The Children's Hospital of Philadelphia®
RESEARCH INSTITUTE

CENTER FOR APPLIED GENOMICS

RNAseq (Bulk) sample submission/guideline

RNA sequencing (RNA-Seq) uses the capabilities of high-throughput sequencing methods to provide insight into the transcriptome of a cell. RNA-Seq technology allows for quantification of gene expression, facilitates the discovery of novel transcripts, and identifies alternatively spliced genes and the detection of allele-specific expression. This technology can classically investigate polyadenylated messenger RNA (mRNA) transcripts, but also can be applied to investigate different populations of RNA, including total RNA, pre-mRNA, and noncoding RNA, such as microRNA and long ncRNA. A classic RNA-Seq experiment consists of design/planning, isolating RNA, converting RNA to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on a Sequencer. Nevertheless, many experimental details, dependent on a study objective, should be considered before performing RNA-Seq experiments. These include the use of biological and technical replicates, depth of sequencing, read length (single read vs paired end), library preparation type, selection of cases and controls, and desired coverage across the transcriptome.

Please follow the requirements below when planning the experiments:

1. Isolation of RNA

The first step in transcriptome sequencing is the isolation of RNA. The RNA quality is typically measured using the Nanodrop spectrophotometer for concentration and purity check (260/280 ratio = 1.8 to 2.2). The “Gold Standard” method for integrity check is capillary electrophoresis. The measurement of the integrity of the RNA and a more precise measurement of the concentration should be performed using capillary electrophoresis device. (Bio-Rad Experion, Agilent Bio-Analyzer/TapeStation, or PerkinElmer LabChip GX). These methods produce an RNA Integrity Number (RINe) between 1 and 10 with 10 being the highest quality and showing the least degradation. The RINe refers to the sample integrity using gel electrophoresis and analysis of the ratios of 28S to 18S ribosomal bands. Note that the RINe measurements are based on mammalian organisms and certain species with abnormal ribosomal ratios (i.e., insects) may erroneously generate poor RINe scores. Be aware that low-quality RNA (RINe < 6) can considerably affect the sequencing results (e.g., irregular gene coverage, 3′–5′ transcript bias, etc.) and lead to arbitrary biological conclusions. Unfortunately, high-quality RNA samples may not be available in some cases (FFPE, autopsy, long surgical procedures, laser capture microdissection) and the effect of degraded RNA on the sequencing results should be prudently considered and discussed with the CORE staff.

CAG quantity and purity criteria for RNAseq are:

a. The Nanodrop OD 260/280 ratio should be above 1.8 to 2.2, the OD 260/230 ratio should be above 2.0
   - OD260/230 nm <1.6: indicates potential contamination with Guanidinium isothiocyanate or other chaotropic agents absorbing at 230 nm. This is seen if the wash buffer is carried through in column purifications.
   - OD260/ 280 nm < 1.6: indicates potential contamination with phenol absorbing at 270 nm. This is seen if part of the phenol phase is aspirated when collecting the aqueous phase in a phenol:chloroform extraction.

b. RQI/RINe-score is > 6.0. (Measurement with capillary electrophoresis). Open to discussion with the CORE staff.
   - RIN 7-10: Little to no degradation; whole transcriptome or mRNA RNA sequencing possible.
   - RIN value 5-7: Partial degradation; whole transcriptome and mRNA sequencing possible. However, mRNA sequencing will be poor since mRNAs will start to lose their 3´ poly-A tails.
   - RIN value <5: Degraded RNA; only whole transcriptome sequencing is possible but partial sequencing of degraded rRNA material unavoidable.

   WARNINGS:

   - To avoid bias, all samples must be treated in similar conditions.
   - If there is evidence of gDNA presence, DNAse treatment should be considered.
   - Large differences in RINe values between samples must be evaluated (example: one biological group with high quality RNA, one with degraded RNA).
   - RNA method used for extraction and EDTA content. EDTA can be a compromising solution if in excess for cDNA and library generation. EDTA can inhibit enzymatic fragmentation and 1st/2nd strand cDNA synthesis.

c. Concentrations and volumes according table below:

Automated Library Preparations			
Library Preparation	RINe	Concentration (Total RNA Input)	Comments
Illumina Stranded mRNA	7-10	50ng – 1 ug	stranded, polyA based
Illumina Stranded Total RNA	7-10	100 ng – 1 ug	stranded, ribodepletion
Takara	4-10 (+FFPE)	250 pg- 10 ng	low input/poor RINe

2. How many biological replicates do I need?

The number of biological replicates needed for whole transcriptome or mRNA sequencing varies and will depend on the aims of the experiment. We do not have a specific number, but it is prudent that at least three biological replicates per sample group are included. Replicates allow for statistical tests and data comparisons to be performed, but ideally, we suggest more if monetarily and experimentally feasible.

With the recent advances in technology, the reproducibility and robustness of technical replicates are well proven and we suggest prioritizing biological replicates over technical replicates for the majority of studies. However, we STRONGLY recommend discussions with the CAG staff about the experimental design to define the replicate requirements.

3. Which Library and RNA type will be used?

The library will be chosen based on the QC performed on the samples: integrity and concentration. Also, the library recommendation can be discussed with the CORE director based on the experimental design and QC results as a whole.

- Whole transcriptome sequencing: this option permits the characterization of all RNA transcripts for a given system including both the coding mRNA and non-coding RNA larger than 170 nucleotides in length independent of polyadenylation (snRNAs and snoRNAs larger than 170 nucleotides in length are also included). However, since ribosomal RNA (rRNA) corresponds to the majority of the transcriptome content (~90%), the ideal library preparation is aimed to remove the ribosomal RNA (AKA: ribodepletion) prior to sequencing (oligo-based, capture/hyb pulling). This increases the depth of sequencing for the relevant part of the transcriptome instead of wasting the reads with rRNA content. The library preparation also keep strandedness information of the RNA transcripts which characterize from which of the two DNA strands a given RNA transcript resulted from. This provides increased assurance in transcript annotation and enables the detection of antisense transcript expression.

- mRNA sequencing: targets all polyadenylated (poly-A) transcripts of the transcriptome. The mRNA library (also called PolyA enrichment) preparation targets poly-A tailed transcripts from total RNA as part of the library preparation protocol. The poly-A tailed transcripts include the mRNAs (the coding part of the genome), which only accounts for 1-4 % of the whole transcriptome. The enrichment of poly-A tailed transcripts results in a higher depth of sequencing. The increased sequencing depth improves the sensitivity in identifying under-expressed mRNA transcripts. The library preparation also aims to keep strandedness information of the RNA transcripts. This provides increased confidence in transcript annotation. Notes: This method is not ideal for poor quality samples (low abundancy of viable transcripts when degraded). Remember that mitochondrial poly-A tailed transcripts are highly abundant sequences and can be filtered out with bioinformatic tools before mapping takes place.

4. Sequencing options: Number of reads, read length and single or paired end reads?

- Depth: Depth of sequencing is one of the most critical factors with respect to both differential expression analysis and discovery of novel transcripts. The final number of reads is continually questioned in the NGS scientific community and many studies demonstrate that for Poly-A enriched experiments, 30 million reads is the minimum for most tissues. However, for detection and analysis of low expression mRNAs, over 100 million reads may be needed. CAG recommends a minimum of 20 million reads per sample for most applications. Any custom read depth should be discussed beforehand. Keep in mind that splicing investigation requires a higher depth of sequencing which needs to be discussed with the Bioinformatics team prior to sequencing. However, if needed, it is possible to re- sequence the same library and adjust the number of reads or add extra reads at an additional cost (please contact us if you are interested in this option).

- Read length: The number of sequencing cycles defines the read length. The length of the reads is related to the alignment specificity but longer reads can also affect the overall quality since the quality score (Q-score) drops as the reads get longer. Therefore, CAG recommends read lengths of 50bp or more for most applications. In general, paired-end (PE) sequencing (where sequencing is performed from both ends of the molecule) is considered superior compared to single-end sequencing. Paired-end sequencing facilitates better alignment and mapping (by decreasing the rate of alignment ambiguity) which increases transcript assembly confidence. Paired-end sequencing is therefore recommended for discovery applications such as detecting and characterizing novel splice isoforms. Notes: There are a number of warnings associated with comparing libraries of different read lengths and depth, therefore it is recommended to use the same for all samples of the study. During your consultation with CAG, we will discuss the specific goals of your project and make recommendations concerning the ideal number of reads per sample, read length and single or paired end reads.

5. Preparing your samples to the CAG

- The isolated RNA solution should be transferred to an *RNAse free* 96-well plate or *RNAse* free low-bind 1.5 mL tubes
- To avoid seeing batch differences in the final data, it is important to put the samples on the plate in a random order, starting with position A1, then B1, etc. Do not leave any blank wells blank between the samples. Make sure the right seal is used.
- Record the sample IDs (DO NOT USE PATIENT NAMES) in a special XLS file according to the criteria established by the CORE (model attached).
- Please use a permanent marker to annotate each plate on both sides.
- The plates should be stored and transported at –80ºC dry ice. Deliver the samples on dry ice to the Genome Analysis Facility lab or send your samples by courier to the CAG Facility.
- Aliquot ~5 ul of sample for QC in case you have not performed QC in your own lab.
- Communicate with the CAG CORE team with any questions and concerns.

CAG Address: (Hours: M-F 9am-5 pm)
Center for Applied Genomics (CAG) – NGS LAB
The Children's Hospital of Philadelphia (CHOP)
3615 Civic Center Boulevard, Room 1014
Philadelphia, PA 19104-4318
(267) 426-0181 (T)
billingsj@chop.edu
CAG NGS iLAB page

6. Data analysis

The data analysis plays a very important role in the NGS RNAseq pipeline, not only to evaluate the overall quality of the run but to perform a thoughtful biological interpretation of the results. CAG has a very skilled team of Bioinformaticians who provide comprehensive data analysis appropriate for specific experiments and project goals. The bioinformatics team is an integrated part of our NGS platform and our scientists have a strong background in both the experimental and analytical aspects of NGS. With that in mind, we consider each project independently to determine the most appropriate analysis tools and pipeline to answer the relevant biological questions. Ours services are offered as a fee for service model or collaboration efforts. Our data analysis includes:

- Comprehensive QC of sequencing data
- Mapping: Alignment of reads to the specified reference genome (Human, Mouse, Rat, drosophilia, etc.)
- Identification of splice-junctions
- Identification of alternate splicing and splice variants
- Identification of antisense transcripts
- Quantification of known transcripts (both Ensembl and UCSC are supported)
- Prediction and quantification of novel transcripts
- Test for differential expression at gene and transcript level as well as tests for significant changes in promotor usage
- Normalization and group comparison (unsupervised clustering: principle component analysis and heatmap)
- Gene Ontology Enrichment Analysis (GO Analysis)