

miRNA Expression
Normalization Algorithms and Amplification Biases
Progress Report
John Steill

1 Proposal updates

There are no significant scope updates. I am committed to recommending a normalization algorithm optimized for miRNA Differential Expression Analysis. I intend to quantify the effects of amplification biases. Explanatory efforts with CNNs remain a stretch goal.

2 Progress

- I have obtained data from sequencing the same biological sample with 18 technical replicates, 2 replicates each at 3 different starting quantities (3ng, 6.25ng, and 25ng) AND 3 different PCR settings (10, 20, and 20 cycles.)
- I have decided upon a 100 ng sample to be used as a gold standard.
- I have programmed a custom miRNA alignment algorithm, in which families of miRNA transcripts which share the first 15 bases are treated as a single library. I have obtained expected counts.
- I have programmed two normalization routines, CPM and Upper Percentile, and have demonstrated that they produce markedly different profiles. (Trimmed Mean of M-Values remains to be completed.)
- I have observed strong evidence of systemic bias, in which a library's relative expression rises nearly monotonically with decreasing initial quantity and increasing pcr cycles.

3 Challenges

- It will not be easy to decide upon a metric to evaluate normalization algorithms. It will likely turn out that the optimal choice for analyzing top expressors and low expressors will be different. I will mitigate this by both continuing my literature review and understanding choices other authors have made. Also, I will consult with lab mates to define as tightly as possible their experimental aim and optimize for that.
- The examples I've seen for training neural nets has thus far been classification problems. The function here I am trying to learn is $bias = f(library, amt, N_{cycles})$, which is real valued. The signal maybe relatively weak, and designing the layers to capture hairpin-inducing sequences will be complicated. I will mitigate this risk by treating this part of

the project as exploratory, and possibly lower my expectations of being able to reliably "de-bias" normalized expression values values.