INTRO

Total RNA Sequencing at NYGC has been developed on the HiSeq 2500 using 2x50bp or 2x125bp read length. We offer library preparation utilizing the KAPA Stranded RNA-Seq Kit with RiboErase for rRNA depletion (human, mouse, and rat species-specific), which employs an enzymatic depletion followed by strand-specific cDNA synthesis. 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). Samples are recommended to be submitted in duplicate or triplicate for statistically significant results. For RNA sequencing of species other than human, mouse, or rat, please refer to our mRNA 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. 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 but in that case NYGC takes no responsibility for sub-optimal results.

The sample submission requirements are as follows:

TOTAL RNA SAMPLE REQUIREMENTS

• A minimum of 2µg of total, DNase-treated RNA is required
• Samples should be submitted in a total volume of 15µl to 30µl nuclease-free water
• Samples should have absorbance values of OD260/280=>1.9
• Sample quality should be ascertained by BioAnalyzer (or equivalent) with a RIN value >=7
• Samples should be quantified by RiboGreen (or equivalent)
• A minimum submission of 8 samples

LIBRARY PREPARATION

RNA will be prepared using the KAPA Stranded RNA-Seq Kit with RiboErase. 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 DNA by PicoGreen.

SEQUENCING

Sequencing will be performed on the HiSeq 2500 instrument; libraries will be loaded onto the HiSeq 2500 flowcell for clustering on the cBot using the instrument-specific clustering protocol. The HiSeq 2500 generates roughly 200M-250M passed filter sequencing reads per flow cell lane after alignment and duplicate removal; samples will be sequenced to the target number of reads discussed at the start of the project. Each instrument processes two flow cells (16 lanes) simultaneously, and the run time is approximately 5 days. The NYGC currently operates 12 HiSeq 2500 instruments.
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)

DELIVERABLES
The files delivered at the completion of a project include:
• BAM format file containing all passed filter reads and quality scores with corresponding BAI file
• Splice junction and non-canonical alignments tables
• 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)
• 3 months of data storage, unless otherwise specified

TURNAROUND TIME
NYGC estimates turnaround time from the date samples pass QC in the NYGC laboratory. Typical turnaround times for projects of <500 samples is about 6-8 weeks dependent on the queue when samples arrive.