**Outline of general quality control procedures for**

**Add Health RNA profiling at UCLA Social Genomics Core Lab**

This memo outlines QC procedures and metrics for RNA profiling in the UCLA Social Genomics Core Lab and UCLA Neuroscience Genomics Core Lab, including 1) key procedures to optimize sample and assay quality, and 2) measures of achieved quality. Most key procedures constitute a criterion that is met before subsequent assay steps commence (e.g., normalizing sample density, balancing the mass of samples that are pooled for multiplex assay, etc.) and thus do not yield any outcome metric. Measures of empirical sample quality and achieved assay quality typically do generate metrics, and are noted below (METRICS).

**1. Sample processing and RNA extraction**

**Preshipment administrative control:** Secure an electronic manifest (e.g., Excel spreadsheet) listing all sample IDs for all samples expected to ship (and no more) PRIOR to sample shipment. This allows us to meet administrative validity requirements to recognize shipping errors that might lead to assaying samples that are not properly consented for RNA analysis.

**Shipment receipt:** Verify general size of shipment (expected number of boxes), arrival timing (samples expected to arrive within 24 h of shipping), sample temperature (samples expected to arrive on dry ice), and physical condition (no damaged boxes or other evidence of compromise). Photodocument any anomalies while transferring samples into storage at -80C AS SOON AS POSSIBLE (to minimize temperature deviation/freeze-thaw cycling). Record, decontaminate, and discard immediately any open/exposed samples to minimize cross-contamination.

METRICS: Photodocumentation of physical shipment anomalies. (Note that we so far have not had any such issues with Add Health samples.)

**Shipment storage:** Maintain all samples in -80C freezer with constant temperature monitoring & temperature anomaly alarm. Maintain an empty back-up freezer in case of primary freezer failure; in the event of primary freezer failure, transfer all samples to back-up freezer before samples reach -20C.

**RNA extraction:** Retrieve and thaw samples (PAXgene RNA tubes) in sequential batches of 12 (or multiples, e.g., 24 or 48) commensurate with available capacity of automated nucleic acid processing systems (Qiagen QIAcubes).

Conduct thaw procedures in laminar flow biosafety cabinet (to minimize contamination and blood-borne pathogen exposure). During thaw, verify sample IDs against manifest. DO NOT process unlisted samples (instead return to “Quarantine Box” in freezer). Note absent samples (present in manifest but not in expected sequential position). Notify Add Health administration of any mismatch relative to manifest.

Follow standard Qiagen protocol (using Qiagen supplies) to extract total RNA from PAXgene RNA tubes in batches of 12 (<https://www.qiagen.com/us/resources/resourcedetail?id=dc649975-b6ba-43e7-a3e9-e9986c612d2e>). All protocol procedures are conducted at Biosafety Level 2 with standard “PCR clean/RNA safe” procedures (including double gloves and face masks to prevent sample contamination; use of RNAse-free tubes, pipette tips, and other plasticware in all procedures; 1-time use & disposal of all pipette tips to prevent cross-contamination; conduct of procedures within a laminar flow biosafety cabinet whenever feasible; minimize duration of sample vessel exposure to ambient air & temperature; maintain samples within secondary containment whenever feasible, e.g., air-capped centrifuge rotors). Immediately on completion of extraction protocol, quantify the RNA density and purity by spectrophotometry (Nanodrop ND1000) and freeze the remainder at -80C.

METRICS: 1) Inventory any sample anomalies noted on working manifest. 2) Extracted RNA mass (ng), density (ng/ul), and purity (A260/A280) for each sample is added to the Working Manifest.

No QC sample gating (e.g., discarding suboptimal samples) takes place at this stage.

**Sample normalization & plating:**

The RNA sequencing procedures used in this study take approximately 360 ng of input RNA in a 12 ul volume (2 ul of which is consumed by additional QC assays outlined in the next section, and the remaining 10 ul of which serves as input into the cDNA library synthesis for sequencing) delivered in wells of a specified 96-well plate. In this procedure we thaw individual Extracted RNA samples, and generate 12 ul Assayed RNA samples that are adjusted to a density of 30 ng/ul (or as close to that target as feasible given sup-optimal Extracted RNA mass), and dispense the Assayed RNA samples into 96-well plates. In the majority of cases Extracted RNA samples have a density > 30 ng/ul and are thus diluted to achieve target RNA density. In the less frequent cases where RNA density is notably less than the target (i.e., < 20 ng/ul) we use the Qiagen RNeasy MinElute Cleanup system to concentrate all available RNA mass in 12 ul (for details see <https://www.qiagen.com/us/resources/resourcedetail?id=06f3d3ff-5926-4915-bffe-4adcf99266aa>). Following this concentration procedure we quantify the resulting RNA density and purity by spectrophotometry (Nanodrop ND1000) and dispense the entire available RNA mass as the Assayed RNA sample. In the rare case where RNA density remains below the limit of detection after concentration, samples are excluded from further assay procedures. Because the RNA sequencing workflow we use often yields high-quality data with detectable but sub-optimal input RNA (sub-optimality factors as low as 20% of target input), and there is a strong interest in avoiding any form of sample bias in Add Health, we do not generally delete samples with any detectable RNA mass from assay at this stage. Instead, we assay all technically feasible samples, and we use post-assay data validity metrics described below (Endpoint sample QC) to determine which samples end up being empirically valid for analysis (put another way, pre-assay QC metrics have not proven highly predictive of post-assay data validity except in the 0 or near-0 RNA mass, so we do not attempt to forecast ultimate validity from midstream QC metrics given their weak predictive relationship).

METRICS: 1) Final Assayed RNA mass (ng) and density (ng/ul), and 2) the “sub-optimality factor” indicating how closely the available Assayed RNA mass approached the target of 360 ng (values ranging from 0-1, corresponding to achieved fraction of ideal Assayed RNA mass).

**2. RNA sequencing**

Batches of 4 plates (384 samples; or the closest approximation for the final batch for a given year’s shipment) are delivered to the UCLA Neuroscience Genomics Core Laboratory (UNGC) for cDNA synthesis, using the Lexogen QuantSeq 3’ FWD system, and cDNA library sequencing, using an Illumina HiSeq 4000 instrument, both following the manufacturer’s standard protocols for this workflow.

**Assayed sample QC:** RNA sequencing procedures begin by quantifying Final Assayed sample RNA density using an additional RNA-binding dye measure (PicoGreen Fluorescence) and sample RNA quality (i.e., RNA integrity vs. degradation) using a standard electrophoresis system (Agilent TapeStation) with integrity quantified by standard RNA Integrity Number (RIN). RIN values range from 0-10, and the Lexogen QuantSeq 3’ FWD system specifies a RIN value > 3 as optimal input. (Note that this system tolerates a much wider range of input RIN values than do most other mRNA-targeted cDNA synthesis systems because this system targets only a small (65 bp) 3’ portion of the mRNA strand rather than attempting to sequence the entire transcript; as such it is immune to RNA degradation events that occur outside the 3’ 65-bp region adjacent to the poly-adenylation tail).

METRICS: Assayed Sample 1) PicoGreen RNA mass and 2) RIN.

**RNAseq assay QC:** The Lexogen QuantSeq 3’ FWD cDNA library synthesis system involves multiple QC steps within the library synthesis protocol (see <https://www.lexogen.com/wp-content/uploads/2019/03/015UG009V0251_QuantSeq_Illumina.pdf> for more details). Briefly, poly-A-selected Assayed Sample RNA is converted to cDNA and amplified by quantitative PCR to produce a “qPCR library”, with the specific number of PCR amplification cycles selected based on a quantitative PCR side-reaction that is used to optimize settings for an entire batch of samples. The resulting qPCR library is assayed by electrophoresis (Agilent TapeStation with hsDNA tapes) to verify that DNA population size distribution is appropriate for a cDNA library derived from mRNA, and quantified using DNA-binding dye assay (PicoGreen) to verify expected yield >1 ng/ul. Based on exact yield measures, individual qPCR libraries are normalized to 3-5 pMol concentration, and pooled into batches (typically 24 samples) for multiplex sequencing on the Illumina HiSeq 4000 instrument. Assays target >10 million 65-nt single-stranded sequencing reads per sample. qPCR library pools are optimized to a DNA concentration of 3 pMol for the key clustering reaction that drives the Illumina “sequencing by synthesis” assay chemistry.

The Illumina HiSeq 4000 produces a large array of technical quality control parameters for instrument technical function, and the key parameter for assessing the integrity of the sample assay process itself is the “Q-Score,” which measures the accuracy with which individual nucleotide base calls can be resolved. The Q-score is a variant of the classical Phred quality score and is defined by the same metric of -log10(probability of erroneous call), with Q scores >30 indicating >99.9% basecalling accuracy. (See <https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf> for more details on this metric.) The overall performance of a sequencing run is quantified by the fraction of sequenced DNA fragments (“reads”) with Q-score > 30 (“%Q30”, which is typically expected to be 75%-95%). Another subordinate measure of successful multiplexing is relative balance across samples in the total number of reads acquired per sample (e.g., expecting each sample to yield approximately 1/24 of the total number of reads in a 24-sample pool, with absolute read frequencies targeting > 10 million reads/sample). In general, if a sequencing run fails on any of theses integrated performance metrics then the sequencing process is repeated. In other words, we do not receive from the UNGC data that has not passed these integrated sequencing run performance metrics.

Note that the “run metrics” above pertain to the quality of the overall assay process for a pool of samples, and do not provide information about assay quality for individual samples (or the individual reads within a sample). However, read-specific QC measures are provided in the assay output data files in the form of Phred quality scores that are included within the “FASTQ” data files that are produced by the assay process itself. The utilization of those FASTQ quality measures is discussed in the next section (Post-assay data processing).

**3. Post-assay data processing and Endpoint QC**

**Alignment & transcript abundance quantification:** Raw read data in the form of FASTQ files (i.e., 65-nt strings of A’s, C’s, T’s, and G’s) are aligned to the reference human transcriptome sequence (ENSEMBL hg38) using the STAR aligner (see Dobin et al in <https://www.ncbi.nlm.nih.gov/pubmed/23104886> for more details) to quantify the number of individual reads that map to each human gene transcript in each sample. The STAR aligner makes use of the individual nucleotide call reliabilities (Phred quality scores) that are stored in parallel with the read nucleotide sequence within each FASTQ file (see https://en.wikipedia/wiki/FASTQ\_format for more detail on FASTQ file format and content). As such the read-level QC data enters directly into the read-mapping process itself and the results of that process thus take into account assay QC metrics. Given that, there is no need for any additional QC gating on individual reads. If a read is consistently or predominately poor in quality, STAR will not judge it to align to any human transcriptome region and will allocate the read to an unmapped read bin during the alignment process. If a read is only partially poor, then the quantitative degree of uncertainty will be taken into account by STAR in judging which if any human gene transcript the read maps to. The STAR aligner documentation reference above provides more information on how FASTQ quality scores enter into the alignment computations and the threshold at which a read is declared unmapped.

In general, for optimal RNA samples (e.g., from PAXgene RNA tubes), we expect successful read alignment rates >90% (i.e., < 10% of reads unmapped). For suboptimal RNA samples with poorer biological signal-to-noise ratios (e.g., dried blood spots) valid data can have read alignment rates that average 80% and reach 70% or lower.

It may be useful to take the quantitative variation in sample-specific alignment rates into account in data analyses (e.g., as a validity weighting factor or technical covariate), although much of the validity information available from this measure is also picked up by downstream endpoint validity metrics such as the sample correlogram / profile consistency measure described below, and so alignment rate information typically does not add much useful information. However, it is reasonable to gate out samples from analysis if alignment rates are quite low for a particular sample (e.g., < 80%) but other samples in the data set routinely achieve higher alignment rates (e.g., > 90%). If all samples hover in the area of 80% mapped, however, that may reflect the challenging sample-collection environment (e.g., average 80% is typical for dried blood spots) rather than any sample-specific deficiency. Put another way, alignment rates are most useful as a relative QC metric for identifying technically aberrant samples (although NOT biologically aberrant samples), and the average alignment rate across samples can be taken as a general signal-to-noise reliability metric in the traditional psychometric sense (i.e., low alignment indicates measurement noise, but does not indicate bias/invalidity in the psychometric sense).

Because there is no absolute threshold for invalidity in mapping percentages and there is a strong interest in avoiding bias in Add Health, we conservatively provide whatever read mapping data is available from STAR for all samples and leave it to Add Health data users and/or data administrators to determine which samples should be gated out in any given analysis; note that different gating thresholds are appropriate for different analytic purposes because the relative costs of sampling bias vs. (un-)reliability vary across substantive settings. However, we do provide guidance to Add Health users and administrators regarding the practices we typically use for QC gating in our own analyses (including the 90%/PAXgene and 70%/DBS mapping criteria we typically apply).

METRICS: 1) Total Reads per sample and 2) sample-specific Mapping Rate (or Alignment Rate).

**Read count normalization**

RNA sequencing yields absolute read counts for each gene transcript in each sample, and those absolute counts will vary based on the total number of reads acquired for each sample (or “sequencing depth”) as well as the fraction of those reads that are “lost” as unmapped. In order to remove such non-substantive “nuisance” influences, RNA sequencing data are typically normalized to a “rate” metric that expresses reads for a given transcript as a fraction of the total number of mapped reads acquired for that sample (i.e., across all transcripts). This is typically normalized to a metric of transcript-specific reads per million total mapped reads (TPM/transcripts per million mapped reads) for interpretive convenience and to avoid numerical computation problems that emerge with small decimal fractions (e.g., if normalized as a fraction of all reads then the vast majority of values would fall below 10-4).

For some research purposes it may be useful to further normalize the TPM values to equate samples on the average expression level of one or more reference genes that are believed to be constantly expressed across cells (sometimes called “housekeeping genes”). This can be particularly useful for RNA sequencing data because the total number of mapped reads (the TPM normalization denominator) can be significantly affected by the expression of a small number of highly abundant RNAs that are variably expressed across samples and/or cell subsets (e.g., ribosomal genes, which often vary in expression as a function of BMI, or hemoglobin genes, which vary in expression as a function of metabolic and RNA degradation processes in red blood cells, etc.). To support such analyses we also provide a “Reference gene-normalized” data set that normalizes samples based on average expression of 11 standard reference gene transcripts (see <https://www.ncbi.nlm.nih.gov/pubmed/23810203> for background).

To stabilize sampling variation in genes that differ by 10,000-fold or more in average expression level, it is also conventional to log2 transform the TPM values (or reference gene-normalized TPM values) prior to statistical analysis. 0 count values cannot be log transformed, so it is common to substitute a small value at the lower limit of reliable detection for 0 read counts when computing log2 TPM values. That 0-substitute is typically chosen to be the lower limit of reliable detection in the current assay context, and all values below that are raised to the 0-substitute value (i.e., expression values are “floored” at the minimum level of reliable detection). In general absolute read counts < 10 are highly variable (unreliable) in RNA sequencing, so one numerically conservative convention is to floor low TPM values at the numerical value of 10 / expected number of reads per sample. Add Health RNAseq assays target 10 million reads per sample, yielding 1 TPM as a conservatively realistic lower limit of detection. This results in a log2 TPM floor value of 0 (which is also convenient for leaving all log2-transformed data in non-negative values and making minimal expression levels easily recognized as 0 values in the log2 metric). Flooring has the effect of obscuring any systematic effects that may exist in very weakly expressed genes (risking false negatives) but is conservative in the sense of reducing the risk of spurious false positive results that are numerically unreliable.

Log2 TPM values are convenient for statistical analysis, but they have the disadvantage of concealing the total number of reads from which the rate value was estimated (e.g., a rate estimated from 10 million total reads is likely to be more accurately estimated than is a rate estimated from 10 thousand total reads). To allow statistical analyses that take into account this source of measurement precision, the total number of reads acquired for a sample is reported as a QC metric (see Alignment and Transcript quantification above). If desired, the total number of mapped reads can be derived by multiplying a sample’s total number of acquired reads by its mapping rate.

The normalization procedures outlined above successfully remedy the primary sources of technical noise in RNA sequencing data. However, they are not the only possible approach to such issues and data analysts may prefer different strategies for normalizing RNA seq read counts. To support such analyses, we also provide to Add Health raw read count data in addition to log2 TPM data and log2 Reference Gene-Normalized TPM data.

**Endpoint sample QC**:

None of the QC metrics listed above provides a comprehensive assessment of the entire assay pipeline from start (Input RNA Sample) to finish (normalized TPM values); problems can occur at any stage of the assay pipeline and good input QC (optimal RNA mass and RIN value) is not sufficient to guarantee that the ultimate data are valid and reliable. What is needed to judge ultimate validity is an “Endpoint QC” measure that provides an integrated assessment of all sample processing, assay, and data processing steps.

The best endpoint QC metric that we have been able to identify is a metric called “Average Sample Profile Correlation,” which measures the average correlation (similarity) between numerical gene expression values for a given sample and numerical gene expression values for every other sample in a given data set (or if that’s not feasible, a large number of other samples from the same data set, e.g., the other 95 samples assayed on the same plate, or a random subset of 100 other samples from the same study). The intuition underlying this metric is that a given sample’s data should be broadly similar to other sample’s data, particularly if they’re in the same study (i.e., come from the same organism and tissue source, are processed through the same assay pipeline, etc.). If a sample’s RNA expression profile is similar to others in the same data set (e.g., average pairwise correlation with other samples r > .85) that provides “convergent validity” evidence.

If a sample’s Average Profile Correlation is NOT high (i.e., this sample is NOT similar to most others in the same data set), then the question arises why. One possibility (not the only one) is that the sample and/or assay process is corrupted or degraded at some stage/s, in which case the average correlation with other samples will be quantitatively reduced (because noisy/unreliable measurement leads to an attenuation of the inter-sample correlation coefficient).

Quantitative attenuation of inter-sample profile correlations could reflect a reduction in general assay performance or it could reflect sample-specific technical issues (e.g., RNA degradation, insufficient RNA input, genetic or other biological idiosyncrasies that hamper assay efficiency). However, correlation attenuation can also occur when the sample and assay processes are valid but the input sample is biologically aberrant (e.g., the blood donor has leukemia). Poor assay performance typically affects all samples similarly, so the average profile correlation will decline for all samples (e.g., from a typical r > .90 to perhaps r = .50 to .70). By contrast, sample-specific issues affect only that sample (or only a few samples), and so other samples will continue to show high inter-sample correlations with their neighbors (e.g., r = .90) whereas the affected sample/s will show consistently lower correlation with neighbors (e.g., r = .70). The ability to distinguish between global assay performance issues vs. sample-specific issues is the reason we compute inter-sample correlations between individual samples (as opposed to correlating an individual sample with the average expression profile across all samples, which yields only a single correlation for each sample and makes it harder to determine whether a low correlation is local to the sample or stems from global assay problems). This logic is difficult to describe but easy to visualize if we compute a correlation matrix across samples, with cells containing the pairwise correlation between each sample (computed across the ~20,000 genes assayed or the > 50,000 gene transcripts assayed) and then color the cells of that correlation matrix using a “heat plot” color scheme (e.g., higher values colored more red and lower values colored more blue, with intermediate values noncolored/white). In this example, there will be a red “major diagonal” (where each sample correlates r = 1.0 with itself) and off-diagonal values will typically be in the red-to-white (intermediate) spectrum of color. If the whole assay is misperforming, all of the off-diagonal elements will be bluish and there will be little differential coloring across samples (i.e., little color heterogeneity across rows or columns of the matrix). On the other hand, if a given sample is misperforming but the assay is performing well in general, most of the off-diagonal elements will still be red-to-white, but that specific sample will have all or mostly blue intercorrelation values (i.e., low inter-sample correlation with most other individual samples in a context where higher samples are technically achievable in other samples). That has the effect of creating a blue-colored row and a blue-colored column in the part of the heat plot corresponding to that sample, with those 2 blue lines intersecting at the diagonal position of the sample row (or column). This “plus shape” pattern is visually striking and allows a QC reviewer to quickly scan large numbers of samples to identify a few aberrant individuals. Numerically, the “plus” pattern corresponds to one or a few low values of average inter-sample correlation coefficients amidst a generally higher-valued distribution for other samples, and so can also be observed by plotting a histogram of average inter-sample correlation coefficients (in which case individually aberrant samples appear as low-end outliers). This histogram provides a useful basis for discriminating between individual sample issues and global assay issues (or global sample issues, such as the general reduction in signal-to-noise ratios that stem from using dried blood spots as input).

In addition to providing a basis for sample-specific QC adjustment (e.g., screening and/or use as a quantitative reliability weight or technical covariate), analysis of Average Profile Consistency values can also be used to detect subtle variations in assay performance (e.g., by testing for association with processing sequence, assay plate, assay batch, etc.).

Given Add Health’s interest in avoiding any form of bias, we do not screen out any samples based on the endpoint Average Profile Correlation metric and leave that to downstream users who may have different preferences for balancing bias vs. noise. However, we do note that in our own downstream statistical analyses we typically screen out any sample with an Average Profile Correlation r < .85 (assuming higher values can be realistically expected, e.g., using PAXgene input RNA). In the case of dried blood spot RNA, where the limited RNA volume unavoidably leads to greater noise, we use data set-specific thresholds derived from the histogram-based outlier analysis described above, which typically yields cut-off Average Profile Correlation thresholds in the range of r = [.50, .80].

METRICS: Average Profile Correlation (Pearson or Spearman r metric).

**Internal validity analyses:** These procedures test for the presence of pre-established gene expression effects in a newly derived data set to provide “ground truth” validity information. These analyses are useful for selecting optimal normalization options and verifying that no administrative misalignment has occurred during the linkage of gene expression data to other study data (e.g., survey data in the case of Add Health).

The most important of these analyses tests for expected differences in expression of X- and Y-linked genes as a function of study data indicating participant sex. Several strongly sex-polarized gene transcripts exist and should generally show non-overlapping distributions in samples from females vs. males (e.g., X-linked XIST, which is involved in X inactivation in XX genomes, and Y-linked ribosomal protein RPS4Y1). Failure of such transcripts to show consistent non-overlapping segregation with indicated/reported sex may reflect administrative errors in matching RNA sample IDs to the external study data. However, due to the extreme effect size of these sex differences, this internal validity measure often lacks the sensitivity needed to help resolve which normalization strategy is most appropriate for a given data set. As such, it is most useful for ground truth validation and detecting administrative errors.

Other internal validity “standards” with more modest effect sizes are generally most useful for determining whether regular TPM normalization is sufficient to control the effects of technical or nuisance variation, or whether additional normalization is helpful (e.g., reference gene normalization). For this purpose it is often helpful to estimate a model that controls for some major demographic and behavioral influences on gene expression and tests those covariate-adjusted data for, 1) increased expression of pro-inflammatory genes with higher BMI, 2) relatively higher expression of Type I interferon response genes in individuals of African ancestry and relatively lower expression of those genes in individuals of European ancestry. Both of these effects are fairly consistent across studies (i.e., reflect true biological signals) but are relatively modest in effect size and therefore provide a more nuanced criterion for determining which normalization approach best optimizes signal-to-noise ratios for known benchmark findings. In the context of Add Health, internal validity analyses found reference gene normalization to yield superior signal-to-noise ratios (i.e., statistically significant associations of the expected form) whereas simple TPM normalization yielded weaker parameter estimates and statistical significance (i.e., poorer signal-to-noise ratios).

**Summary**

Available QC metrics that are potentially useful to account for in Add Health RNA data analysis (e.g., as sample exclusion criteria or quantitative reliability weights) include sample input metrics (Input Sample RNA mass and RIN), RNA sequencing read metrics (total reads per sample and alignment rate), and, most importantly, endpoint sample reliabilty criteria (average profile correlation; typically should be > .85 for PAXgene samples and data set-dependent, outlier-defined for the dried blood spot samples). In general we do not recommend gating out samples based on input sample metrics because 1) the Lexogen QuantSeq / Illumina HiSeq 4000 assay pipeline often succeeds in generating reliable data from sub-optimal input RNA, and 2) anomalies upstream and downstream of input sample RNA characteristics can also impact ultimate data validity, so input RNA metrics alone cannot accurately predict ultimate data validity. In our experience the most sound strategy is to 1) exclude samples based on outlier analysis of the Average Profile Correlation endpoint QC metric, and potentially outlier low total read counts (e.g., < 5 million reads/sample in Add Health), and in the data that remain, 2) use the Average Profile Correlation statistic and possibly also total (or mapped) read counts and/or read mapping rate and/or input RNA mass as additional technical covariates (or validity weights) in substantive statistical analyses.