# SUPPLEMENTARY MATERIAL FOR “RNA-SEQ GENE EXPRESSION ESTIMATION WITH READ MAPPING UNCERTAINTY”

## Details of the EM algorithm for our RNA-Seq model

We use the EM algorithm to find the the values of $\theta$ that maximize the observed data likelihood:

$$P(r|\theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \theta_i P(r_n|G_n = i),$$

where

$$P(r_n|G_n = i) = \frac{1}{\ell_i} \sum_j P(r_n|Z_{nij} = 1),$$

and where we have assumed a uniform RSPD and a strand-specific protocol for ease of presentation (our experiments in this paper use a non-strand-specific model, for which the equations are similar).

Key to the EM algorithm is the expected value of the complete data log likelihood function, given current values for the parameters. The complete data log likelihood may be written as

$$\log P(r, z|\theta) = \sum_{n,i,j} z_{nij} \log \left( \frac{\theta_i}{\ell_i} P(r_n|Z_{nij} = 1) \right).$$

The $Q$ function for the EM algorithm is thus

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{Z|r,\theta^{(t)}}[\log P(r, Z|\theta)] \\ &= \sum_{n,i,j} E_{Z|r,\theta^{(t)}}[Z_{nij}] \log \left( \frac{\theta_i}{\ell_i} P(r_n|Z_{nij} = 1) \right). \end{aligned}$$

During the E-step we calculate the expected values of the $Z_{nij}$ variables:

$$\begin{aligned} E_{Z|r,\theta^{(t)}}[Z_{nij}] &= P(Z_{nij} = 1|r, \theta^{(t)}) \\ &= \frac{(\theta_i^{(t)}/\ell_i) P(r_n|Z_{nij} = 1)}{\sum_{i',j'} (\theta_{i'}^{(t)}/\ell_{i'}) P(r_n|Z_{ni'j'} = 1)}. \end{aligned}$$

Finally, the M-step has us maximize $Q(\theta|\theta^{(t)})$ with respect to $\theta$:

$$\theta_i^{(t+1)} = \frac{E_{Z|r,\theta^{(t)}}[C_i]}{N},$$

where

$$C_i = \sum_{n,i,j} Z_{nij}.$$

The EM algorithm consists of alternating between the E and M steps until convergence. We start the algorithm with $\theta^{(0)}$ corresponding to uniform $\tau$ isoform expression levels and stop when the log likelihood is no longer increasing significantly.

When a non-uniform RSPD is used, we also estimate the RSPD parameters during the EM algorithm. The parameters of the RSPD, $\phi_1 \ldots \phi_B$, are initially set to a uniform distribution, and are updated during the M-step by

$$\phi_b^{(t+1)} = \frac{E_{Z|r,\theta^{(t)},\phi^{(t)}}[C'_b]}{\sum_{b'} E_{Z|r,\theta^{(t)},\phi^{(t)}}[C'_{b'}]},$$

where

$$C'_b = B \cdot \sum_{n,i,j,i\neq 0} Z_{nij} \cdot u \left( \begin{array}{c} \min\left(\frac{j}{\ell_i}, \frac{b}{B}\right) - \\ \max\left(\frac{j-1}{\ell_i}, \frac{b-1}{B}\right) \end{array} \right)$$

and $u(x) = x$ if $x \geq 0$ and $u(x) = 0$ otherwise.

## Proof of concavity of likelihood function

We prove that the likelihood function for our model is concave with respect to $\theta$ and thus that the EM algorithm is guaranteed to reach a global maximum in the parameter space of $\theta$. Our proof is similar to that given in (Jiang and Wong, 2009) for the concavity of their likelihood function. Our log likelihood function is

$$\log P(r|\theta) = \sum_{n=1}^{N} \log \sum_i \theta_i P(r_n|G_n = i).$$

Because the sum of concave functions is also a concave function, we need only prove that

$$f(\theta) = \log \sum_i \theta_i P(r_n|G_n = i)$$

is concave. Let $H(\theta)$ be the Hessian matrix for $f(\theta)$:

$$\begin{aligned} H_{jk}(\theta) &= \frac{\partial^2 \log \sum_i \theta_i P(r_n|G_n = i)}{\partial\theta_j \partial\theta_k} \\ &= -\frac{P(r_n|G_n = j) P(r_n|G_n = k)}{(\sum_i \theta_i P(r_n|\ G_n = i))^2}. \end{aligned}$$

We may express $H(\theta)$ as $-c(\theta)x'x$, where

$$\begin{aligned} c(\theta) &= \frac{1}{(\sum_i \theta_i P(r_n|G_n = i))^2} \\ x &= [P(r_n|G_n = 0), \ldots, P(r_n|G_n = m)]. \end{aligned}$$

Noting that $c(\theta) > 0$, it follows that $\forall y = [y_0, \ldots, y_m]$,

$$\begin{aligned} yH(\theta)y' &= y(-c(\theta)x'x)y' \\ &= -c(\theta)(yx')(yx')' \\ &= -c(\theta)(yx')^2 \\ &\leq 0. \end{aligned}$$

Therefore, $H(\theta)$ is negative semidefinite and both $f(\theta)$ and $P(r|\theta)$ are concave.

## Treatment of repetitive read sequences

In practice, we encounter a significant number of repetitive reads, which have many alignments to the reference sequences. Because of the large amount of uncertainty in their mapping, these reads contribute little information regarding expression levels, yet require significant additional computation. Thus, to reduce compute time, we filter out all alignments for reads that map to $T$ or more genes ($T = 100$, typically). To compensate, we mark all potential read start positions that align to these repetitive reads. These positions are not allowed to generate reads in the model, effectively reducing the length of each isoform (although we do not filter other non-repetitive reads that start at these positions). We then correct for this heuristic by boosting the estimated expression levels of each isoform according to the ratio of its true length, $\ell_i$, and its reduced length $\ell'_i$. This adjustment can be thought of as a generalization of the “mappability” correction used in (Morin *et al.*, 2008), which corresponds to setting $T = 1$. If the initial estimate (via EM) of the fraction of reads mapping to isoform $i$ is $\theta'_i$, we estimate the true fraction, $\theta_i$, as:

$$\theta_i = \frac{\ell_i}{\ell'_i}\theta'_i.$$

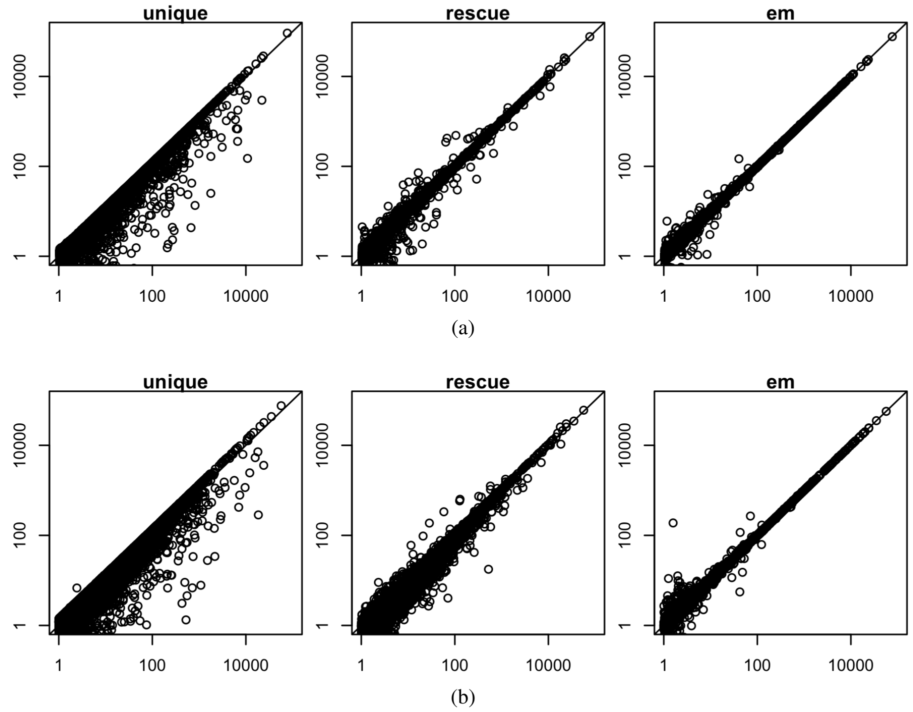
The noise fraction, $\theta_0$, is then computed as

$$\theta_0 = 1 - \sum_i \frac{\ell_i}{\ell'_i}\theta'_i.$$

## REFERENCES

Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**(8), 1026–1032.

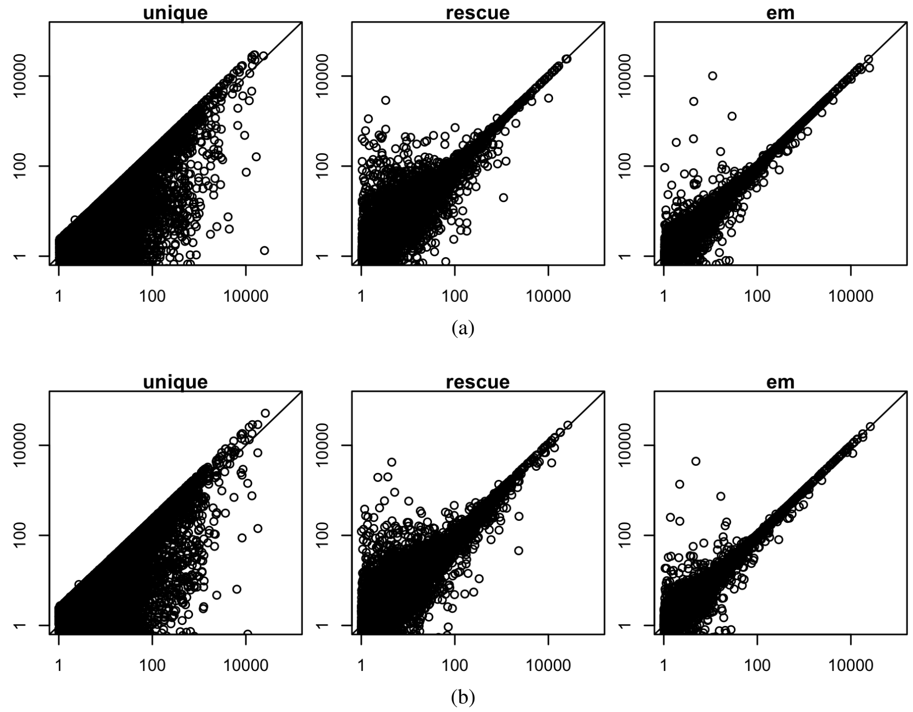
Morin, R. D., Marra, M. A., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., and Jones, S. J. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, **45**(1), 81–94.



(a)

(b)

Fig. 1: Gene expression estimates (y-axis) vs. sample values (x-axis) for the simulated mouse liver RNA-Seq data set. Comparisons are given for both (a) $\nu$ and (b) $\tau$. Values on the axes are given in terms of NPM ($\nu$) or TPM ($\tau$).



(a)

(b)

Fig. 2: Gene expression estimates (y-axis) vs. sample values (x-axis) for the simulated maize RNA-Seq data set. Comparisons are given for both (a) $\nu$ and (b) $\tau$. Values on the axes are given in terms of NPM ($\nu$) or TPM ($\tau$).

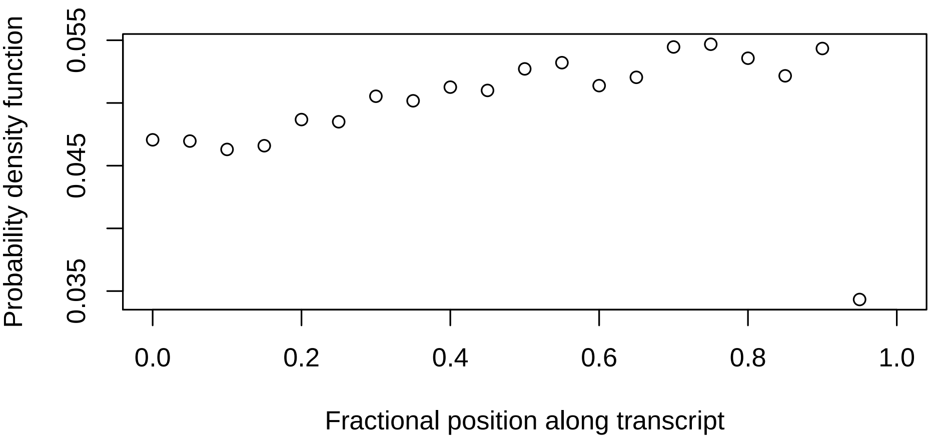


Fig. 3: The RSPD estimated by our method from all reads in the mouse liver data set. This RSPD indicates a 3’ bias in the protocol used for this data set. The single data point at the 3’ end is lower most likely because it is difficult to map short reads that extend significantly into a poly(A) tail.

| | | | | Sample gene expression in NPM ($\nu$) or TPM ($\tau$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $[1, 10)$ | $[10, 10^2)$ | $[10^2, 10^3)$ | $[10^3, 10^4)$ | $[10^4, 10^5)$ | All |
| (A) | $\nu$ | | N | 5242 | 5720 | 1116 | 115 | 6 | 12199 |
| | | MPE | em uniform | 2.2 | 0.9 | 0.5 | 0.3 | 0.2 | 1.1 |
| | | | em rspd | 2.2 | **0.8** | **0.4** | **0.2** | 0.2 | **1.0** |
| | | EF | em uniform | 18.6 | 1.7 | 0.6 | 0.9 | 0.0 | 8.9 |
| | | | em rspd | 18.7 | 1.8 | 0.8 | 0.9 | 0.0 | 9.0 |
| | $\tau$ | | N | 5976 | 4584 | 984 | 114 | 14 | 11672 |
| | | MPE | em uniform | 2.4 | 1.1 | 0.6 | 0.5 | 0.7 | 1.5 |
| | | | em rspd | **2.3** | **1.0** | **0.4** | **0.3** | **0.3** | **1.3** |
| | | EF | em uniform | 25.4 | 5.8 | 1.1 | 0.9 | 0.0 | 15.4 |
| | | | em rspd | 24.7 | 5.1 | 1.2 | 0.9 | 0.0 | 14.8 |
| (B) | $\nu$ | | N | 8237 | 7165 | 1234 | 129 | 10 | 16775 |
| | | MPE | em uniform | 3.4 | 1.4 | 0.7 | 0.6 | 0.8 | 2.1 |
| | | | em rspd | **3.3** | **1.3** | **0.6** | **0.4** | **0.5** | **1.9** |
| | | EF | em uniform | 40.1 | 19.1 | 10.0 | 6.2 | 10.0 | 28.7 |
| | | | em rspd | 39.6 | 17.6 | 5.5 | 2.3 | 0.0 | 27.4 |
| | $\tau$ | | N | 8596 | 6401 | 1102 | 124 | 9 | 16232 |
| | | MPE | em uniform | 3.5 | 1.6 | 0.9 | 0.7 | 1.1 | 2.3 |
| | | | em rspd | **3.3** | **1.4** | **0.7** | **0.4** | **0.7** | **2.1** |
| | | EF | em uniform | 40.8 | 19.2 | 9.6 | 3.2 | 11.1 | 29.8 |
| | | | em rspd | 40.1 | 17.3 | 5.0 | 1.6 | 0.0 | 28.4 |

**Table 1.** Error of the `em uniform` and `em rspd` estimated gene expression levels with respect to sample expression values from simulations of mouse (A) and maize (B) RNA-Seq data with a non-uniform RSPD learned from mouse liver data. Bold values indicate that the estimates are significantly ($p < 0.05$) more accurate, as assessed by a paired Wilcoxon signed rank test.

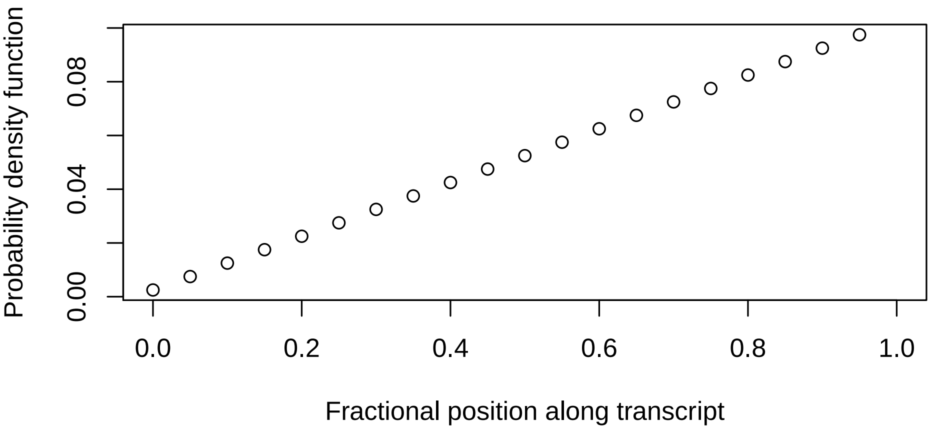
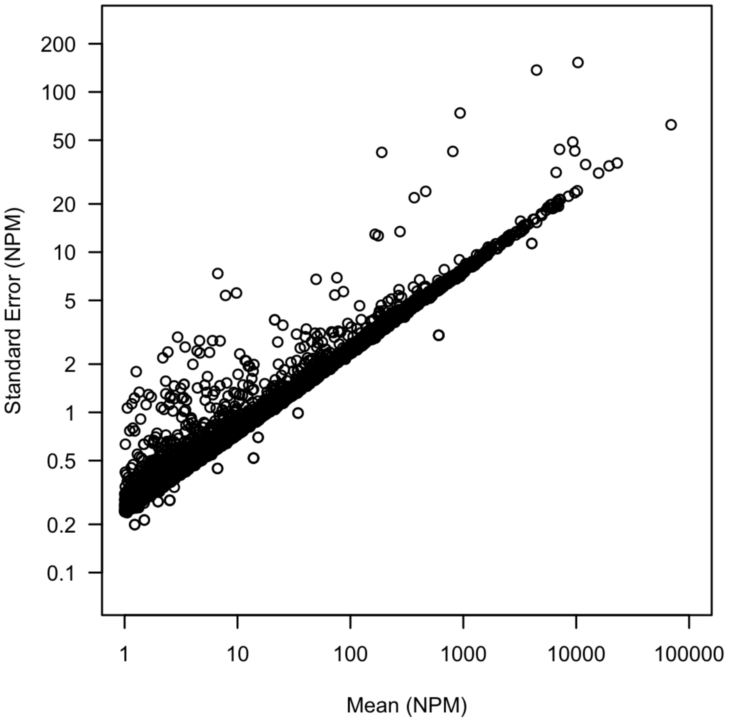


Fig. 4: A synthetic extreme 3' biased RSPD used for testing the benefits of estimating expression with a non-uniform RSPD model.

| | | | | Sample gene expression in NPM ($\nu$) or TPM ($\tau$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $[1, 10)$ | $[10, 10^2)$ | $[10^2, 10^3)$ | $[10^3, 10^4)$ | $[10^4, 10^5)$ | All |
| (A) | $\nu$ | | N | 5518 | 5286 | 1036 | 112 | 9 | 11961 |
| | | MPE | em uniform | 2.7 | 1.1 | 0.6 | 0.5 | 0.6 | 1.5 |
| | | | em rspd | **2.6** | **1.0** | **0.5** | **0.3** | 0.4 | **1.3** |
| | | EF | em uniform | 27.1 | 5.0 | 2.8 | 2.7 | 11.1 | 15.0 |
| | | | em rspd | 24.3 | 3.0 | 0.8 | 0.9 | 0.0 | 12.6 |
| | $\tau$ | | N | 6274 | 4026 | 887 | 111 | 15 | 11313 |
| | | MPE | em uniform | 3.9 | 2.3 | 1.8 | 1.5 | 2.1 | 2.9 |
| | | | em rspd | **2.7** | **1.2** | **0.7** | **0.5** | **0.6** | **1.7** |
| | | EF | em uniform | 42.2 | 26.3 | 13.3 | 7.2 | 13.3 | 33.9 |
| | | | em rspd | 30.8 | 7.5 | 1.0 | 1.8 | 0.0 | 19.9 |
| (B) | $\nu$ | | N | 8957 | 4732 | 995 | 120 | 14 | 14818 |
| | | MPE | em uniform | 4.6 | 1.6 | 0.7 | 0.5 | 0.3 | 3.0 |
| | | | em rspd | **4.0** | **1.3** | **0.6** | **0.5** | 0.6 | **2.5** |
| | | EF | em uniform | 47.8 | 26.8 | 17.0 | 12.5 | 21.4 | 38.7 |
| | | | em rspd | 43.3 | 17.9 | 10.1 | 5.8 | 21.4 | 32.6 |
| | $\tau$ | | N | 9227 | 4941 | 1040 | 113 | 12 | 15333 |
| | | MPE | em uniform | 6.6 | 3.1 | 2.3 | 2.3 | 1.5 | 4.6 |
| | | | em rspd | **5.0** | **1.7** | **0.7** | **0.6** | **0.6** | **3.0** |
| | | EF | em uniform | 56.1 | 37.6 | 27.1 | 23.9 | 16.7 | 47.9 |
| | | | em rspd | 50.2 | 21.1 | 9.2 | 7.1 | 16.7 | 37.7 |

**Table 2.** Error of the `em uniform` and `em rspd` estimated gene expression levels with respect to sample expression values from simulations of mouse (A) and maize (B) RNA-Seq data with a synthetic 3'-biased RSPD. Bold values indicate that the estimates are significantly ($p < 0.05$) more accurate, as assessed by a paired Wilcoxon signed rank test.



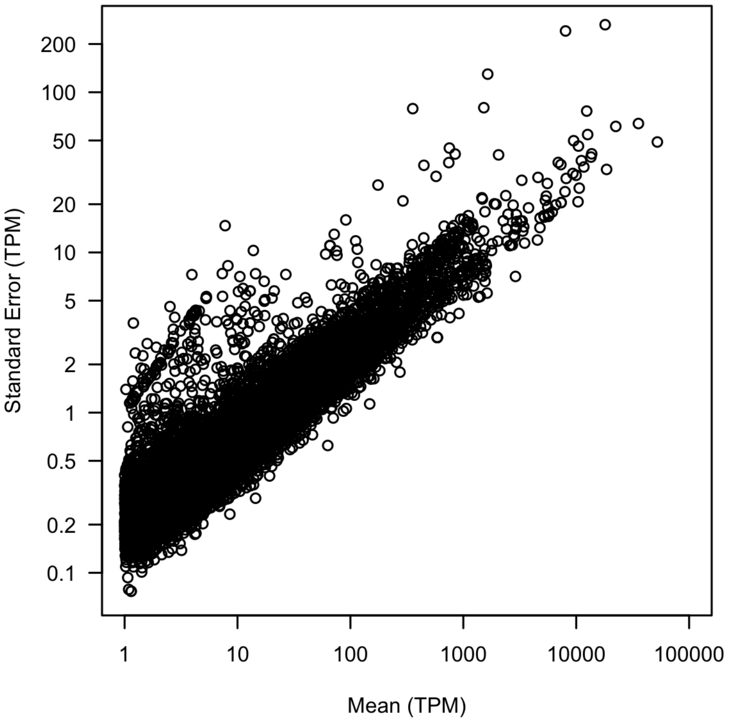


Fig. 5: Standard error vs. mean gene expression level estimate for genes with expression greater than 1 NPM (top) or 1 TPM (bottom) from the mouse liver data. Standard errors were computed by running the EM algorithm on 1000 non-parametric bootstrap samples. For a multinomial model, we expect a linear relationship between the variance and mean of a $\nu$ estimate, which presents itself as a line in log-log coordinates of a standard error vs. mean plot. The standard errors for genes with the same mean $\tau$ estimate are more variable because of varying gene length.

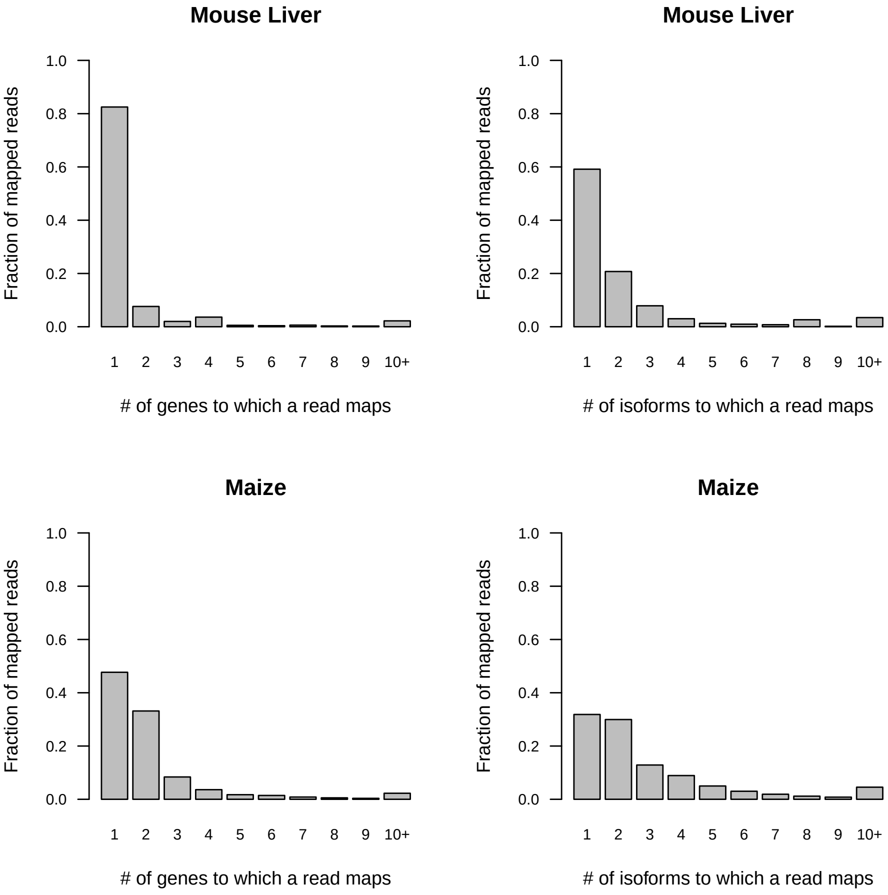


Fig. 6: Distribution of the number of genes and isoforms mapped to by the reads in the mouse liver and maize simulations.