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. 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.

\[
p_{\text{ref}}(y|\theta) = \sum_{i=1}^{\text{Comp.}4} \pi_{\text{Comp.}i} \times \text{Skellam}(y|x_i, \lambda_i) = \sum_{i=1}^{\text{Comp.}4} \pi_{\text{Comp.}i} \times \text{Skellam}(y|x_i, \lambda_i)
\]

where \(y\) is the read count, \(\theta\) the unknown parameters, \(\pi\) the mixture proportions, \(x\) the mean of the mixture components, and \(\lambda\) the precision of the mixture components. The Skellam distribution is a discrete distribution that models the difference of two independent Poisson variables.

RESULTS

Parameter Estimates of Skellam Mixture:

\[
\begin{array}{cccc}
\text{Component} & \text{Mean} & \text{Precision} & \text{Proportion} \\
1 & 103.033 & 22.776 & 0.1101 \\
2 & 132.028 & 48.3224 & 0.1101 \\
3 & 22.7526 & 41.5387 & 0.1101 \\
4 & 53.2318 & 0.0702 & 0.1101 \\
\end{array}
\]

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 genomes, with the goal of applying this methodology genome-wide.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences (U21GM105365), the US National Science Foundation (DMS-1719888), and the US National Cancer Institute (R01-CA132132).