Standards, Guidelines and Best Practices for RNA-Seq
V1.0 (June 2011)
The ENCODE Consortium

I. Introduction: Sequence based assays of transcriptomes (RNA-seq) are in wide use because of their favorable properties for quantification, transcript discovery and splice isoform identification, as well as adaptability for numerous more specialized measurements. RNA-Seq studies present some challenges that are shared with prior methods such as microarrays and SAGE tagging, and they also present new ones that are specific to high-throughput sequencing platforms and the data they produce.

RNA-Seq is not a mature technology. It is undergoing rapid evolution of biochemistry of sample preparation; of sequencing platforms; of computational pipelines; and of subsequent analysis methods that include statistical treatments and transcript model building. This document is part of an ongoing effort to provide the community with standards and guidelines that will be updated as RNA-Seq matures and to highlight unmet challenges. The intent is to revise this document annually to capture new advances and increasingly consolidate standards and best practices.

RNA-Seq experiments are diverse in their aims and design goals, currently including multiple types of RNA isolated from whole cells or from specific sub-cellular compartments or biochemical classes, such as total polyA+ RNA, polysomal RNA, nuclear ribosome-depleted RNA, various size fractions of RNA and a host of others. The goals of individual experiments range from major transcriptome "discovery" that seeks to define and quantify all RNA species in a starting RNA sample to experiments that simply need to detect significant changes in the more abundant RNA classes across many samples for the purpose of cell phenotyping. The guidelines and standards discussed here do not exhaustively cover the entire matrix of this experimental space, but instead emphasize best practices designed to support "reference quality" transcriptome measurements for major RNA sample types.

Different study aims and RNA types will therefore call for appropriate adjustments in standards developed for reference measurements. However, other parts of the standards recommended, such as providing proper meta-data to describe the sample and processing should be widely applicable.

II. Information to be supplied with each sample used for an RNA-seq experiment.

To be useful to the scientific community, RNA-seq data should be accompanied by information concerning the biological source of the RNA and protocols used to extract and prepare the RNAs.

1. For cell lines the following information should be recorded and provided:
    a) Cell line source and lot number.
    b) Growth time/passage number.
    c) Cell density.
    d) Cite protocol used to culture cell lines.
    e) Cite results of tissue culture contaminant (e.g.mycoplasma/ wolbacia) tests if conducted
    f) Confirmation of freezing cell aliquots of examined lines

2. For sub-cellular compartments, tissues, organs or whole organisms, the following should be recorded and provided:
    a) animal strain and genotype.
    b) individual genome sequence source data, where available
    c) Protocols tissue or cell type preparation, such cell sorting markers etc, embryo staging, etc. Estimates of purity from cellular enrichments should be provided.
    d) Amounts of starting material (tissue/organ weights, cell number from which sub-cellular compartments were isolated, etc).
    e) Estimate of enrichment or homogeneity of sample from other associated biological elements (e.g. degree of nuclear enrichment compared to associated cytosolic elements, percent homogeneity of CD8+ cells, fraction of animals of the stated stage).

3. Identification of the type of RNA targeted (size range, poly A+ or A-, 5' capped or uncapped, polysomal, etc).

4. Protocols used to isolate RNAs (size range, 5'/5'-3'tags, poly A+/A-). While each RNA type has unique issues associated with the intended purification, evidence of the enriched status of the targeted RNA type should be recorded. This could include length profiles of the isolated RNAs, the amount of ribosomal RNA present in poly A+ samples or evidence of 5' cap modifications.

5. Methods used to quantify RNAs prior to sequencing: An appropriately sensitive and precise measurement of RNA input is critical. Current implementations of widely used flourimetric or uv spectrophotometric methods adapted for small sample inputs offer mutually supporting complementary data, and agreeing measurements by two different methods are advised. A few applications do not permit such measures (i.e. single cell RNA-Seq).

III. RNA Sequence Experiment Design: Replication and sequencing depth

1. Replicate number: Experiments should be performed with two or more biological replicates, unless there is a compelling reason why this is impractical or wasteful (e.g. overlapping time points with high temporal resolution). A biological replicate is defined as an independent growth of cells/tissue and subsequent analysis. Technical replicates from the same RNA library are not required, except to evaluate cases where biological variability is abnormally high. In such instances, separating technical and biological variation is critical. In general, detecting and quantifying low prevalence RNAs is inherently more variable than high abundance RNAs. A typical R^2 (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between 0.92 to 0.98. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.

2. Sequencing depth. The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample. Experiments whose purpose is to evaluate the similarity between the transcriptional profiles of two polyA+ samples may require only modest depths of sequencing (e.g. 30M pair-end reads of length > 30NT, of which 20-25M are

mappable to the genome or known transcriptome, Experiments whose purpose is discovery of novel transcribed elements and strong quantification of known transcript isoforms requires more extensive sequencing. The ability to detect reliably low copy number transcripts/isoforms depends upon the depth of sequencing and on a sufficiently complex library. For experiments from a typical mammalian tissue or in which sensitivity of detection is important, a minimum depth of 100-200 M 2 x 76 bp or longer reads is currently recommended. [Specialized studies in which the prevalence of different RNAs has been intentionally altered (e.g. "normalizing" using DSN) as part of sample preparation need more than the read amounts (>30M paired end reads) used for simple comparison (see above). Reasons for this include: (1) overamplification of inserts as a result of an additional round of PCR after DSN and (2) much more broad coverage given the nature of A(-) and low abundance transcripts.

IV. Information to report: Sample preparation metadata

1. Method of Preparation of cDNA. The method of preparing cDNAs made from the targeted RNAs for sequencing must be described in the metadata supplied with sequencing data. The specific of cDNA priming (oligodT or random) and the length of RNA at the time of cDNA copying when random priming is used affect the outcome significantly and should be reported. Information concerning the cDNA preparation should include information concerning all details involved in making cDNAs for sequencing, including the use of bar-codes for multiplex sequencing.
2. Quantitative standards (spike-ins). It is highly desirable to include a ladder of RNA spike-ins to calibrate quantification, sensitivity, coverage and linearity. Information about the spikes should include the stage of sample preparation that the spiked controls were added, as the point of entry affects use of spike data in the output. In general, introducing spike-ins as early in the process as possible is the goal, with more elaborate uses of different spikes at different steps being optional (e.g. before poly A+ selection, at the time of cDNA synthesis, or just prior to sequencing). Different spike-in controls are needed for each of the RNA types being analyzed (e.g. long RNAs require different quantitative controls from short RNAs). Such standards are not yet available for all RNA types. Information about quantified standards should also include:
    a) how many individual spike-ins used
    b) source of the spike-ins (home-made or commercial or NIST)
    c) amount added for each individual spike-in
    d) spike sequences for primary read data use by others
    e) the concentration of each of the spike-ins in the pool used.

3. cDNA and sequencing design: This information should indicate whether the method allowed for the generation of strand specific or unstranded data, whether the sample consists of pooled and bar coded RNA targets, sequencing platform used, depth of sequencing (e.g. number of reads obtained), length of sequence reads, whether the reads are in single or paired-end format.

V. Information to report: Post-sequencing mapping, read statistics, quality scores

1. Mapping of sequence data: Multiple short read mapping algorithms are currently available to map reads to genome assemblies and to transcript model

collections. Because they have not as yet been subjected to systematic comparisons (though this is planned), data producers should select the program they feel provides the best results. Information specifying the mapping needs to be provided in sufficient detail to be reproduced, including: the program and version used, parameters employed, etc. In addition, information concerning the sequence version of the reference genome and, as appropriate, collections of transcript model sequences or splice junction collections (i.e. Refseq, Genecode, UCSC). Treatment of reads mapping equally well to more than one site in the reference genome or model set needs to be specified.

2. Specifying thresholds used:
   a) number of allowed mis-matches, minimal score, etc.
   b) treatment of multiple mapping reads: Was there a cap on number of loci to which multiple reads were distributed (e.g. only loci with <10 reads are reported). Specify the algorithm by which multi-mapping reads were distributed.
   c) quality scores used to filter the reads
   d) paramters used to trim reads (e.g. those based on quality scores and presence of linkers)
   e) for "split reads", whether there are constraints regarding the location of the splits (i.e. within the same chromosome; within a certain genomic interval) and regarding the sequences at the split (allowed only at canonical junctions, etc.)

3. Information concerning mapping strategy. Specify if mapping was performed relative to the genome and/or transriptome or a combination, and if so, information on the version of the genome used, and the transcriptome of reference (RefSeq, ENSEMBL, GENCODE, etc). Specify the order of steps in the mapping pipeline: simultaneous genome and transcriptome mapping, or stepwise mapping, and whether novel annotation derived from the data itself is part of the final mapping.

4. Information concerning mapped results: Several baseline statistics should be provided for each sample. Reporting 4a-4d below are expected, while 4e-4g are recommended when applicable.

a) total number of uniquely mapped reads (i.e. occurs once in reference genome)

b) if paired-end reads are used, report the number of mapped read pairs (involving the mated pair or each read) and the number of mapped single reads.

c) When appropriate, provide quantitation of various mapped elements including exons, splice sites, CAGE and PET tags, transcripts. Published approaches of normalization (RPKM/FPKM) are among those that might be applied.

d) Reproducibility of replicates. Evaluation of reproducibility for the existence and quantification of different RNA types (e.g. long vs. short RNAs, CAGE vs., splice sites) currently require treatments suited to the specific data-type and its analysis. Establishing the best practices for each is an active area of research. Algorithmic approaches such as IDR (http://www.encodestatistics.org/publications/IDR101.pdf) can be applied if determined appropriate. The specific implementation and settings

used for any of these algorithms should be explicitly reported and made available. For IDR the result should be used as a metric of reproducibility and not as a metric to derive a false discovery rate. Alternatively, correlation metrics as a function of prevalence can be used. This may prove more sensitive and be more appropriate. For messenger RNA, a target is that biological replicates display > 0.9 correlation for transcripts/ features greater than 1 RPKM in two or more replicates.

e) Estimating depth of coverage for mRNA-Seq. A routine approximation of the sequence coverage achieved for an mRNA-seq experiment can be made from mappable reads (unique and multiple mappers). To estimate the sequence coverage per mRNA of an average length (ignoring that there is actually a broad length distribution) present at 1 copy per cell based on an estimated input of the number of mRNAs the following calculation can be used: (Total sequence NT in the sequencing reaction / Estimate of the Number of Molecules of mRNA/cell) / (1,500NT/mRNA). Example: 10^10 nucleotides sequenced / 2X10^6 mRNAs/cell = 5X10^3NT sequence coverage per/mRNA. 5X 10^3 NT /1.5 X^3 NT/mRNA ~3X sequence coverage of an RNA present at one copy per cell. More prevalent RNAs have proportionately higher coverage. This is, of course, highly sensitive to the number picked for mRNAs/cell, and this number is poorly known for most systems. In addition, in the cases of tissues/organs, whole animals acting as the source of the RNA, an estimate of the number of cells is difficult.

f) Empiric coverage. This evaluation can be performed informatively for spike-in standards and/or for the top ~ 30% of annotated RNA in the prevalence spectrum. While uniform sequence coverage over each transcript is a goal in RNA-Seq, it is widely appreciated that this is not fully achieved using the current dominant technologies. Variables affecting uniformity include the method of cDNA synthesis, secondary structure of individual transcripts, length of the RNA template at the point of cDNA priming, method of priming, and GC content. Evaluating overall coverage (inclusion) and quantitative uniformity of coverage are further complicated for genes with multiple possible transcript isoforms than for genes with a single isoform. It is therefore recommended that assessments of coverage be performed for sub-sets of transcripts having single isoforms. Spike-in standards, in this context, offer the advantage of presenting a single known isoform. A simple average empiric coverage calculation can use the total sequence mapped for spike-ins, or for an appropriately selected endogenous transcript set, divided by the known calculated sequence length of that RNA (i.e. its sequence complexity). Coverage is expected to be a function of the prevalence of the transcript in the sample and of depth of sequencing.

g) The 3'-5' coverage ratio can be used as a metric in mRNA-seq. This is highly relevant to polyA selected templates and/or success of oligo-dT priming, but it is not appropriate for many other transcriptome fractions and sample preparation protocols. In theory, one would like this ratio to approach 1.0 for mRNA, but the existence of multiple transcript isoforms

with differing 3' and 5' exons can complicate this expectation, as will technical issues with end-representation (see below). A 3'/5' ratio is therefore best applied to a sub-set of test loci having unique 3' and 5' exons, and also to spike-in RNAs that are polyadenylated and added prior to any oligo dT selection step. If the 3'/5' ratio is high for such mRNAs or spike-in standards, it suggests problems with degraded RNA input or with cDNA synthesis biases. Given over- and under-representation of sequence coverage of the 5' and 3' 100-200 nucleotides of mRNA during high throughput sequencing (Hillier, et al 2009 Genome Res 19:657) the coverage ratio should be evaluated outside these end effects.

5. Estimating technical and mapping error: Matching millions of short and often error-prone reads against 100's of millions or billions of bases of genomic sequence generates mapping errors. Particularly problematic sources of mapping error in RNA-Seq include sequence reads from multigene families and paralogous genes, when some members are highly expressed in the sample, since these will generate many instances of reads that vary slightly from each other, and whose mapping - with tolerated mismatches - will therefore be in error. The mis-mapping frequency for these is expected to be a complex function of read length and format (paired-end or single reads), sequencing platform biases, mismatch threshold, and expression levels of paralogous genes represented in a transcriptome.

In principle, one could address the problem of a null model for the majority of the transcriptome that is nonparalogous by using known non-transcribed portions of the genome to derive a null model to estimate the extent of mapping error. However, the biology of transcription in most organisms appears to include some transcription of much of the genome, making it difficult to implement this approach with confidence. In addition, genomic or organelle DNA contamination of an RNA prep can thwart interpretation of read mapping error by this approach, when reads from DNA are wrongly attributed to mis-mapped transcript reads. This is a current research challenge, motivating estimates of sequencing error and mis-mapping by simulation and by other methods, that can then inform and help to adjust thresholds of detection at transcript and splice junction levels. It is therefore important to report, in sufficient detail to allow reproduction by others, the nature of any null model and/or the derivation of thresholds by the methods used on a study-by-study basis.

6. Analysis of spike-in standards: Using the reads that map to the spike-in standards it is possible to determine:

a) Individual spike-in detection
b) Percent of spike-in sequence detected (sequence coverage)
c) Correlation of spike-in data with its dosed amount
d) Quantification of antisense sequences to spike-ins for strand specific methods to assess the levels of inappropriate antisense signal
e) For paired format data, quantify erroneous mate-pair frequency in which one read maps to one spike-in and the other read maps to either another spike-in or to the genome to determine a rate of strand-switching. This should particularly be assessed in experiments aimed at discovering chimeric or trans-spliced transcripts.

f) It is recommended that the average coverage be greater than 1x for spike-in transcripts that are > 1 x 10^{-6} in the sample.

VI <u>Novel Elements</u>.

When appropriate to the study, it is desirable to report information concerning novel transcribed elements (minimally defined as clusters of reads that are not connected to any annotated transcript) and their estimated transcript abundances. Some currently available programs assemble and report these as simple read clusters, while others produce detailed models of novel transcript elements from RNA-Seq input. Depending on the software and appropriateness of the data, quantification at the level of individual transcript isoform models is offered or quantification can be at the level of read clusters. The program, version, parameters and, if appropriate, transcriptome annotation and model set used should be reported in a standard GFF format, with transcripts that are compatible with explicitly compatible with curated/reference annotations labeled with the corresponding annotation transcript ID.

While not required for the submission of raw read mappings, care should be taken to explicitly define the level of "novelty" that is claimed in the analysis of the data and the level of compatibility with the existing, curated annotations. For example, a transcript might be "novel" when compared to Refseq yet already be in GENCODE for human in polyA+ samples corresponding to mRNA. Furthermore, novel transcript isoforms of known genes should be sub-categorized as: (a) 5' or 3' extensions of known transcripts, (b) novel splice isoforms [and whether they change the predicted protein-coding sequence], (c) whether they encode an ORF or known subtype of non-coding RNA (snoRNA, etc…), (d) whether they could be the result of incomplete splicing from nuclear RNA, and (e) whether they are related to repeat elements (whether overlapping or adjoining to ribosomal RNA repeats, for example). Elements or transcripts containing non-canonical splice junctions should be explicitly flagged as such. Novel transcripts that are anti-sense to known models in RNA-seq samples from strand-preserving protocols should be filtered for low-level strand-mismapping contamination (which can be measured from the spike-ins). When replicates are available, the novel annotation should be present in both replicates.