Can we use a rarefaction curves to assess how much coverage is needed for RNA-seq?:

In short, rarefaction curves can be used to determine coverage for RNA-seq, but the degree of confidence in the value depends on the application.

Rarefaction curves plot the accumulation of unique transcripts by the abundance of transcripts sequenced/assembled. When the rarefaction curve reaches an asymptote, an increase in transcript abundance (or sampling intensity) does not result in additional expressed transcripts being detected; thus 100% sampling coverage has been met at this point.

When applied to samples with a fully-assembled reference genome, the asymptote should match with the number of transcripts known to occur in the genome, thus sample coverage can be considered quantitative. However, it is generally impossible to know how many transcripts are produced by an organism without a fully genome sequence. As such, the utility of rarefaction curves is purely qualitative simply indicating when samples are close to saturation in de novo assemblies. Due to natural variability in RNA, library preparation, sequencing error, or assembly error, the detection of false positive transcripts may occur and are undifferentiated from expressed transcripts. As new transcripts are detected with increased sampling coverage, the probability of detecting new transcripts decreases while the probability of false positive occurrence remains constant. At some point, the probability of detecting a new transcript will be lower than the probability of detecting a false positive which will continue to inflate rarefaction curves (Fig 1). In order to account for false positives, normalization should be applied either by increasing sensitivity to new transcript detection (i.e. omitting genes with fewer than 5 reads mapped) or by determining the rate of false positive occurrence in known genomes and applying this value to reduce expression of common genes and enhance detection of rare transcripts.

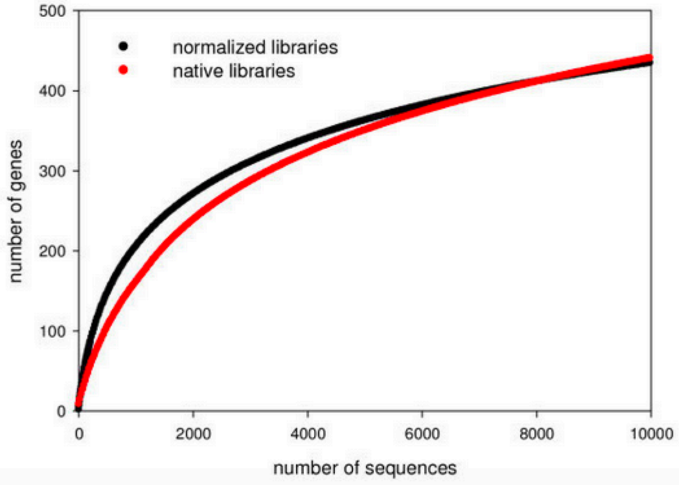

Fig 1. In de novo genome assembly,

Rarefaction curves illustrating identification of new transcripts or genes as a function of sampling intensity (number of sequences). Analysis comparing normalized library to naïve library show increased detection of rare genes such that sampling coverage is met sooner.

From Hale et al. 2009. Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): The relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10(1):203.