# Optimising a Genomic Annotation for the Analysis of RNA-Seq Data.

Ellis Patrick, Michael Buckley and Jean Yee Hwa Yang.

School of Mathematics and Statistics, University of Sydney; CSIRO Mathematics, Informatics and Statistics.

## Introduction

While ideally we would like to work at the transcriptome level when analysing RNA-Seq data, there may still be a need or want to work on a genomic level; we may lack confidence in the quality of the annotation or that we have sequenced to a sufficient depth to make inferences about the abundance of different isoforms within a gene. When performing an expression analysis at a genomic level, a change in the isoforms present within a gene can introduce a length bias when comparing the expression of a gene between samples. The two key aims of this project are:

- Detect a change in isoforms.
- Estimate constitutive exons.

## Assumptions

Assume reads are drawn independently from some underlying distribution. Thus if sequencing depth, gene expression and the proportions of transcripts within a gene are held constant we can expect the exon counts for a particular gene to be uncorrelated. We plan on manipulating this assumption to develop a metric for detecting a change in alternate splicing and will attempt to use this to estimate the constitutive exons.

## Detect a change in isoforms

### Estimating sequencing depth

An intuitive method for estimating sequencing depth, $T_j$, for sample $j$ would simply be to take the total counts for a sample. For robustness we instead use the trimmed means of M-values method (TMM) proposed in [3].

### Estimating $\mu_{ij}$

If we restrict our attention to a particular gene, then let:

- $X_{ij}$ be the counts for the $i^{th}$ exon in the $j^{th}$ sample.
- $n_s$ and $n_e$ be the number of samples and exons.
- $\gamma(j)$ be the group that sample $j$ belongs to.

We can estimate $\mu_{ij}$, the value we expect to observe at $X_{ij}$ as

$$\begin{aligned}\hat{\mu}_{ij} &= (\text{proportion of total counts in group } \gamma(j)) \\ &\quad \times (\text{proportion of group } \gamma(j) \text{ counts in sample } j) \\ &\quad \times (\text{total counts for exon } i) \\ &= \frac{\sum_{k=1}^{n_e}\sum_{m\in\gamma(j)} X_{km}}{\sum_{k=1}^{n_e}\sum_{m=1}^{n_s} X_{km}} \frac{T_j}{\sum_{k\in\gamma(j)} T_k} \sum_{k=1}^{n_s} X_{i,k}\end{aligned}$$

### Estimating the covariance matrix $\Sigma$

Assuming the count data follows a Poisson distribution we could standardize our data

$$Z_{ij} = \frac{X_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

and hence calculate a covariance matrix as

$$\hat{\Sigma} = ZZ^T/(n_s - 1).$$

### A metric for estimating a change in isoforms

$$\tau = \frac{1}{n_e^2}[\sum_i^{n_e}\sum_j^{n_e} \hat{\Sigma}_{ij}^2] - \frac{n_e}{n_e^2}$$

where large values of $\tau$ should correspond to a change in the isoforms present between the two conditions.

## Estimate constitutive exons

In order to focus on the overall expression of a gene, rather than isoform-specific expression, Bullard et al (2010) [1] define a Union-Intersection (UI) annotation for a gene which is simply an attempt to estimate the constitutive exons within a gene. This UI definition, derived entirely from the chosen annotation, is quite restrictive and lead to the exclusion of large amounts of data. We propose a method, inspired by the work of [4] on exon arrays, which attempts to estimate the constitutive exons based on experimental data.

### Method

Using a suitable annotation, then for a given gene we can summarized how many reads lie within each exon of that gene for each sample. We define our method for identifying constitutive exons as follows:

1) Apply average-linkage hierarchical clustering of the exons across all samples using 1- $\Sigma$ as a distance metric, where $\Sigma$ is the covariance matrix described earlier.
2) Cut the clustering dendrogram at some predetermined height.
3) Evaluate all subclusters of the tree using some scoreing metric.
4) The union of all the exons in highest scoring subcluster then becomes then new annotation for the specified gene.

For a scoring metric we take the subcluster which has the highest average coverage. This method does not guarantee to estimate the constitutive exons but should at the least identify a dominant signal.

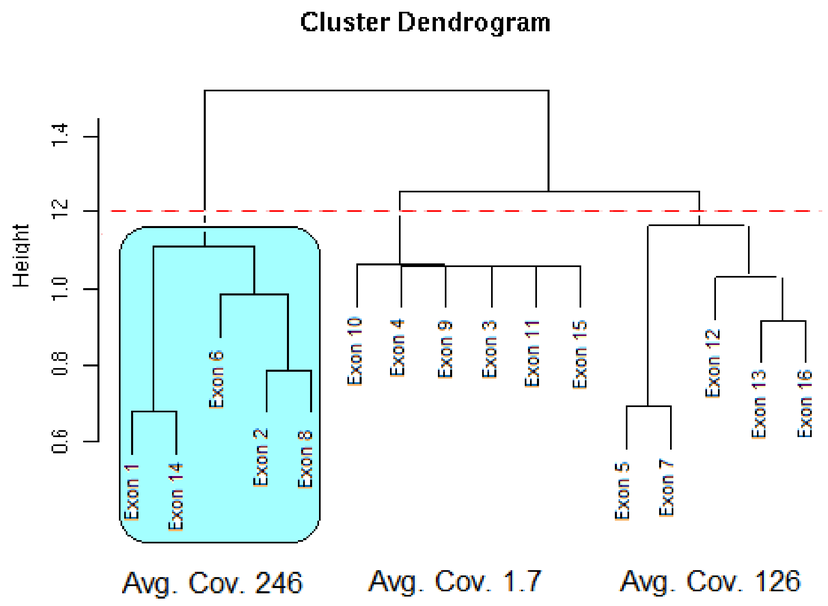


Figure 1: Plot of exons selected by clustering method for a particular gene.

## Public Data

The data consists of two mRNA-Seq datasets from the MicroArray Quality Control Project [2] where 35 base-pair-long reads were obtained using Illuminas Genome Analyzer II high-throughput sequencing system. It compares Ambions human brain reference RNA (Brain) to Stratagenes human universal reference RNA (UHR). Both Brain and UHR were assayed, each using seven lanes distributed across two flow-cells. Accompanying this data set is qRT-PCR data from MAQC-1 which consists of four observations for both Brain and UHR over 1021 genes.

## Experimental Results

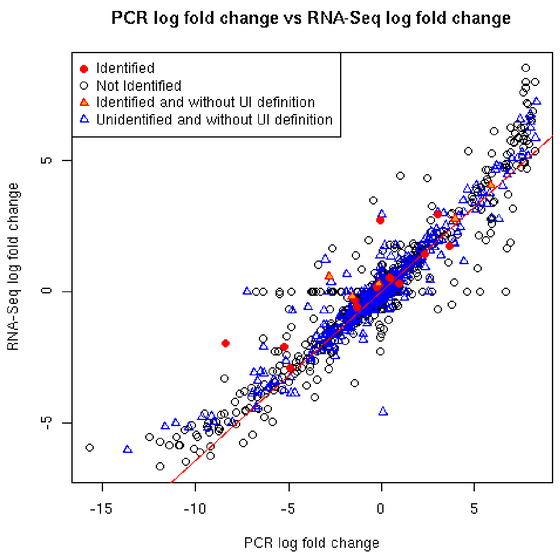


Figure 2: For the RNA-Seq data we use the union of all exons within a gene to summarize our counts. The black circles are those genes for which the UI definition is non-empty. The blue triangles are the undetected genes of the 386 genes for which the UI definition is empty. The red dots and orange triangles are those genes that our method identified as having a change in isoforms and had a non-empty and empty UI definition.

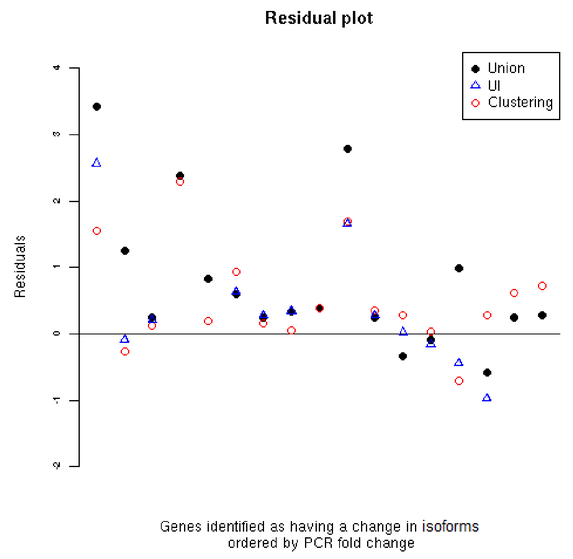


Figure 3: After fitting a straight line through the plot on the left, this figure plots the residuals for the genes identified as having a change in isoforms for 3 different annotations, union of all exons (black dots), UI definition (blue triangles) and our clustering method (red circles).

### Evaluation

- A high proportion of the genes identified as having a change in isoforms in figure 2 appear to be outliers.
- On manual inspection all these genes appear to have been identified correctly however we have no measure of false negatives.
- The UI definition generally seems to make the residuals smaller in figure 3.
- Our clustering method seems to behave consistently with the UI definition but is defined for all of the genes.

## Discussion

When working at a genomic level a change in the proportions of isoforms present within a gene could cause problems with an expression analysis. We have outlined two simple tools to identify the effects of alternate splicing and estimation constitutive exons which should help facilitate the study of novel gene models.

## References

[1] Bullard, J., Purdom, E., Hansen, K. and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11** 94+.

[2] Consortium, M. A. Q. C. (2006). The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24** 1151–1161.

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

[4] Xing, Y., Kapur, K. and Wong, W. H. (2006). Probe selection and expression index computation of affymetrix exon arrays. *PLoS One*, **1** e88.

Supported in part by ft0991918 (YH), the APA (EP) and CSIRO.

Email: ellis.patrick@sydney.edu.au