# NYGC Quality Control

All samples undergo rigorous quality assessment using a comprehensive set of quality measures upon completion of each step of sample processing: 1) sample receipt, 2) library preparation, 3) sequencing, 4) data analysis. Quality control measures are scrutinized by the combined efforts of our Project Management, Laboratory, Sequencing Analytics, and Bioinformatics teams. Samples that do not meet our expected quality criteria are flagged and reviewed in consultation with the investigator prior to initiation of the next step of the sample processing pipeline.

# RNA Preparation

## Extraction

Total RNA is extracted from flash frozen post-mortem tissue. Trizol/Chloroform extraction method is used, followed by Qiagen RNeasy minikit column purification. The column purification step is used to ensure the quality of extracted RNA .

# Sample QC

## Automated volume check

An automated volume check on investigator samples submitted in our 2D matrix rack tubes is performed as part of our initial QC for each sample upon arrival in the laboratory. This information is matched with the sample volume information provided by the investigator to confirm that sample integrity was not compromised during shipment

## RNA Quantification

Total RNA is quantified using fluorescent-based assays (RiboGreen or Qubit) to accurately determine whether sufficient material is available for library preparation and sequencing. A Nanodrop 2000 reading provides a 260/280 ratio, a generally accepted measure of purity of extracted RNA.

## RNA Integrity

RNA sample size distributions are profiled by a Fragment Analyzer (Advanced Analytics) or BioAnalyzer (Agilent Technologies), to assess sample quality and integrity. Samples that contain degraded material and/or DNA contaminants, which could affect library preparation performance, are flagged.

Samples that do not meet our initial criteria for QC undergo further review with our Project Management team in consultation with the investigator. Investigators are provided the opportunity to re-submit new or additional material for samples that do not pass our initial QC criteria.

# Library Preparation

## Manual Library Preparation

Libraries are prepared using KAPA Stranded RNA-Seq Kit with RiboErase and unique Illumina-compatible PCR primers with indexes purchased from BioScientific (NEXTflex RNA-seq Barcodes, cat# 512915, 8nt index).

Libraries are prepared using 500ng of total RNA input, 550bp in length for sequencing on the HiSeq 2500.

## Automated Library Preparation

Libraries are prepared using the KAPA Stranded RNA-Seq Kit with RiboErase (kapabiosystems) in accordance with the manufacturer’s instructions. Briefly, 500ng of total RNA is used for ribosomal depletion and fragmentation. Depleted RNA undergoes first and second strand cDNA synthesis. cDNA is then adenylated, ligated to Illumina sequencing adapters, and amplified with 9 cycles of PCR for sequencing on the NovaSeq 6000.

# Library QC

## Library Quantification

Picogreen or a Qubit™ 2.0 Fluorometer is used to measure the total amount of RNA in the prepared library.

## Library Size distribution

Size distribution profiles of the final libraries are assessed using the Fragment analyzer/Bioanalyzer. Libraries that fall outside of the expected size range and/or contain adapter dimer contaminants are flagged.

# Sequencing

## HiSeq 2500

Sequencing is performed on an Illumina HiSeq 2500 sequencer (v4 chemistry) using 2x125bp cycles targeting ~40 million reads per sample.

## NovaSeq 6000

Sequencing is performed on an Illumina NovaSeq 6000 sequencer using 2x100bp cycles targeting approximately 40 million reads per sample.

# Sequencing QC

All sequencing runs are reviewed for quality by our Sequencing Analytics team. Sequencing runs that do not pass our quality criteria for each of the metrics below are flagged and reviewed in consultation with Illumina Technical Support.

## % Pass Filter (PF) clusters

Library cluster efficiency should fall within the optimal range expected for instrument and flowcell type. PF percentages outside the expected range indicate either incorrect loading concentrations or problems with a particular sequencing run.

## % sample de-multiplexed

All PF reads within a single lane of a flowcell are assigned to a specific barcoded library based on the indexed read. The percentage of reads within a lane that are assigned to each sample after de-multiplexing is assessed to confirm expected sample distribution within the sample pool.

## # of PF reads/sample

The total number of PF reads per sample must meet the expected number of reads for a given sequencing application and analysis type, as discussed upfront with the investigator. Samples that do not meet the expected number of reads are queued for additional sequencing.

## % bases >Q30

To ensure the highest quality sequencing data, FASTQ data in which at least 75% (HiSeq X) or 80% (HiSeq 2500) of bases have an Illumina Quality score >30 (a Phred like score indicating an expected 99.9% base call accuracy) are selected and used in downstream analysis.

## Quality by cycle

Assessment of the quality score by cycle is used to verify that the accuracy of called bases is maintained across the entire length of the sequencing read.

## GC content

GC content is reflective of both sample and library type. GC content can vary between organisms, and can be an indicator of poor sequence quality attributable to biases introduced during library preparation.

## K-mer content/adapter contamination

FASTQC data is examined to identify over-represented sequences in the sequencing data, including k-mers and reads that align to the Illumina adapter sequences, both of which could indicate poor library quality and result in uneven base composition.

# Data Analysis QC - RNA

## Total reads/sample

The total number of reads per sample must meet the expected total number of reads for a given sequencing application and analysis type, as discussed upfront with the investigator. Samples that do not meet the expected number of reads are queued for additional sequencing.

## % rRNA

The proportion of reads in sample that align to ribosomal RNA (rRNA) provides a measure of the success of upfront rRNA (ribo)-depletion during total RNA library preparation, and polyA- enrichment during mRNA library preparation. Higher than expected levels of reads mapping to rRNAs can lower the signal-to-noise ratio, and can have an impact on downstream analysis.

## % Duplicates

PCR amplification during library preparation can give rise to the duplication of reads. Libraries that produce a higher than expected number of duplicate reads result from reduced library complexity and reduced representation of the underlying transcript diversity. Higher than expected duplication rates can also indicate reduced levels of sample complexity attributable to lower amounts of starting material used for library preparation.

## % Aligned

The proportion of reads successfully mapping to the reference genome.

## % Gene assignment

The percentage of mapped reads that are assigned to annotated gene regions are marked as follows: 1) % coding (CDS), 2) % un-translated 3’ and 5’ regions (UTRs), 3) % intronic (non-coding regions within genes), 4) % intergenic (non-coding outside of genes). Samples that show lower or higher than expected assignment to these genomic regions could indicate quality issues related to technical artifacts introduced during library preparation and/or sequencing, initial quality of the RNA sample, or could reveal information about the origin of the sample (tissue type).

## % Strandedness

The percentage of sequenced reads with the correct “strandedness” is verified following preparation of libraries using a stranded library preparation protocol (which preserves DNA strand information).

## 5'/3’ coverage

Read coverage is inspected for uniformity across gene bodies. Ideal experiments show uniform coverage; bias in the representation of reads at either the 5’ or 3’ ends of gene bodies can indicate poor sample quality and/or technical artifacts introduced during library preparation.

## % Mean GC

The mean GC content averaged for each sequenced read is used to flag libraries that show biases introduced through PCR amplification during library preparation and/or sequencing.

## Insert size

The mean inner distance between the end of Read 1 and the start of Read2 are calculated to confirm that the library insert sizes are appropriate for the sequenced read length. Smaller insert sizes lead to overlapping reads and/or sequencing into the adapter sequences, limiting the number of usable bases for mapping and downstream analysis.

## Xist vs chrY gender check

Gender-specific transcript expression from the X and Y chromosomes is used to determine the gender of the sequenced sample and compare to the gender specified in the sample submission form.

## Unsupervised clustering to check for unexpected structure in data (batch effects, sample swapand contamination)

Hierarchical clustering of the RNA sequencing data is performed to validate sample identity and grouping based on experimental design. Unexpected clustering results can reveal potential artifacts in the experiment, such as sample swaps, contamination, sample quality variability and/or technical batch effects, all of which undergo further investigation.

## Tissue markers

A pre-defined set of tissue-specific transcript expression markers are used to determine the source tissue of each sample.

## DNA/RNA Concordance

The identity of the RNA sample is confirmed by evaluating concordance with whole genome sequencing data using Conpair, a tool that uses a set of SNPs common in the human population to determine sample identity.

## Sample contamination

Contamination checks are performed on the sequencing data to verify that the sequenced library originates from a single individual/sample. Samples in which a detectable level of contamination is identified are flagged for review. This information is used to identify sample swaps and to confirm sample identity (in combination with Concordance and Tissue marker analysis).