

RNA-Seq: Isoforms Quantification and the Mixture of Beta Regression

Billy Chang, Rafal Kustra, Quaid Morris

Graduate Student Research Day
April 29, 2010

Background on Alternative Splicing

- Gene contains EXONS and introns.



Background on Alternative Splicing

- Gene contains EXONS and introns.



- Remove introns.



Background on Alternative Splicing

- Gene contains EXONS and introns.



- Remove introns.



- Joint Exons.



Background on Alternative Splicing

- Gene contains EXONS and introns.



- Remove introns.



- Joint Exons.



- Exons can be skipped too.



Background on Alternative Splicing

- Gene contains EXONS and introns.



- Remove introns.



- Joint Exons.



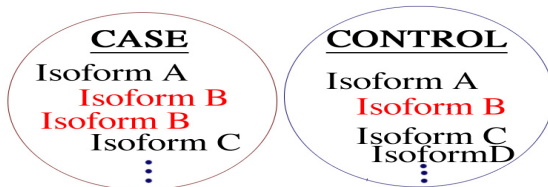
- Exons can be skipped too.



- A single gene can produce multiple proteins.

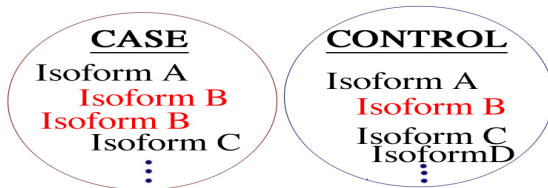
Why Quantify Isoforms?

- Disease associated isoforms.

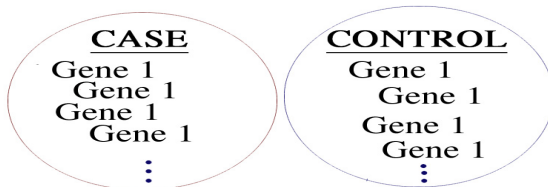


Why Quantify Isoforms?

- Disease associated isoforms.

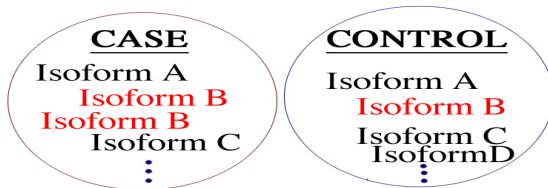


- Gene level...

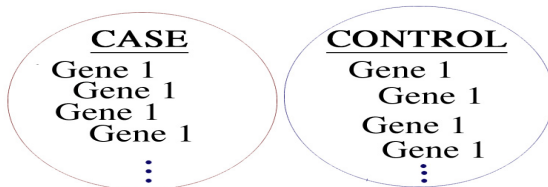


Why Quantify Isoforms?

- Disease associated isoforms.



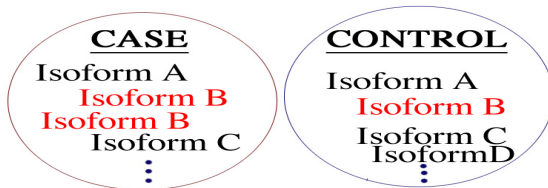
- Gene level...



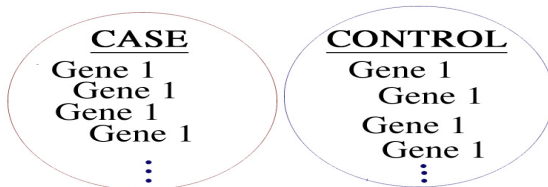
- How to Count?

Why Quantify Isoforms?

- Disease associated isoforms.



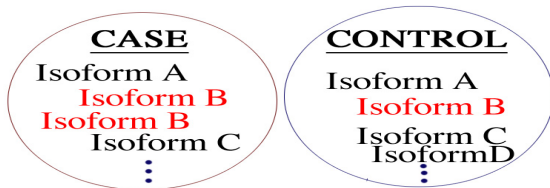
- Gene level...



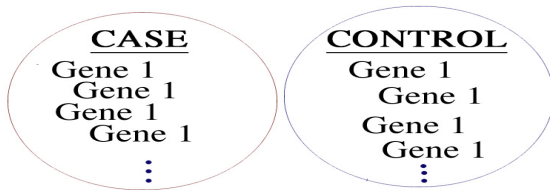
- How to Count?
- Microarray: noisy, design issues.

Why Quantify Isoforms?

- Disease associated isoforms.



- Gene level...





- How to Count?
- Microarray: noisy, design issues.
- Classical Sequencing technology: slow and expensive.



Next Generation Sequencing



Next Generation Sequencing

- 
- Isoform fragmentation (100 - 400 bp):


Next Generation Sequencing

- 
- Isoform fragmentation (100 - 400 bp):

- Sequence one end of each fragment (36 bp).

| | | |
|--------------|--------------|--------------|
| cgtc....ttta | aatt....catg | tata....ggtt |
| cgtc....ttta | aatt....catg | tata....ggtt |

Next Generation Sequencing

- 

- Isoform fragmentation (100 - 400 bp):



- Sequence one end of each fragment (36 bp).



- “Reads”: subsequences of the isoform.



Big Picture

- Isoform from donor, fragment and sequence:

DONOR



Big Picture

- Isoform from donor, fragment and sequence:

DONOR



- Obtain reads.



Big Picture

- Isoform from donor, fragment and sequence:

DONOR



- Obtain reads.

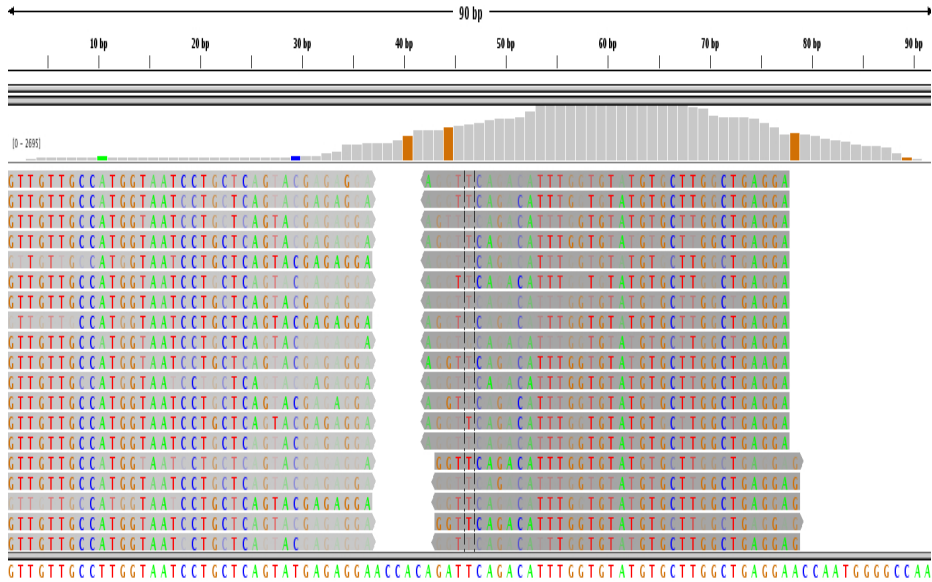


- Align to currently known isoforms (reference)

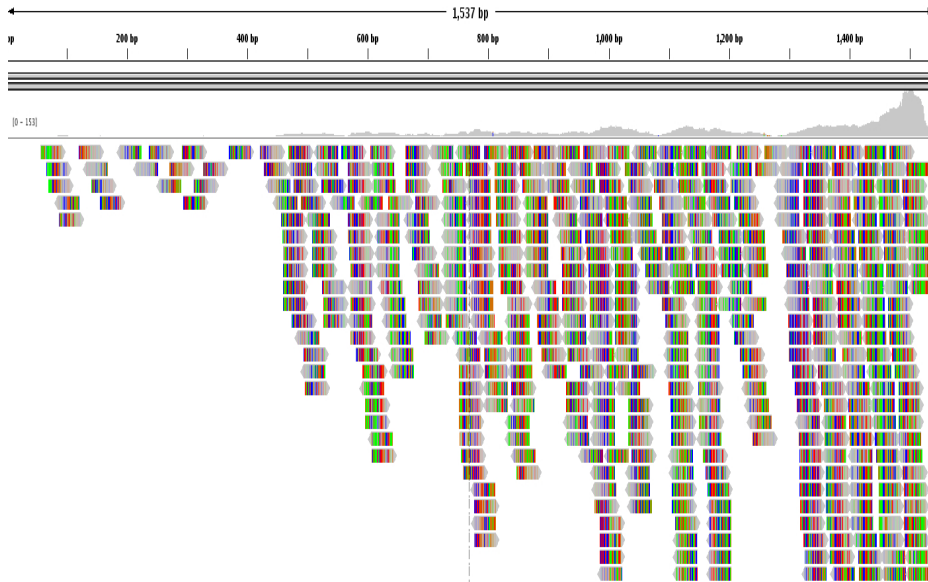
REFERENCE



Alignment: Illustration (Integrative Genomic Viewer (IGV))



Alignment: Illustration



The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.

The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.
- Multi-read: a read that can be aligned to multiple isoforms.

The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.
- Multi-read: a read that can be aligned to multiple isoforms.
 - ▶ Just by chance, repetitive sequences shared by multiple genes.

The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.
- Multi-read: a read that can be aligned to multiple isoforms.
 - ▶ Just by chance, repetitive sequences shared by multiple genes.
 - ▶ Alternative isoforms:



The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.
- Multi-read: a read that can be aligned to multiple isoforms.
 - ▶ Just by chance, repetitive sequences shared by multiple genes.
 - ▶ Alternative isoforms:



The Statistical Problem

- Number of reads aligned to an isoform provide a measure of abundance.
- Multi-read: a read that can be aligned to multiple isoforms.
 - ▶ Just by chance, repetitive sequences shared by multiple genes.
 - ▶ Alternative isoforms:



- Goal: A statistical model to resolve the ambiguity of multi-read.

Statistical Model

- Generative mechanism: T isoforms with proportion $\{\pi_k\}_{k=1}^T$ and length $\{l_k\}_{k=1}^T$

$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim P(r | t_i = k)$$

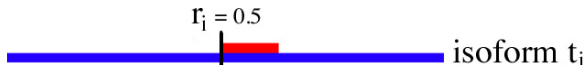


Statistical Model

- Generative mechanism: T isoforms with proportion $\{\pi_k\}_{k=1}^T$ and length $\{l_k\}_{k=1}^T$

$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim P(r | t_i = k)$$



- t_i is unobserved (due to multi-read), marginalize to get $p(r_i)$:

$$P(r = r_i) = \sum_{k=1}^T P(r = r_{ik} | t_i = k) p(t_i = k) = \sum_{k=1}^T P(r = r_{ik} | t_i = k) \pi_k$$

r_{ik} is the mapped location of read i on isoform k .

Statistical Model

- Generative mechanism: T isoforms with proportion $\{\pi_k\}_{k=1}^T$ and length $\{l_k\}_{k=1}^T$

$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim P(r | t_i = k)$$



- t_i is unobserved (due to multi-read), marginalize to get $p(r_i)$:

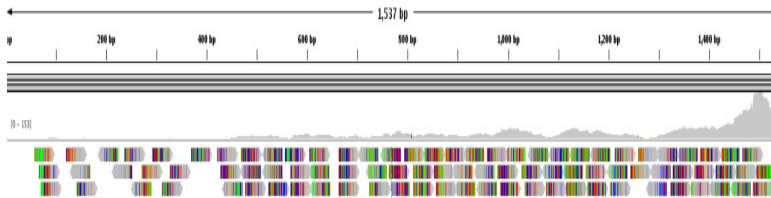
$$P(r = r_i) = \sum_{k=1}^T P(r = r_{ik} | t_i = k) p(t_i = k) = \sum_{k=1}^T P(r = r_{ik} | t_i = k) \pi_k$$

r_{ik} is the mapped location of read i on isoform k .

- $P(r = r_{ik} | t_i = k) = 0$ if read i cannot be mapped to isoform k .

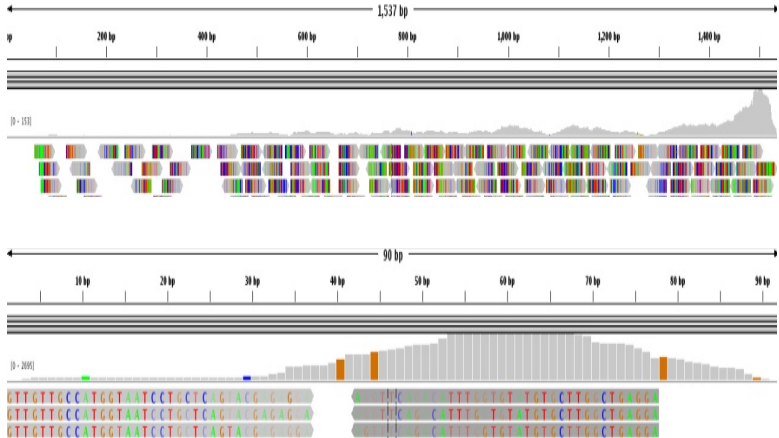
$$P(r_{ik}|t_i = k)$$

- Fragmentation varies as a function of isoform length:



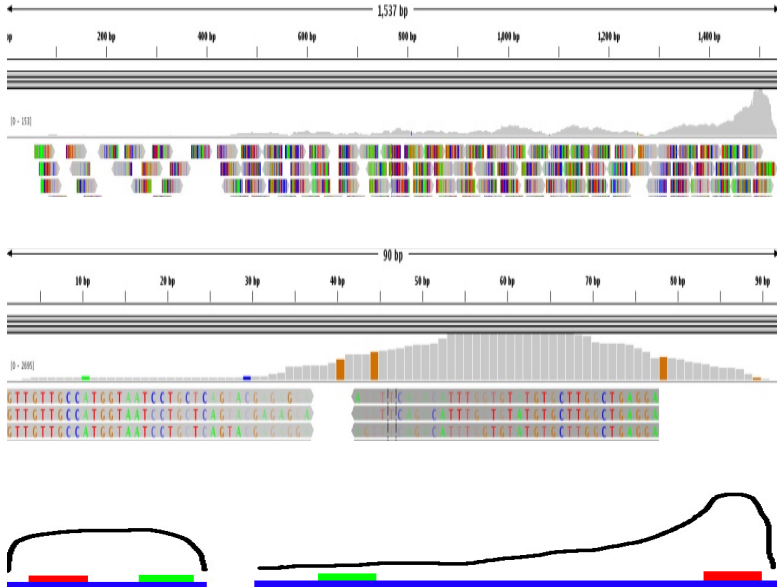
$$P(r_{ik}|t_i = k)$$

- Fragmentation varies as a function of isoform length:



$$P(r_{ik}|t_i = k)$$

- Fragmentation varies as a function of isoform length:



Beta Regression (Ferrari & Cribari-Neto (2004))

- Assume the relative read location r_i follows a beta distribution.

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

Beta Regression (Ferrari & Cribari-Neto (2004))

- Assume the relative read location r_i follows a beta distribution.

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

- w.r.t. usual parametrization $\text{beta}(a_k, b_k)$,
 $a_k = \mu_k \phi_k, b_k = (1 - \mu_k) \phi_k$

Beta Regression (Ferrari & Cribari-Neto (2004))

- Assume the relative read location r_i follows a beta distribution.

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

- w.r.t. usual parametrization $\text{beta}(a_k, b_k)$,
 $a_k = \mu_k \phi_k, b_k = (1 - \mu_k) \phi_k$
- $E(r_i | t_i = k) = \mu_k, \text{var}(r_i | t_i = k) = \frac{\mu_k(1-\mu_k)}{1+\phi_k}$

Beta Regression (Ferrari & Cribari-Neto (2004))

- Assume the relative read location r_i follows a beta distribution.

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

- w.r.t. usual parametrization $\text{beta}(a_k, b_k)$,
 $a_k = \mu_k \phi_k, b_k = (1 - \mu_k) \phi_k$
- $E(r_i | t_i = k) = \mu_k, \text{var}(r_i | t_i = k) = \frac{\mu_k(1-\mu_k)}{1+\phi_k}$
- Link functions:

$$\mu_k = \text{logit}^{-1}(\beta_0 + \beta_1 l_k)$$

$$\phi_k = \exp(\theta_0 + \theta_1 l_k)$$

Beta Regression (Ferrari & Cribari-Neto (2004))

- Assume the relative read location r_i follows a beta distribution.

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

- w.r.t. usual parametrization $\text{beta}(a_k, b_k)$,
 $a_k = \mu_k \phi_k$, $b_k = (1 - \mu_k) \phi_k$
- $E(r_i | t_i = k) = \mu_k$, $\text{var}(r_i | t_i = k) = \frac{\mu_k(1-\mu_k)}{1+\phi_k}$
- Link functions:

$$\mu_k = \text{logit}^{-1}(\beta_0 + \beta_1 l_k)$$

$$\phi_k = \exp(\theta_0 + \theta_1 l_k)$$

- log-likelihood:

$$\begin{aligned} \log P(r_i | t_i = k) &= \log \Gamma(\phi_k) - \log \Gamma(\mu_k \phi_k) - \log \Gamma((1 - \mu_k) \phi_k) \\ &+ (\mu_k \phi_k - 1) \log(r_i) + \{(1 - \mu_k) \phi_k - 1\} \log(1 - r_i) \end{aligned}$$

EM-Algorithm

- The expected complete data log-likelihood is:

$$\sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log P(r_{ik} | t_i = k) + \sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log \pi_k$$

Where $\tau_{ik} = P(t_i = k | r_{ik})$

EM-Algorithm

- The expected complete data log-likelihood is:

$$\sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log P(r_{ik} | t_i = k) + \sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log \pi_k$$

Where $\tau_{ik} = P(t_i = k | r_{ik})$

- E-Step:

$$\tau_{ik} \doteq P(t_i = k | r_{ik}) = \frac{P(r_{ik} | t_i = k) \pi_k}{\sum_{k'=1}^T P(r_{ik'} | t_i = k') \pi_{k'}}$$

EM-Algorithm

- The expected complete data log-likelihood is:

$$\sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log P(r_{ik} | t_i = k) + \sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log \pi_k$$

Where $\tau_{ik} = P(t_i = k | r_{ik})$

- E-Step:

$$\tau_{ik} \doteq P(t_i = k | r_{ik}) = \frac{P(r_{ik} | t_i = k) \pi_k}{\sum_{k'=1}^T P(r_{ik'} | t_i = k') \pi_{k'}}$$

- M-Step

$$\pi_k^{new} = \frac{\sum_{i=1}^N \tau_{ik}}{N}$$

EM-Algorithm

- The expected complete data log-likelihood is:

$$\sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log P(r_{ik} | t_i = k) + \sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log \pi_k$$

Where $\tau_{ik} = P(t_i = k | r_{ik})$

- E-Step:

$$\tau_{ik} \doteq P(t_i = k | r_{ik}) = \frac{P(r_{ik} | t_i = k) \pi_k}{\sum_{k'=1}^T P(r_{ik'} | t_i = k') \pi_{k'}}$$

- M-Step

$$\pi_k^{new} = \frac{\sum_{i=1}^N \tau_{ik}}{N}$$

- For beta regression, solve the weighted beta regression problem:

$$\operatorname{argmax}_{\beta_{0,1}, \theta_{0,1}} \sum_{i=1}^N \sum_{k=1}^T \tau_{ik} \log P(r_i | t_i = k)$$

Simulation

- 500,000 reads, 36 bp.

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- $\phi(\text{length}) = 4\Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- $\phi(\text{length}) = 4\Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- Most isoforms have small sample proportions ($\pi_k \sim 10^{-5}$), some larger (~ 0.001).

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- $\phi(\text{length}) = 4\Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- Most isoforms have small sample proportions ($\pi_k \sim 10^{-5}$), some larger (~ 0.001).
- Recall:

$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- $\phi(\text{length}) = 4\Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- Most isoforms have small sample proportions ($\pi_k \sim 10^{-5}$), some larger (~ 0.001).
- Recall:

$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

- Beta regression: natural cubic splines expansion on $\log(\text{length})$, 4 equally spaced knots.

Simulation

- 500,000 reads, 36 bp.
- 3437 isoforms from the 150 genes with the highest number of isoforms (28-68) from the Ensembl database.
- $\mu(\text{length}) = \Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- $\phi(\text{length}) = 4\Phi^{-1}\left(\frac{\text{length}-500}{5000}\right)$
- Most isoforms have small sample proportions ($\pi_k \sim 10^{-5}$), some larger (~ 0.001).
- Recall:

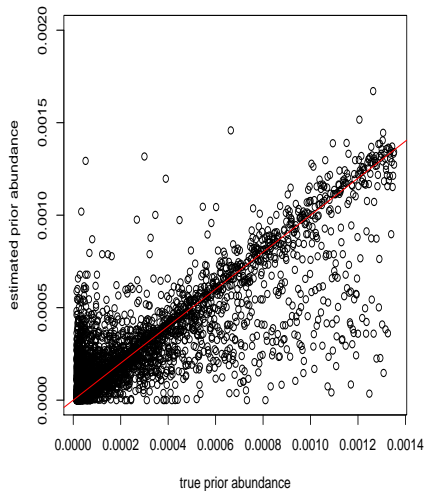
$$t_i \sim \text{multinomial}(\pi_1, \dots, \pi_T)$$

$$r_i | t_i = k \sim \text{beta}(\mu_k, \phi_k)$$

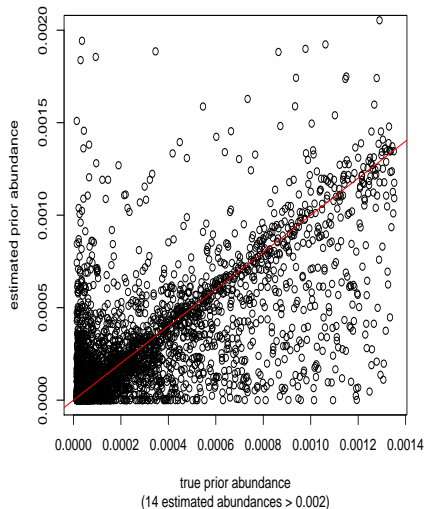
- Beta regression: natural cubic splines expansion on $\log(\text{length})$, 4 equally spaced knots.
- Bowtie (Langmead (2009)) for read alignment. Reads with > 200 mappable locations are discarded. Obtain $\sim 15,000,000$ alignments.

Simulation Results

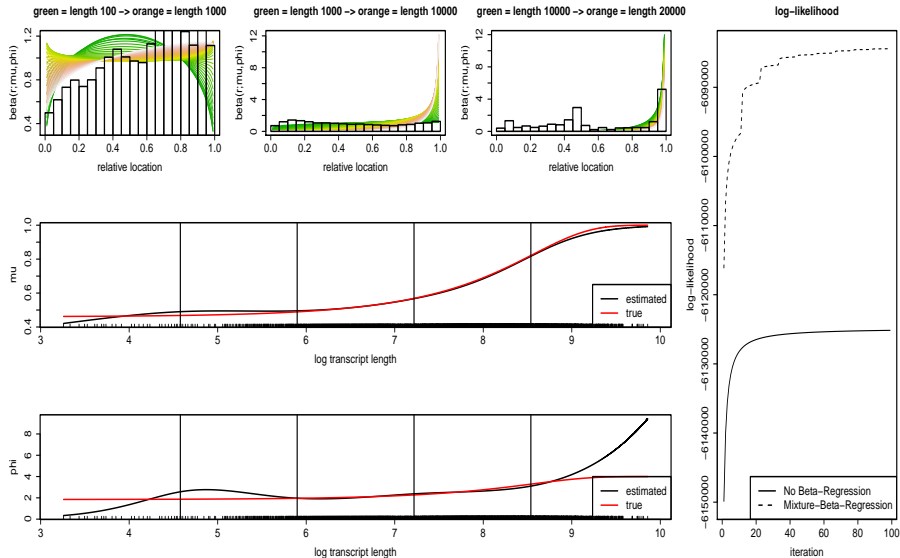
With Fragmentation Estimation



Uniform Fragmentation



Fitted Model



Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.

Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.
- Computational aspects:

Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.
- Computational aspects:
 - ▶ E-step and updates for π_k : sparse matrix manipulation.

Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.
- Computational aspects:
 - ▶ E-step and updates for π_k : sparse matrix manipulation.
 - ▶ Beta regression: slow.

Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.
- Computational aspects:
 - ▶ E-step and updates for π_k : sparse matrix manipulation.
 - ▶ Beta regression: slow.
 - ▶ Trick: ECM algorithm.

Final Thoughts

- Fragmentation estimations improves isoform abundances estimation accuracies.
- Computational aspects:
 - ▶ E-step and updates for π_k : sparse matrix manipulation.
 - ▶ Beta regression: slow.
 - ▶ Trick: ECM algorithm.
- Future work: read errors, GC-content bias.

Reference

- Ferrari, S, & Cribari-Neto, F. (2004) Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):199-815

Reference

- Ferrari, S, & Cribari-Neto, F. (2004) Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, 31(7):199-815
- Langmead B, Trapnell C, Pop M, Salzberg SL.(2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Reference

- Ferrari, S, & Cribari-Neto, F. (2004) Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, 31(7):199-815
- Langmead B, Trapnell C, Pop M, Salzberg SL.(2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
- IGV: Integrative Genomics Viewer.
<http://www.broadinstitute.org/igv/>

Reference

- Ferrari, S, & Cribari-Neto, F. (2004) Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, 31(7):199-815
- Langmead B, Trapnell C, Pop M, Salzberg SL.(2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
- IGV: Integrative Genomics Viewer.
<http://www.broadinstitute.org/igv/>
- Li, B. et. al. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, 26(4):493-500.