ModENCODE RNA-Seq best practices (DRAFT) adapted from Myers/Wold/Gingeras 2009

Send any comments on this document to Brian Oliver

1. Obtain greater than 1X RNA-Seq coverage of known "gold standard" transcripts that are present in a range of abundances. For cell lines this should include transcripts present at less than 1 copy per cell. This standard will be difficult to achieve in complex tissues where rare transcripts are expressed in rare cell types. Cell level purification, fractionation, normalization or other methods for getting rare transcripts from organisms will be required.

2. Primary determination will be performed in biological duplicate, unless there is a compelling reason indicating that this is impractical or wasteful (e.g. overlapping timepoints, Solexa & 454 data). Technical replicates of the same library are not required. Biological duplicate must show > 0.9 correlation for transcripts/ features greater than 1 RPKM.

3. A ladder of RNA spike-ins should be included to calibrate quatification, sensitivity, coverage and linearity. The pools of ERCC mRNAs obtained from NIST are available and should be used in future experiments.

4. A null model for non-transcribed regions is highly problematic as we do not know enough about the genome and how much of it is really transcribed. While we do not have a model for biological transcriptional noise, we should determine the frequency of sequencing and mis-mapping in simulations to determine how much density and junction noise is expected for technical reasons.

5. Read density provides strong support for transcription of a given segment of the genome, but producing a full annotation of all spliced transcripts is complicated by read length and strandedness. Longer and longer RNA-seq read lengths are possible and maximum length paired-end data is highly desirable for determining junctions linking one or more exons. Overlapping genes should be distinguished by data from stranded libraries. Maximal discovery and annotation may require a mix of RNA-Seq strategies.

6. RNA-Seq data currently provides very high quality exon-pair or "genelet" data, but deduced full-length transcript models produced exclusively by RNA-Seq data are suspect. Exon-pair/genelet annotation should be reported. Full-length sequence should be pursued whenever feasible to provide connectivity.

7. These guidelines will continue to evolve as the technology improves.