ABSTRACT:

With the rapid development of second-generation sequencing technologies, RNA-Seq has become a popular tool for transcriptome analysis. It offers the chance to detect novel transcripts by obtaining tens of millions of short reads. After mapped to the genome and/or to the reference transcripts, RNA-Seq data can be summarized by a tremendous number of short-read counts. The huge number of short-read counts enables researchers to make transcript quantification in ultra-high resolution. Recent work found that short-read counts have significant sequence bias, which makes simple transcript quantification methods questionable. Thus, more elaborate statistical models that can effectively remove the sequence bias of the short-read counts are highly desirable to make transcript quantification more accurate. In this talk, I will present some statistical analysis for bias correction in RNA-Seq short-read counts. Since the sample size is over tens of millions, routine statistical computing is infeasible. Our statistical computing is conducted using a subsampling method called leveraging. I will present some statistical properties of the leveraging algorithm. Real RNA-Seq examples will also be presented to demonstrate the empirical performance of our method.