REMC Standards and Guidelines for RNA-sequencing

DRAFT v3.0

I. Introduction

Next-generation sequence based transcriptome methodologies (broadly referred to as RNA-seq) were initially developed in 2007 for massively parallel short read sequencing platforms. RNA-seq involves purification of RNA, followed by either selection of poly-A(+) RNA or depletion of ribosomal RNA. RNA is then converted into cDNA by one of two methods; 1) random priming, followed by cDNA fragmentation, end-repair and Illumina/SOLiD linker ligation or, 2) Enzymatic or chemical RNA fragmentation followed by linker ligation and cDNA generation. Following PCR amplification of tailed cDNA fragments with primers suitable for solid phase (Illumina) or emPCR (SOLiD) clonal amplification RNA-seq libraries are subjected to sequencing. Sequence alignment software is then used to compare the short sequence reads to reference genome and transcriptome databases, and exon-exon junction databases. From this analysis paradigm emerges data that is used for a variety of purposes, including the measurement of gene-level and exon-level expression abundance; detection of base changes (mutations and polymorphisms) relative to reference datasets; measurement of alternative splicing events; identification of gene fusion events; and identification of RNA editing events.

In parallel to RNA-seq, the small RNA content of a transcriptome can be captured and sequenced (miRNA-seq). miRNA-seq library construction involves; 1) extraction of total RNA; 2) 3' pre-adenylated RNA linker ligation 3) 5' ATP dependent RNA linker ligation; 4) RT using a primer which hybridizes to the 3' RNA linker followed by PCR with primers suitable for solid phase (Illumina) or emPCR (SOLiD) clonal amplification. Following PCR the desired insert size range, typically 18-30 bp, is purified away from ligation and PCR by-products by gel electrophoresis. Taking advantage of the fact that the complexity of miRNA-seq libraries is significantly less than that of RNA-seq libraries, miRNA-seq libraries are typically indexed and pooled prior to sequencing.

Due to the numerous advantages of RNA-seq over array-based platforms for transcriptome characterization, RNA-seq has been embraced by the research community and stabilized, production scale RNA-seq methodologies have been developed. Large scale projects such as The Cancer Genome Atlas, TARGET and the Cancer Genome Characterization Initiative have now largely switched to RNA-seq for transcriptome profiling. However, RNA-seq remains an immature technology and it is expected that advances in molecular techniques and sequencing technologies will continue to shape its development in the months and years to come. To ensure that the recommendations in this document remain relevant it should be reviewed and updated on a yearly basis.

II. Total RNA extraction and QC guidelines.

Some commercially available RNA extraction kits utilize columns that can deplete the small RNA fraction during total RNA extraction making the resulting RNA unsuitable

for miRNA-seq library preparation. A standard Trizol based extraction methodology or a column run under non-selecting conditions (for example Ambion mirVANA (the TCGA standard)) is recommended for REMC total RNA extraction.

Following extraction, RNA integrity must be determined, recorded and provided. High quality RNA is essential for RNA-seq library construction. Degraded RNA can lead to increased noise in the resulting mRNA-seq library as measured by transcript coverage, mitochondrial content, ribosomal content and gene diversity (Figure 1). It is recommended that a minimum threshold RIN 7, as measured by an Agilent bioanalyzer, or equivalent, be applied to all REMC RNA samples.

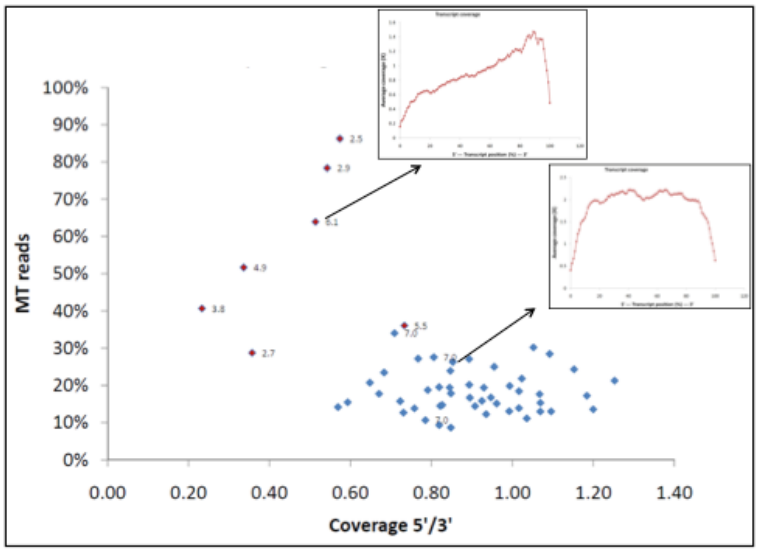


Figure 1. Mitochondrial and average trancript coverage versus RIN value. mRNA-seq libraries constructed from RNA with RIN >=7 (blue) and RIN <=7 were assessed for mitochondrial and average transcript coverage. Insets provide example coverage plots used to calculate the 5'/3' coverage metric. Libraries were sequenced to an average depth of 120 million reads (6Gb aligned) using PE50 sequencing on an Illumina platform.

III. Performance of the RNA-seq Sequence Experiment: Library construction protocol and sequencing depth.

1. Due to protocol specific biases introduced during RNA or cDNA amplification it is recommended that neither RNA nor full-length cDNA amplification be performed for REMC RNA-seq libraries.

2. It is recommended that polyA+ RNA be used for REMC RNA-seq libraries.

3. Both random primed and RNA fragmentation based methodologies are suitable for REMC RNA-seq library preparation. If a random primed method is used, it is recommended that the subsequent cDNA be rendered single stranded prior to PCR amplification to maintain strand specificity.

4. Paired-end sequencing should be the method of choice for RNA-seq library sequencing

due to the improvements in gene, exon and exon-junction detection enabled by the paired read (Figure 2).

Figure 2. Gene and exon detection by single end (SET) or paired end (PET) sequencing. A RNA-seq library was sequenced using paired-end 75 base chemistry on an Illumina platform. For the comparison, the second read was discarded (SET 75) or for PET, trimmed back to 50 bases. Alignment was performed independently to genome, transcriptome and exon-exon junction resources using bwa and exon and gene coverage enumerated.

5. The number, type and length of sequence reads required to sample a RNA-seq library is dependent on the library preparation methodology and analysis type to be performed (see Figure 3). For samples that will form REMC complete epigenomes, where the goal is to comprehensively represent polyadenylated transcription within a cell or tissue type, a minimum depth of ~200 million paired-end reads per replicate (representing 100 million cDNA fragments) of 75 nucleotides in length is required. In cases where more than two replicates of a tissue or cell type will be performed ~50-100 million paired-end reads per additional replicate (representing 25-50 million cDNA fragments) of 75 nucleotides is sufficient.

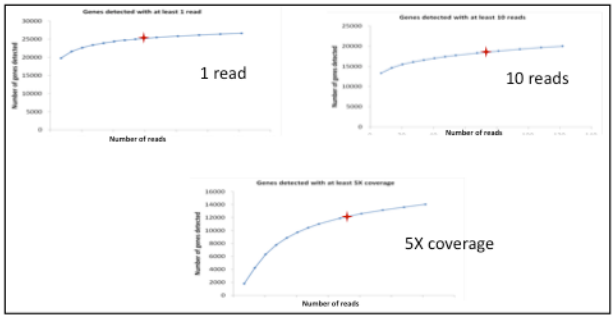


Figure 3. Read depth vs. gene detection. Red star indicates 100 million aligned reads. PolyA+ RNA-seq libraries sequenced as PE50s on the Illumina platform.

IV. RNA-seq Sequence Experiment QC Metrics.

The following QC metrics should be determined and monitored to ensure that REMC RNA-seq libraries are of high quality:

1. Exon:Intron ratio;
   -assessed to detect potential genomic contamination (the mRNA-seq protocol should include a standard DNAse treatment step).

2. Overall transcript coverage;
   -assessed to detect mRNA degradation and degree to which full-length mRNA was randomly sampled. Uniform coverage should be obtained for 1Kb transcripts

3. Total number of duplicate reads (identical forward and reverse read starts);
   -a measure of library diversity. Tissue and cell type dependent but outliers should be investigated as potential library failures.

4. Fraction of reads mapping to mitochondrial transcripts;
   -assessed to detect RNA degradation. Tissue and cell type dependent but outliers should be investigated as potential library failures.

5. Fraction of reads mapping to ribosomal transcripts;
   -assessed to detect degree of polyA enrichment. Tissue and cell type dependent but outliers should be investigated as potential library failures.

6. Fraction of reads mapping to intergenic regions; and
   -assessed to detect potential genomic contamination (the RNA-seq protocol should include a standard DNAse treatment step).

7. Percentage of reads inappropriately aligned to anti-sense strands should be <=1%. This can be determined by enumerating the fraction of exon-exon junction reads aligned to the anti-sense strand. Alternatively known poly-adenylated foreign RNA standards can be spiked into the RNA prior to library construction and used to determine the level of artifact anti-sense transcription.

V. miRNA-seq Sequence Experiment QC Metrics.

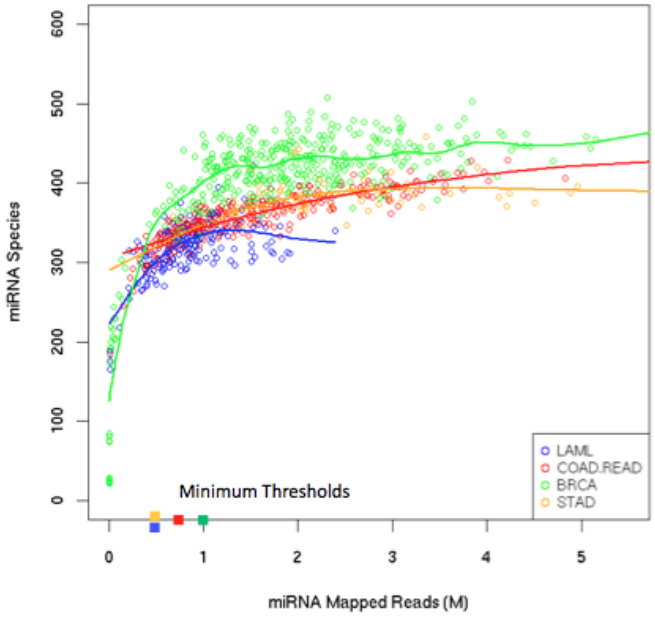
The following QC metrics should be determined and monitored to ensure that REMC miRNA-seq libraries are of high quality.

1. Saturation plots of miRNA species diversity to assess sequence depth (Figure 4). miRNA diversity appears to be tissue dependent.

2. Fraction of total reads that are adapter dimer (the major miRNA-seq library construction biproduct). Dimer fraction increases as RNA quality and miRNA quantity decrease in a given sample.

3. Fraction of aligned reads that align to other small RNA species (tRNAs, snoRNAs etc..).

4. Fraction of aligned reads that align to mature miRNA, miRNA* and pre-miRNA sequences.



Each point represents a miRNA-seq library sequenced in pools of 8 on the Illumina GAiix platform. The resulting plots are used to set thresholds on a depth by tissue.

Samples with reads below threshold are re-pooled for more sequencing.

LAML = Acute Myeloid Leukemia
COAD,READ = Colon/Rectum Adenocarcinoma
BRCA = Breast Invasive Carcinoma
STAD = Stomach Adenocarcinoma

Figure 4. Read depth vs. miRNA detection