mRNA SEQUENCING

INTRO

mRNA Sequencing at NYGC can be performed on the HiSeq 2500 using 2x125bp read length or Novaseq using the 2x100bp read length. We offer library preparation utilizing the Illumina TruSeq Stranded mRNA sample preparation protocol, which employs a polyA enrichment followed by strand-specific cDNA synthesis. It is critical for the success of this preparation method that the RNA polyA tails not be degraded, which is why we cannot accept samples with RIN values <8 (details below). The service is inclusive of sample QC, library prep, sequencing, and standard analysis. Delivery includes aligned .bam files as well as count matrix files and differential expression results (details below). For RNA sequencing of any species without a polyA tail, please refer to our total RNA Sequencing service.

INPUT REQUIREMENTS

Upon receipt of samples, NYGC will perform QC first by fluorescence-based quantification using RiboGreen, and second by measuring the RNA integrity on the BioAnalyzer or Fragment Analyzer depending on submission size. Investigators will be notified of samples that fall below the required total mass or that are degraded and not suitable for library preparation. Samples that do not meet the requirements may still be processed for sequencing based on customer decision but in that case NYGC takes no responsibility for sub-optimal results.

The sample submission requirements are as follows:

mRNA SAMPLE REQUIREMENTS

- A minimum of 2µg total, DNase-treated RNA is required
- Samples should be submitted in a total volume of 50 µl to 100µl nuclease-free water
- Samples should have absorbance values of OD260/280=>1.9
- Sample quality should be ascertained by BioAnalyzer with a RIN value >=7
- Samples should be quantified by RiboGreen

LIBRARY PREPARATION

RNA will be prepared using the Illumina TruSeq Stranded mRNA sample preparation kit**.** All samples submitted within a study will be prepared in a single batch by one operator to eliminate preparation batch effect. The library QC will include a measurement of the average size of library fragments using the FragmentAnalyzer and estimation of the total concentration of RNA by picogreen.

SEQUENCING

Sequencing can be performed on the HiSeq 2500 or Novaseq instruments. The HiSeq 2500 generates roughly 200M-250M passed filter single end reads per flow cell lane and Novaseq S2 generates roughly 1.4-1.6 billion single end passed filter sequencing reads per flow cell lane. For RNA projects in Novaseq we use 2x100bp read length.

QUALITY CONTROL METRICS

Assessment of the quality of the sequencing data will include multiple metrics at several steps of the analysis pipeline. Following the completion of a sequencing run, a QC specialist will review the sequencing quality metrics including: number of pass filter reads per sample, base quality per cycle, percent base content per cycle, and the overall distribution of base quality scores. If the raw sequencing data passes quality control threshold, it will be placed into the alignment pipeline.

Post-alignment, Picard and RSeQC will be used to generate a sample specific metrics report. For RNA sequencing, relevant metrics include alignment statistics, %GC, gene body coverage, insert size, rRNA contamination, duplication rate, unsupervised clustering, batch effect analysis, library complexity, etc.

ANALYSIS

Steps in the NYGC RNA analysis pipeline include:

- Alignment of raw reads to hg19 using STAR aligner
- Quality control using RSeQC and Picard (%GC, %duplicates, gene body coverage, unsupervised clustering, library complexity, etc.)
- Picard MarkDuplicates
- Gene quantification using featureCounts
- Gene annotation using Gencode
- Normalize count matrix with DESeq2
- Differential expression with DESeq2
- Transcript quantification with Kallisto and differential isoform expression with Sleuth
- Fusion gene discovery with FusionCatcher (for somatic studies only)

DELIVERABLES

The files delivered at the completion of a project include:

- Expected number of reads per sample as specified in the service description
- >80% of bases sequenced with a quality score above Q30
- BAM format file with corresponding BAI index file
- Tables of splice junction and non-canonical alignments
- Quality control metrics
- Raw count matrix with all genes and samples
- Normalized count matrix
- Regularized count matrix
- R object containing the results of the Kallisto/Sleuth pipeline
- FusionCatcher table (for somatic studies only)
- 3 months of data storage, unless otherwise specified

TURNAROUND TIME

Turnaround time for projects with 200 or less samples is 6 weeks from the date samples pass QC in the NYGC laboratory. If a project is greater than 200 samples, NYGC would deliver 100 additional samples per week. Please discuss any expedited turnaround needs with your Project Manager.