mbi Mathematical Biosciences Institute at The Ohio State University

The Ohio State University Center for Pharmacogenomics

The Ohio State University College of Public Health
Ohio's first and only accredited college of public health

Signal Decomposition of Allelic Expression RNA-Seq Data Using Skellam Mixture Model

Rong Lu, Ryan M. Smith, Michał Seweryn, Audrey C. Papp, Amy Webb, Wolfgang Sadee, & Grzegorz A. Rempala

INTRODUCTION

Differences in RNA expression across two alleles in the same individual (allelic expression imbalance; AEI) is a powerful phenotype for identifying functional regulatory SNPs, allele-specific epigenetic programming, and loss-of-heterozygosity in cancer, among other things. The advent of high-throughput RNA-Seq allows us to survey the entire transcriptome for AEI. However, existing studies [1-3] highlight significant challenges that obscure reliable AEI detection, especially for modest differences in expression across the two alleles (1.5 to 2-fold AEI). Most notably, detecting AEI with standard methods requires high read depth, which cannot be easily obtained by whole transcriptome RNA-Seq for the majority of genes. Therefore, we need to maximize our utilization of the information generated by RNA-Seq to more reliably detect modest AEI. Here, we test whether a Skellam mixture model is suitable for detecting robust AEI.

AIM

1. Estimate probabilities of SNP-wise AEI assuming finite Skellam mixture distribution
2. Fish/classify SNPs with strong evidence of AEI
3. Identify variables that characterize the subsets of SNPs with strong evidence of AEI

METHODS

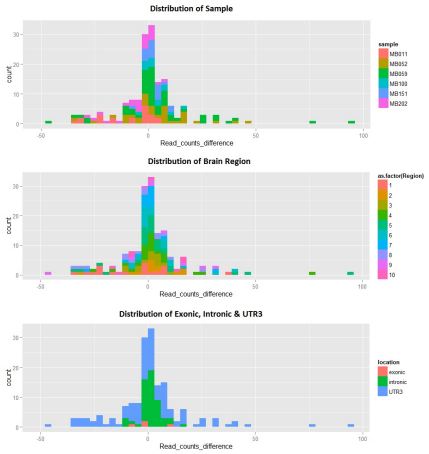
Dataset

- Human brain tissue from 10 different regions (5 cortical, 5 subcortical/brainstem) in 10 subject (100 total samples)
- Whole-transcriptome RNA-Seq using SOLiD technology
- Aligned using IUPAC ambiguity codes to alleviate inherent allelic bias
- Tested here on a single gene (*SLC1A3*)
- Only single nucleotide polymorphisms (SNPs) confirmed heterozygous from high-density array included for analysis
- At least 3 reads/allele required for analysis
- Includes both intronic and exonic SNPs

Statistical Methods

In general, the "reference" reads ("ref") and the "variant" reads ("var") are highly correlated due to common regulatory effect (for a given gene). We focus on modeling the difference of read counts not only to avoid dealing with the high correlation between "ref" and "var" reads but also to reduce possible correlation across SNPs and confounding effects that impact on "ref" and "var" in the same way.

Figure 1: Data Visualization



•Read_counts_difference = "ref" – "var". Cor("ref", "var")=0.893. Total number of SNPs = 163.
•Number of SNPs by sample: 21, 33, 53, 13, 19, and 24 correspond to sample "MB011" through "MB202".
•Number of SNPs grouped by brain region: 18, 12, 14, 23, 28, 21, 16, 14, 9, and 8 correspond to brain region 1 through 10.
•Number of SNPs grouped by location: 4, 53, and 106 correspond to exonic, intronic, and UTR3.

Let Y_1 and Y_2 be the "ref" and "var" read counts respectively. We assume $Y = Y_1 - Y_2$ follow a Skellam mixture distribution with unknown fixed number of mixture components (K). That is,

$$Y_1 = X_1 + X$$

$$Y_2 = X_2 + X$$

where $$Y = Y_1 - Y_2$$

$$f_{Y_1,Y_2}(y_1,y_2) = \sum_{x=0}^{\min(y_1,y_2)} \left[\sum_{i=1}^{K} \pi_i \,\text{Poisson}(y_1 - x|\lambda_{i1})\text{Poisson}(y_2 - x|\lambda_{i2}) f_X(x|\lambda_{iX})\right]$$

$$p_Y(y|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_i \times \text{Skellam}(y|\lambda_{i1},\lambda_{i2})$$

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$$

$$\boldsymbol{\theta} = \left(\begin{pmatrix}\lambda_{11}\\ \lambda_{12}\end{pmatrix}, \begin{pmatrix}\lambda_{21}\\ \lambda_{22}\end{pmatrix}, \ldots, \begin{pmatrix}\lambda_{K1}\\ \lambda_{K2}\end{pmatrix}\right)$$

where π_i's are the proportions of mixture components. $\text{Poisson}(\lambda_{i1}), \text{Poisson}(\lambda_{i2}),$ and X are independent across all mixture components $i = 1, 2, \ldots, K$. It's worth noting that X can be any distribution or any mixture of distributions as long as it is independent with all other distributions.

We use BIC as the criteria of choosing the optimal number of mixture components:

$$\text{BIC} = -2 \times \log(\text{likeli}) + \log(n)(3K - 1)$$

RESULTS

Parameter Estimates of Skellam Mixture:

	Component 1	Component 2	Component 3	Component 4
pi	0.0124 (0.0112, 0.0124)	0.8073 (0.79, 0.81)	0.1101 (0.109, 0.122)	0.0702 (0.064, 0.073)
lambda1	132.0283 (129.87, 139.79)	22.7526 (20.64, 24.03)	13.0100 (12.68, 22.08)	53.2318 (48.32, 58.44)
lambda2	48.3224 (46.092, 56.336)	22.0783 (19.93, 23.10)	41.5387 (40.77, 50.26)	21.9563 (16.49, 25.84)

# of Comp.	loglik	BIC
3	-663.185	1367.12
4	-652.226	1360.482
5	-652.851	1377.015
6	-653.236	1393.067

All fitting are done using EM algorithm with 10,000 sets of initials. For each set of initials, local maximum of log-likelihood is found when the relative log-likelihood difference between successive iterations is less than 10^(-3).

SNPs Classification:

Figure 2. Classification of SNPs Using Skellam Mixture

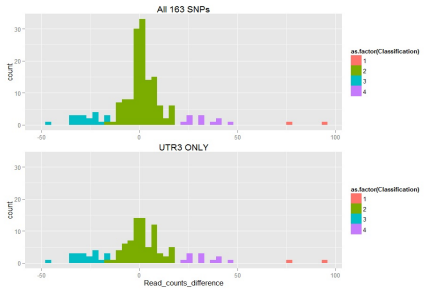
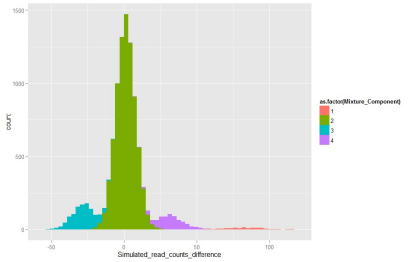


Figure 3. Simulation Sample From Fitted Skellam Model



Sample	Classification (UTR3 ONLY)			
	C1	C2	C3	C4
MB011	0	3	7	0
MB052	0	22	3	4
MB059	2	20	5	7
MB100	0	10	0	0
MB151	0	0	0	0
MB202	0	19	4	0

CONCLUSIONS & FUTURE DIRECTIONS

By applying a Skellam mixture model to our allelic RNA expression data, we are capable of classifying individual SNPs into distinct components. By testing how well SNPs grouped within individual samples fit the estimated proportions of the mixture model, we find it possible to identify samples that exhibit strong AEI, such as MB011. This analysis did not adjust for sequencing depth before fitting the mixture model, which is necessary when read depths across samples are highly variable and is part of our future workflow. This novel approach provides a statistical framework for testing for AEI in RNA-Seq data. Our future studies will test the feasibility of constructing Skellam mixture models or mixture of negative binomial differences from larger datasets that include multiple genes, with the goal of applying this methodology genome-wide.

BIBLIOGRAPHY

1 Nothnagel *et al.* (2011). Statistical Inference of Allelic Imbalance from Transcriptome Data. *Hum Mutat*, 32:98-106.

2 Heap *et al.* (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*, 19:122-124.

3 Smith *et al.* (2013). Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, 14:571.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences (U01GM092655), the US National Science Foundation (DMS-1318886), and the US National Cancer Institute (R01-CA152158).