miRNA Expression:
Normalization Algorithms and Amplification Biases

John Steill

March 22, 2018

1 Introduction

Small noncoding RNAs known as microRNA (miRNA) play a crucial role in gene expression regulation. Each miRNA can target many genes, and many genes can be effected by a variety of miRNA[6]. However, there is still no consensus on how expression data should be pre-processed before downstream analysis.[1] I will survey and evaluate existing normalization methods and attempt to characterize and correct for amplification biases. If I am successful, we will be better able to contrast the regulatory effects of miRNA profiles from different biological conditions.

2 Resources

I have access to raw fast files for hundreds samples from dozens of sequencing submissions. A data set I am especially interested in is the same sample in nine different combinations of starting amount and amplification cycles. In addition, I am able to gather GEO data if needed. I have access to both human and mouse miRNA libraries. Finally, I have access to a suite of multi-core clusters.

3 Methods

I intend to investigate at least simple cpm normalization, upper percentile normalization[2], and Trimmed Mean of M-values [3] normalization. I hope to model amplification bias as a function of miRNA library, starting qty, and number of amplification cycles using Caffe CNNs[5]. For DE, I will use EBSeq.[4]

4 Anticipated Results

I expect non-trivial normalization algorithms to be sample-set dependent. Thus the input will be a table of raw counts, and the output will be a table of normalized values. I will evaluate the results with DE analysis: I hope to minimize false positives by examining samples which are biological replicates but not technical replicates.

An ideal result would be a satisfactory normalization method as well as an explanatory CNN. For example, I hypothesize that hairpin structures in the miRNA, caused be nearby reverse palindromic sequences, might inhibit amplification, so these species will be under-represented with low starting quantities and more amplification cycles. It would be ideal to find a convolutional layer that detected this condition. In addition, I will be looking for more integral patterns, such as CG vs AT content.

References

[1] Tam, Shirley, Ming-Sound Tsao, and John D. McPherson. "Optimization of miRNA-seq data preprocessing." Briefings in bioinformatics 16.6 (2015): 950-963.

[2] Bullard, James H., et al. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." BMC bioinformatics 11.1 (2010): 94.

[3] Robinson, Mark D., and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data." Genome biology 11.3 (2010): R25.

[4] Leng, Ning, et al. "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." Bioinformatics 29.8 (2013): 1035-1043.

[5] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.

[6] Dong, Haifeng, et al. "MicroRNA: function, detection, and bioanalysis." Chemical reviews 113.8 (2013): 6207-6233.