ENCODE Guidelines and Best Practices for RNA-Seq:
Revised December 2016

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. 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 periodically 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.

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. Metadata to be supplied with each RNA-seq experiment.

In an effort to keep the ENCODE and affiliated NHGRI data structured into a common format the ENCODE DCC has devised a series of JSON files to capture relevant metadata for each step. Metadata can be input both programmatically in bulk and also using a web-interface. This process is human facilitated and curated at the DCC wrangler - data provider interface to ensure accurate and complete (meta)data modeling and transfer. For a detailed description of the underlying database and metadata schemas see (http://dx.doi.org/10.1101/044578). It should be noted that the schemas are living documents and evolve alongside the project to capture current needs. For reference, some of the key properties to note are:

1. **Donor Information:** All human donor should be de-identified but when available the following properties should be recorded:

    a. Sex
    b. Age
    c. Ethnicity
    d. Any familial (parent, child, twin) relationships to other known donors also being profiled.
    e. Any allowable (not protected) external links to other donor information and data sources (e.g. 1,000 Genomes donors, PGP project donors, etc…).

2. **Sample Information**: Donors contribute samples in the form of tissue, or primary and immortalized cell types. The DCC has extensive metadata schemas to capture various properties about the samples as often this is key to being able to interpret the data. This can be done in both a structured (JSON) format as well as a more loose collection of accessory PDF files (FACs analysis, plasmid maps,) provided by the data provider. Some things that are important to note are:

a. What kind of material it is should be noted: Tissue, cell line, primary cell type, etc…
b. It's ontology term (a DCC wrangler will work with you to obtain this)
c. If any treatments or genetic modifications (TALENs, CRISPR, etc…) were done to the sample prior to RNA isolation.
d. If it's a subcellular fraction or derived from another sample. If derived from another sample, that relationship should be noted.
e. Some sense of sample abundance: RNA-Seq data from "bulk" *vs*. 10,000 cell equivalents can give very different results, with lower input samples typically being less reproducible. Having a sense of the amount of starting material here is useful.
f. If you received a batch of primary or immortalized cells, the lot #, cat # and supplier should be noted.
g. If cells were cultured out, the protocol and methods used to propagate the cells should be noted.
h. If any cell phenotyping or other characterizations were done to confirm it's identify, purity, etc.. those methods should be noted.

3. **RNA Information**: RNAs come in all shapes and sizes. Some of the key properties to report are:

    a. Total RNA, Poly-A(+) RNA, Poly-A(-) RNA
    b. Size of the RNA fraction: we typically have a + 200 and – 200 cutoff, but there is a wide range, i.e. microRNA-sized, etc…
    c. If the RNA was treated with Ribosomal RNA depletion kits (RiboMinus, RiboZero): please note the kit used.

4. **Protocols**: There are several methods used to isolate RNAs with that work fine for the purposes of RNA-Seq. For all the ENCODE libraries that we make, we provide a document that lists in detail:

    a. The RNA isolation methods,
    b. Methods of size selections
    c. Methods of rRNA removal
    d. Methods of oligo-dT selections
    e. Methods of DNAse I treatments

    These "Production Documents" are found as a .pdf file that is appended to each experiment and we supply the DCC with one document per library to describe the methods leading to its construction. These are all available for download at encodeproject.org.

5. **RNA Quantification and Quality Control**: When working with bulk samples, throughout the various steps we periodically assess the quality and quantity of the RNA. This is typically done on a BioAnalyzer. Points to check are:

    a. Total RNA
    b. After oligo-dT size selections
    c. After rRNA-depletions
    d. After library construction

6. **Library Construction**: Throughout ENCODE, several methods have been used to make RNA-Seq libraries, from homebrew protocols to kits. They all work to varying degrees have different biochemical signatures. It is important to note the kit being used for library construction and report it. For RNA-Seq, it's particularly useful to note if the protocol generates stranded or unstranded data.

7. **Sequencing**: There are several sequencing platforms and technologies out there being used. It is important to provide the following pieces of information:

    a. Platform: Illumina, PacBio, Oxford Nanopore, etc…
    b. Format: Single-end, Pair-end,

c. Read Length: 101 bases, 125 bases, etc…
d. Unusual barcode placement and sequence: Some protocols introduce barcodes in non-customary places. If you are going to deliver a FASTQ file that will contain the barcode sequences in it or other molecular markers – you will need to report both the position in the read(s) where they are and their sequence(s).
e. Please provide the sequence of any custom primers that were used to sequence the library.

III. RNA Sequence Experiment Design: Replication, sequencing depth, spike-ins

1. Replicate number:

In all cases, 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. As part of the ENCODE pipeline, annotated transcript and genes are quantified using RSEM and the values are made available for downstream correlation analysis. Replicate concordance: the gene level quantification should have a Spearman correlation of >0.9 between isogenic replicates and >0.8 between anisogenic replicates.

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. Experiments whose purpose is discovery of novel transcribed elements and strong quantification of known transcript isoforms requires more extensive sequencing.

- Each Long RNA-Seq library must have a minimum of 30 million aligned reads/mate-pairs.
- Each RAMPAGE library must have a minimum of 20 million aligned reads/mate-pairs.
- Each small RNA-Seq library must have a minimum of 30 million aligned reads/mate-pairs.

3. 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) A FASTA (or other standard format) file containing the sequences of each spike in.
b) Source of the spike-ins (home-made, Ambion, etc..)
c) The concentration of each of the spike-ins in the pool used.

IV. Post-Sequencing QC: Post-sequencing mapping, read statistics, quality scores

1. **Mapping of Sequence Data**: https://www.encodeproject.org/rna-seq/long-rnas/

Multiple short read mapping algorithms are currently available to map reads to genome assemblies and to transcript model collections. ENCODE has worked up a common pipeline that can be used for ENCODE RNA-Seq assays, using the STAR mapper. It is deployed on the DNANexus Environment and also

available in a GitHub (https://github.com/ENCODE-DCC/long-rna-seq-pipeline). However, people are also welcome to also provide their own data, mapped how they chose outside of the pipeline and work with the DCC to capture the relevant File, Software (version, parameters) and workflow.

When mapping RNA-Seq data, ENCODE makes the following recommendations:

a. It is preferable to include not only the genome reference but also a set of annotations into the mapping set since this greatly increases the specificity of mapping to splice junctions.
b. The sequences for any exogenous RNA spike-ins that were added should also be included to obtain mapped spike-in reads.
c. The ENCODE 3 data was predominantly mapped using RNA-STAR. One of the outputs of STAR is a log.final.out file which reports several summary stats (mapping rate, # of splice junctions, etc…) and this can be used as an initial gauge to assess performance between various libraries.