s(\boldsymbol{\theta}) = 1 + \sum_{j=1}^{|\nu|-1} e^{\theta_j}$, then $G_{dl} =$

$$\begin{array}{c} \dfrac{-e^{\theta_l}}{s(\boldsymbol{\theta})^2}, \text{ if } d = 1 \\[2ex] \dfrac{-e^{\theta_{d-1}+\theta_l}}{s(\boldsymbol{\theta})^2} + \mathrm{I}(l = d-1)\dfrac{e^{\theta_l}}{s(\boldsymbol{\theta})}\text{ , if } d \geq 2 \end{array} \qquad (12)$$

and $H_{dlm} =$

$$\begin{cases} \dfrac{2e^{\theta_l+\theta_m}}{s(\boldsymbol{\theta})^3} - \mathrm{I}(l=m)\dfrac{e^{\theta_l}}{s(\boldsymbol{\theta})^2} & \text{, if } d=1 \\ \dfrac{2e^{\theta_{d-1}+\theta_l+\theta_m}}{s(\boldsymbol{\theta})^3} - \mathrm{I}(l=d-1)\dfrac{e^{\theta_l+\theta_m}}{s(\boldsymbol{\theta})} & \text{, if } d \geq 2, m \neq l, m \neq d-1 \\ \dfrac{-2e^{2\theta_m}}{s(\boldsymbol{\theta})^2} + \dfrac{2e^{3\theta_m}}{s(\boldsymbol{\theta})^3} + \dfrac{e^{\theta_m}}{s(\boldsymbol{\theta})} - \dfrac{2e^{2\theta_m}}{s(\boldsymbol{\theta})^2} & \text{, if } d \geq 2, m = l, m = d-1 \\ \dfrac{-e^{\theta_{d-1}+\theta_l}}{s(\boldsymbol{\theta})^2} + \dfrac{2e^{\theta_{d-1}+\theta_l+\theta_m}}{s(\boldsymbol{\theta})^3} & \text{, otherwise.} \end{cases} \tag{13}$$

Acknowledgements

D.R. was partially supported by grants R01 CA158113-01 from the NIH (USA) and MTM2012-38337 from the Ministerio de Economía y Competitividad (Spain). C.S. was partially supported by AGAUR Beatriu de Pinós fellowship BP-B 00068 (Spain). M.K. and A.S. were partially supported by Bayerisches Arbeitsministerium foreign trainee fellowship (Germany). The authors wish to thank Modesto Orozco for useful discussions.

References

A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten. Global and unbiased detection of splice junctions from rna-seq data. *Genome Biology*, 11(3): R34, 2010.

Benjamin J. Blencowe. Alternative splicing: New insights from global analyses. *Cell*, 126:37–47, 2006.

G. Casella and R. L. Berger. *Statistical inference*. Duxbury Press, 2 edition, June 2001.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–40, Oct 2004.

M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M.J. Koziol, A. Gnirke, C. Nusbaum, J.L. Rinn, E.S. Lander, and

A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnoly*, 28(5):503–10, May 2010.

R.A. Holt and S.J.M. Jones. The new paradigm of flow cell sequencing. *Genome Research*, 18(6):839–46, Jun 2008.

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

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282): 457–481, 1958.

Y. Katz, E.T. Wang, E.M. Airoldi, and C.B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 12 2010.

V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron. Exact transcriptome reconstruction from short sequence reads. In *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, WABI '08, pages 50–63, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87360-0. doi: 10.1007/978-3-540-87361-7_5. URL `http://dx.doi.org/10.1007/978-3-540-87361-7_5`.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–67, 2009.

A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628, 2008.

S. Pepke, B. Wold, and Mortazavi A. Computation for chip-seq and rna-seq studies. *Nature Methods*, 6(11 Suppl):S22–S32, 2009.

J. Salzman, H. Jiang, and W.H. Wong. Statistical modeling of rna-seq data. *Statistical Science*, 26(1):62–83, 2011.

T. Therneau and T. Lumley. *survival: Survival analysis, including penalised likelihood.*, 2011. URL `http://CRAN.R-project.org/package=survival`. R package version 2.36-10.

C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3):562–78, Mar 2012.

J. Wu, M. Akerman, S. Sun, W.R. McCombie, A.R. Krainer, and M.Q. Zhang. Splicetrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 27(21):3010–6, Nov 2011a.

Z. Wu, X. Wang, and X. Zhang. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics*, 27(4): 502–8, Feb 2011b.

Y. Xing, T. Yu, Y. N. Wu, M. Roy, J. Kim, and C. Lee. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*, 34(10):3150–60, 2006.