

BETA CELL BIOLOGY CONSORTIUM

RNA-Seq Minimal Standards and Guidelines



Version 1 - Last modified on May 3, 2011



This document is derived from ENCODE and modENCODE standards and is provided by the BCBC Bioinformatics and Epigenomics Workgroup. Please contact the workgroup with suggestions or questions.

CONTENTS

A.	Background.....	1
B.	Information to be supplied with each sample used for RNA-seq experiment.....	1
C.	Performance of RNA Sequence Experiment: Number of replicates and sequencing depth.....	2
D.	Information supplied concerning steps taken prior to the sequencing reactions.	3
E.	Information supplied concerning post-sequencing mapping, read statistics and quality scores	3
F.	BCBC3.0 Addendum.....	6
G.	Checklist.....	6

A. BACKGROUND

The utilization of RNA sequencing (RNA-seq) experiments to characterize transcribed elements and quantify RNA expression using RNA has increased in response to improvements in sequencing technologies, to reduction in costs and to improvements in the computational tools that assist in interpreting the results. To guide the generation of high quality data this document aims to provide uniform standards and guidelines for RNA-seq experiments. Due to rapidly changing and emerging sequencing technologies and data types, this standards document should be revised annually. RNA-seq experiments are already used to characterize several types of RNA isolated from either whole cells or sub-cellular compartments. These include long RNAs (i.e. >200nts), short RNAs (i.e. <200nts), polyA+, polyA-, 5' end CAGE tags and 5' and 3' end PET tags. They can also be used to multiplex targeted regions such as those amplified by RT-PCR, RACE, etc. A variety of different sequencing technologies are used as are a number of different analysis methods. Standards are likely to need some modification depending on the sample, the RNA-seq methodology and the analysis methods. The following attempts to capture the types of information that should be reported on each, and provides guidelines for experimental design.

B. INFORMATION TO BE SUPPLIED WITH EACH SAMPLE USED FOR RNA-SEQ EXPERIMENT

RNA-seq data should be accompanied by information concerning the biological source of the RNA and the protocols used to extract 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. Protocols for synchronization of animals and methods used for purification of tissue or cell types (e.g. FACS),
 - b. Amounts of starting material (tissue/organ weights, cell number from which sub-cellular compartments were isolated, etc.),
 - c. 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),
 - d. Lines of mice used and their genetic background.
 3. Identification of the type of RNA targeted (size range, poly A+ or A-, 5' capped or uncapped).
 4. Protocols used to isolate RNAs (size range, 5'/5'-3'tags, poly A+/A-). While each of the RNA types have unique issues associated with obtaining as enriched a sample as possible, evidence of the enriched status of the targeted RNA type should also be presented. 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. Whether the experiment generates strand-specific information should be explicitly stated. The use of amplification strategies and expected impact on strand or 3' bias must be recorded.
 6. If an RNA amplification method is used, supporting information about the method and estimated fold of amplification should also be provided.

C. PERFORMANCE OF RNA SEQUENCE EXPERIMENT: NUMBER OF REPLICATES AND SEQUENCING DEPTH.

1. Replica number: In order to ensure that the data are reproducible, experiments should be performed with at least two biological replicates, unless there is a compelling reason indicating that 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 of the same library are not required. Low copy number RNAs will be inherently more variable – subject to stochastic noise. Also, some variation will exist from sample to sample prepared under the same conditions. A typical R^2 correlation of gene expression between two biological replicates using cell lines is normally 0.92 to 0.98. For biological replicates between individuals or samples from transgenic animals/embryos correlations can be lower, depending upon the sample. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained and values reported.

2. Sequencing depth. The amount of sequencing to be done on a sample is determined by the goals of the experiments. Experiments whose purpose is to evaluate the similarity between the transcriptional profiles of two samples may require only modest depths of sequencing (e.g. one lane yielding 30M of which 20-25M reads are mappable). Experiments whose purpose is discovery of novel transcribed elements will require more extensive sequencing. The ability to detect reliably low copy number transcripts depends upon the depth of sequencing. For experiments involving complex transcriptomes in which sensitivity of detection is important a minimum depth of 100-200 M reads is recommended.

D. INFORMATION SUPPLIED CONCERNING STEPS TAKEN PRIOR TO THE SEQUENCING REACTIONS.

1. Method of Preparation of cDNA Samples for Sequencing. The method of preparing cDNAs made from the targeted RNAs for sequencing must be described in the metadata supplied with the sequencing results. The current sequencing methods provide either long (>200 nt) or short (<200nt) lengths of contiguous sequence. For each of these methods, there are optimum levels of sample to be analyzed (e.g. cluster numbers for Illumina). Information concerning the cDNA preparation should include information concerning all details involved in making cDNAs for sequencing.
2. Information about the use of known quantitative standards (spike-ins) added to the sample to be sequenced. A ladder of RNA spike-ins should be included where available to calibrate quantification, sensitivity, coverage and linearity. Information about the spike-ins should include the stage of sample preparation that the spiked controls were added (e.g. before poly A+ selection, just prior to beginning sequencing protocol). It should be recognized that different spike-in controls will be needed for each of the RNA types being analyzed (e.g. long RNAs require different quantitative controls from short RNAs). However, such standards are not available for all RNA types as yet. 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. Amounts added for each individual spike-in.
3. Method of treatment of cDNA sample for sequencing. This information should indicate whether the method allowed for the generation of either stranded or unstranded data, whether the sample consists of pooled and bar coded RNA targets, what the depth of sequencing (e.g. number of lanes that were run), lengths of the reads, whether ϕ X control was utilized and whether the reads are intended to be single or paired end) should be provided.

E. INFORMATION SUPPLIED CONCERNING POST-SEQUENCING MAPPING, READ STATISTICS AND QUALITY SCORES

1. Mapping of sequence data. There are multiple short read mapping algorithms currently available. Because these mapping programs have not as yet been subjected to a systematic comparison (though this is planned) data producers may utilize the program they feel provides the most comprehensive and accurate set of results. The potential use of multiple mapping algorithms requires that information concerning: what program was utilized, the parameters employed, etc. In addition, information concerning the sequence version of the reference genome should be

provided. All FastQ files should be deposited into the BCBC databases.

2. Each algorithm will require decisions concerning thresholds utilized during mapping. Information concerning such thresholds includes:
 - a. Number of allowable mis-matches, minimal score, etc.
 - b. Treatment of multiple mapping reads. Was there any cap on read number placed on number of loci to which multiple reads are reported to map (e.g. only loci with <10 reads are reported).
 - c. Quality scores used to filter the reads
 - d. Parameters 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 (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. This includes whether mapping is performed relative to the genome and/or transcriptome mapping, and if so, information on the version of the genome used, and the transcriptome of reference (RefSeq, ENSEMBL, GENCODE, etc). This also includes the order of the steps in the mapping pipeline: simultaneous genome and transcriptome mapping, or stepwise mapping.
4. Information concerning mapped results. There are several baseline statistics that should be provided for each sequenced sample. These include:
 - a. Total number of unique reads (i.e. occurs once in reference genome),
 - b. Mappable read pairs. If paired-end reads are used, report the number of mappable read pairs (involving the mated pair or each read of the pair individually),
 - c. Relative quantification of various mapped elements. When appropriate, an estimate of relative quantification of various mapped elements including exons, splice sites, CAGE and PET tags, transcripts. Published approaches of normalization (RPKM/FPKM) might be used,
 - d. An estimate of the sequence coverage achieved by an RNA-seq experiment can be specified in terms of sequence coverage obtained 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 / 2×10^6 mRNAs/cell = 5×10^3 NT sequence coverage per/mRNA. 5×10^3 NT / 1.5×10^3 NT/mRNA $\sim 3 \times$ sequence coverage of an RNA present at one copy per cell. 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 and will also likely make such estimates inaccurate.
 - e. Evenness (uniformity) of coverage. This measurement can be performed for both the spike-in standards and the top 30% of annotated RNA in the prevalence spectrum to ensure that there are enough statistically significant reads. Such measurements can be attempted in several ways depending on the reference. From longest to shortest, nucleotide lengths of genome (total or

non-repeat portion), of transcripts or exons can serve as the denominators for coverage calculations.

- f. 3'-5' coverage ratio. This is highly relevant to polyA selected templates and/or oligo-dT priming, but not nearly as accessible for other kinds of transcriptome fractions and inputs. One would like this ratio close to 1.0 for mRNA. If the 3'/5' is high, it can signal problems such as degraded RNA input or serious cDNA synthesis biases. Given that there has been reported 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 made outside these end effects.
1. Reproducibility of replicates. Evaluation of the reproducibility of different types of RNA types (e.g. long vs. short RNAs, CAGE vs., splice sites) will require individualized analysis. Use of algorithmic approaches like IDR (<http://www.encodestatistics.org/publications/IDR101.pdf>) can be employed. However, the setting needed and used for any of these algorithms should be explicitly provided. It is important that the results be interpreted as a metric of reproducibility not as a metric to discover a false discovery rate. Alternatively, correlation metrics as a function of prevalence can be used. This is more sensitive; arguably more appropriate but also is more difficult. It is recommended that biological replicates must show > 0.9 correlation for transcripts/ features greater than 1 RPKM.

Estimates of technical and mapping errors. Matching millions of short error prone reads against 100's of millions or billions of bases of genomic sequence can result in mapping errors. Low levels of DNA can also contaminate an RNA sample. Ideally, known non-transcribed portions of the genome could be used to derive a null model that could be used to estimate the extent of such mapping problems. However, deriving such null model can be problematic since we often cannot identify non-transcribed regions with confidence. Nonetheless, we should attempt to estimate the frequency of sequencing and mis-mapping in simulations or by other methods to determine how much density and junction noise is expected for technical reasons and adjust thresholds for detection accordingly. [use KO data to derive]

5. Information concerning alignment of reads to spike-in standards. Using the reads that map to the spike-in standards it is possible to determine:
 - a. How many of spike-ins were detected
 - b. What percent of the spike-in sequence was detected
 - c. Correlate the detected spike-in with its dosed amount
 - d. Detection of antisense sequences to spike-ins in strand specific preparations to determine the amount of inappropriate antisense synthesis during reverse transcriptase cDNA synthesis.
 - e. Frequency of mate-pairs where 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.

It is recommended that the average coverage is greater than 1x for spike-in transcripts that are present in a range of abundances.

6. When appropriate, information concerning novel transcribed elements and estimated transcript abundances. Programs exist that predict novel transcript elements, and that attempt quantification at the level of individual transcript isoform. If these are used, the program, version, and parameters used should be reported.

F. BCBC3.0 ADDENDUM

The purpose of this document is to provide a set of minimal standards for the technical quality of RNA-seq data being generated for BCBC 3.0 by defining the fields of information that need to be recorded for ALL RNA-seq experiments.

1. The number of replicates recommended here is only to ensure the technical quality of the RNA-seq experiments. It is understood that, based on the context of specific biological comparisons, more biological replicates will often be necessary to achieve statistically robust and biologically meaningful interpretations.
2. The purpose of the experiment determines the appropriate sequencing depth:
 - a. 20-25 M is recommended for comparing samples based on known genes,
 - b. 100-200 M for genome-annotation: novel transcripts, isoforms, etc.
3. Spiked in control is strongly recommended. A common set of spike-ins for all BCBC investigators will be discussed.
4. Raw FASTQ files will be a deliverable to Betacell Genomics.
5. The method of RNA-seq data processing—mapper software, parameters, reference transcriptome, etc., will be recorded. The final processed data (gene or transcript coordinates used and the normalized number of reads assigned to them) will also be a deliverable to Beta Cell Genomics.
6. In addition, information must be provided about the experiment (who did the experiment and what was the objective) and the design (experimental factors, relationships between samples and between samples and sequence runs). This information will be collected along with details of the samples, sample protocols, and data protocols in a standardized spreadsheet (MAGE-TAB).

G. CHECKLIST

Experiment:

- Contact person
- Objective
- Experimental factors

Samples:

- Descriptions
- Protocols used for cell isolation
- Estimated cell purity
- Identify biological vs. technical replicates
- Amplification method and amount, if used

Sequence:

- Protocols used (a protocol should be made available on the BCBC3.0 website).
- Relation to samples
- Spike-in used
- Machine (version)
- Sequence length
- Single or paired end reads
- Strand specific?
- Sequence length
- Bar codes?
- Data file of sequence reads with quality scores

Sequence analysis:

- Quality assessment (correlations between biological replicates)
- Reference genome, transcriptome
- Mapping software (version) and parameters used
- Numbers of reads, whether trimming was done: total, mapped, unique
- File of gene/ transcript coordinates and normalized number of assigned reads (RPKM/ FPKM)