RNA SEQ (COUNT DATA) 2016 HEALTH AND RETIREMENT STUDY VENOUS BLOOD STUDY (VBS)

HRS Documentation Report

December 2025

Report prepared by:
Gokul Seshadri, University of Minnesota
Trey Smith, University of Michigan
Erik Klopack Indiana University,
Eileen Crimmins, University of Southern California
Bharat Thyagarajan, University of Minnesota
Jessica Faul, University of Michigan

Contents

Overview 2
Sample 2
Collection 2
Laboratory 2
Laboratory Procedures – Sample preparation and processing 2
Quality Control 3
Data Files 10
Log2cpm Calculation 10
If You Need to Know More 11
HRS Internet Site 11
Contact Information 11
Citing this Document 11
References 11

Overview

This data release includes raw counts and log2 counts-per-million (log2cpm) values from RNASeq analysis performed on a subsample of participants who consented to the HRS 2016 Venous Blood Study (n=3748).

Sample

RNASeq was performed on a representative subsample of HRS participants who participated in the 2016 Venous Blood Study. The sample includes all the participants of the 2016 Healthy Cognitive Aging Project (HCAP) who provided blood samples, younger participants designated for future HCAP assessments, and a subsample of HCAP non-participants. This subsample fully represents the entire HRS sample. The same sample that was selected for DNA methylation analysis was selected for RNA sequencing. HRS processed a total of 3748 unique RNA samples.

Collection

The blood collection was managed by Hooper Holmes Health & Wellness. Hooper Holmes, now ExamOne, was provided with the names, addresses, and phone numbers of consenting respondents and contacted respondents to set appointments. Collection materials were mailed to the phlebotomists' homes in advance of the scheduled visit. The phlebotomy service reported to HRS on respondents who declined to schedule appointments or missed scheduled appointments, and HRS staff followed-up up to detect any problems and to attempt to reschedule appointments. Every attempt was made to schedule the blood draw within 4 weeks of the HRS core interview. Fasting was recommended and preferred but not required. Phlebotomists noted the fasting status of the samples. We collected 50.5 mL of blood in 6 tubes – 1 8 mL CPT tube, 3 10 mL double gel serum separator tubes (SST), 1 10 mL EDTA whole blood tube, and a 2.5 mL PAXgene RNA tube. The Paxgene tubes were shipped at room temperature to the Advanced Research and Diagnostics Laboratory (ARDL) at the University of Minnesota. The Paxgene tubes were stored at -80°C until further analysis. More information on the blood collection procedures can be found here: https://hrsdata.isr.umich.edu/data-products/2016-venous-blood-study-vbs

Laboratory

All assays were performed at the University of Minnesota Advanced Research and Diagnostic Laboratory (ARDL) and the University of Minnesota Genomics Center (UMGC). The ADRL is CLIA certified laboratory established for improved coordination and centralization of laboratory activities of multi-center research studies contracted by the Department of Laboratory Medicine and Pathology at the University of Minnesota. ARDL served as the analytic laboratory for the HRS 2016 VBS. UMGC performed the RNA sequencing in this study.

Laboratory Procedures – Sample preparation and processing

RNA Extraction: RNA was extracted from whole blood stored in Paxgene tubes by using the Paxgene Blood miRNA Kit. Extracted RNA is then stored at -80°C until further analysis. Total RNA isolates were quantified using a fluorimetric RiboGreen assay.

Library Preparation: Total RNA samples were treated with the Globin-Zero Gold rRNA Removal Kit (Illumina Inc.) to deplete ribosomal RNA and globin prior to creating sequencing libraries using Illumina's stranded mRNA Sample Preparation kit (Cat. # RS-122-2101). One microgram of total RNA was oligo-dT

purified using oligo-dT coated magnetic beads, fragmented and then reverse transcribed into cDNA, fragmented, blunt-ended, and ligated to indexed (barcoded) adaptors and amplified using 15 cycles of PCR. Indexed libraries were then normalized, pooled and size selected to 320bp +/- 5% using Caliper's XT instrument.

RNA Sequencing: Samples were sequenced using 2*50 bp paired-end reads to a minimum of 20 million reads per sample using the Illumina NovaSeq 6000 at the University of Minnesota Genomics Center. All samples were processed through the HRS RNAseq QC analysis pipeline at the University of Minnesota, this is an extended version of the TopMed/GTEX analysis pipeline (https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md). The STAR aligner was used for alignment of the sequence reads to the GRCh38 human reference genome along with GENCODE 30 annotations. All quality control analyses were performed using an updated version of RNASeQC 2.3.4 and estimated quality control metrics to obtain the final data. The read counts from each sample were combined into a count file.

Normalization/Transformation: We used edgeR calcNormFactors() function and used RLE (relative log expression) normalization to account for compositional differences between the libraries. RLE is the scaling factor method; where the median library is calculated from the geometric mean of all columns and the median ratio of each sample to the median library is taken as the scale factor. We then used cpm() function in edgeR on the normalized DGEList object to estimate the log2 counts-per-million (log2cpm) with a prior.count=2, and the transformed counts were rounded to six decimal places.

Quality Control

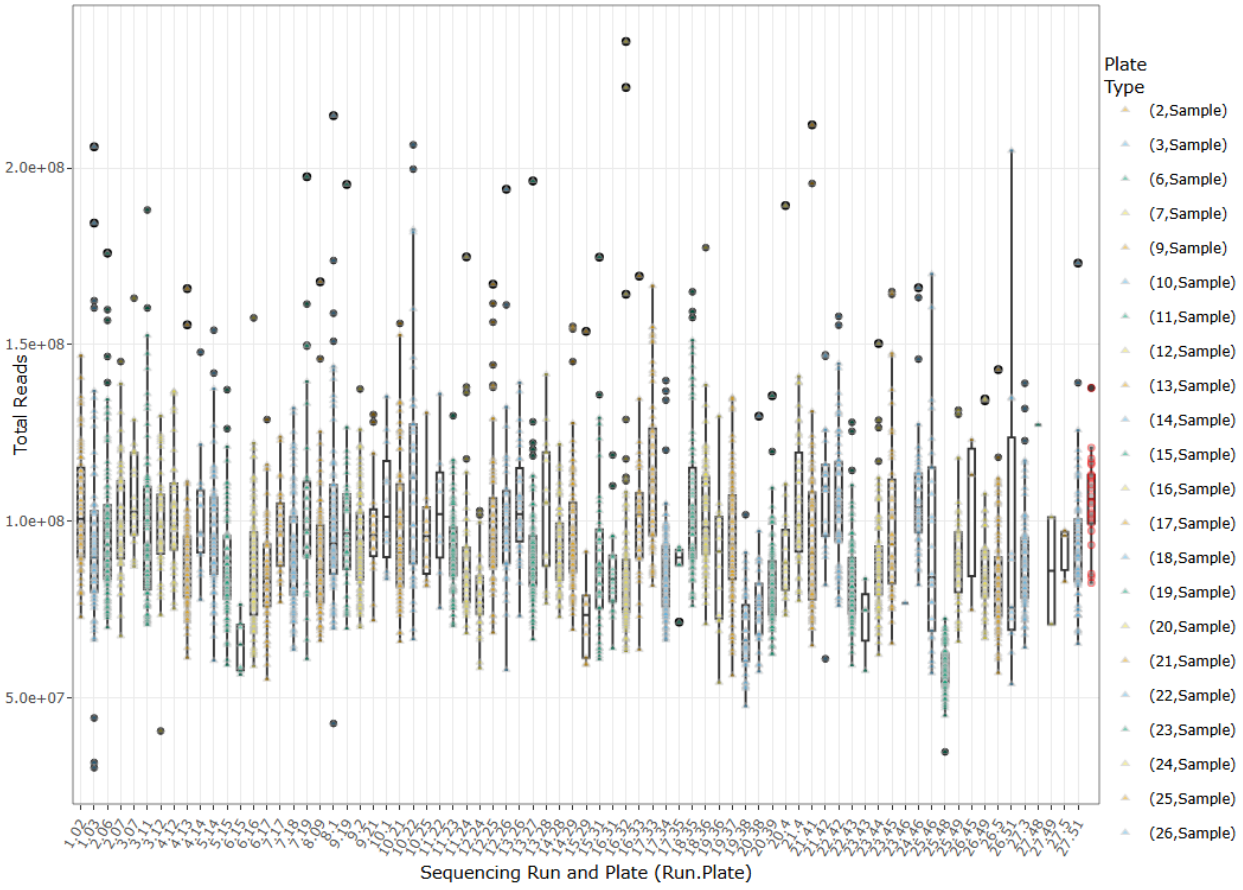
This section outlines the QC process we conducted, including summary plots and statistics, for the final set of HRS RNA Seq data. There are four different metrics that we analyzed,

1. Mapping Distributions
2. Intergenic Rate
3. Ribosomal RNA depletion
4. Expression profiling efficiency & estimated library complexity

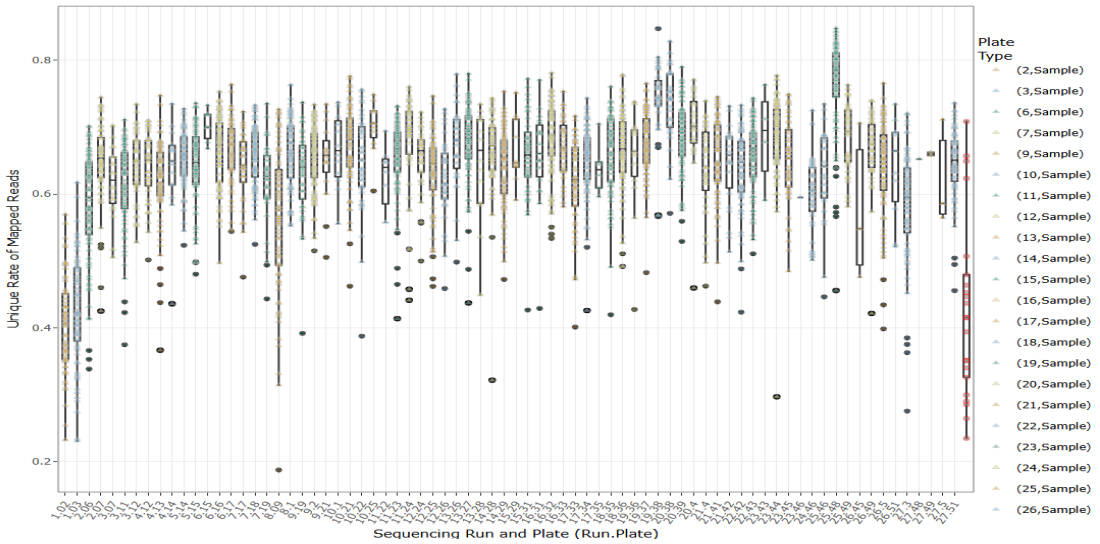
HRS processed a total of 3748 RNA-Seq libraries representing data for 3748 unique RNA samples.

1. Mapping Distributions

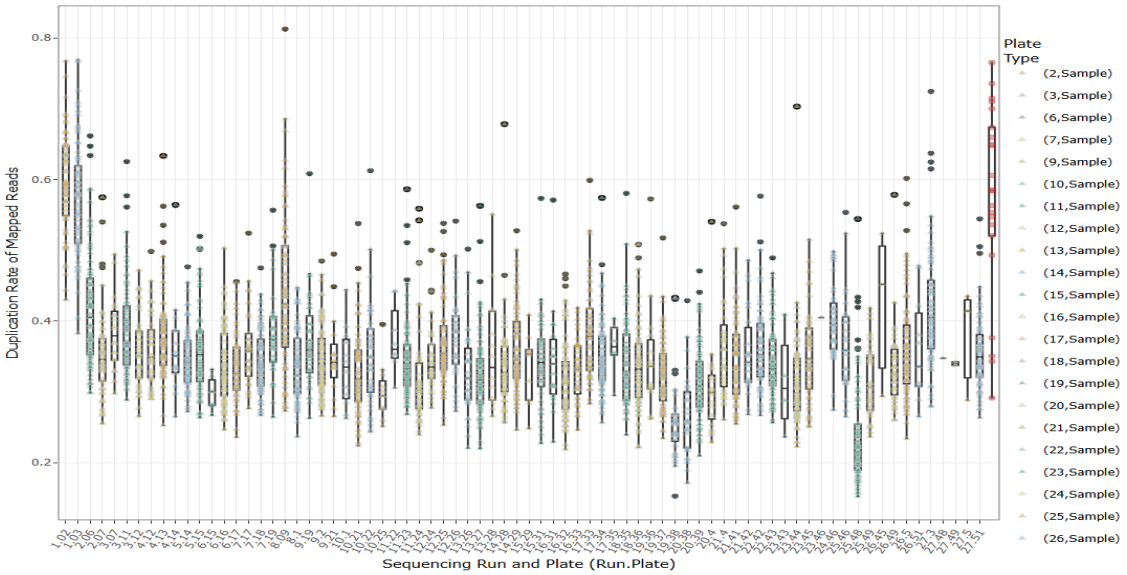
- **Total Reads**
  Total reads are the total number of reads aligned; this is not filtered in any way. The median read count across samples is 90M, with a range of 30M–236M reads. All samples have more than 20M reads.



- **Unique Rate of Mapped Reads**
  Unique rate of mapped reads is the fraction of uniquely mapped reads of all mapped reads. The median number of uniquely mapped reads is 35M, with a range of 7.5M to 85M.

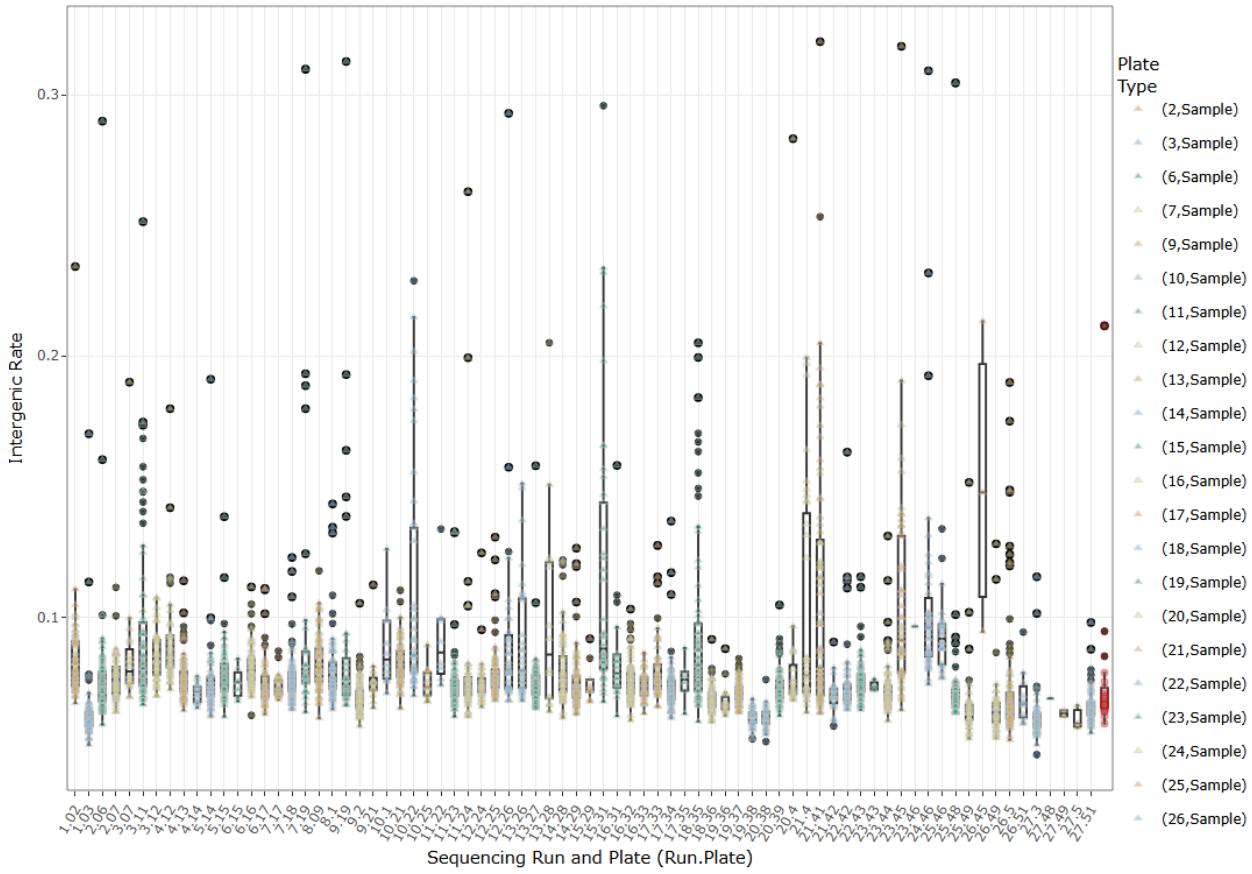


- **Duplication Rate of Mapped Reads**
  'Duplication Rate of Mapped Reads' is the fraction of non-uniquely mapped reads of all mapped reads, includes multi-mappers and optical duplicates). The median duplication rate is 34%, with a range from 15% to 81%.



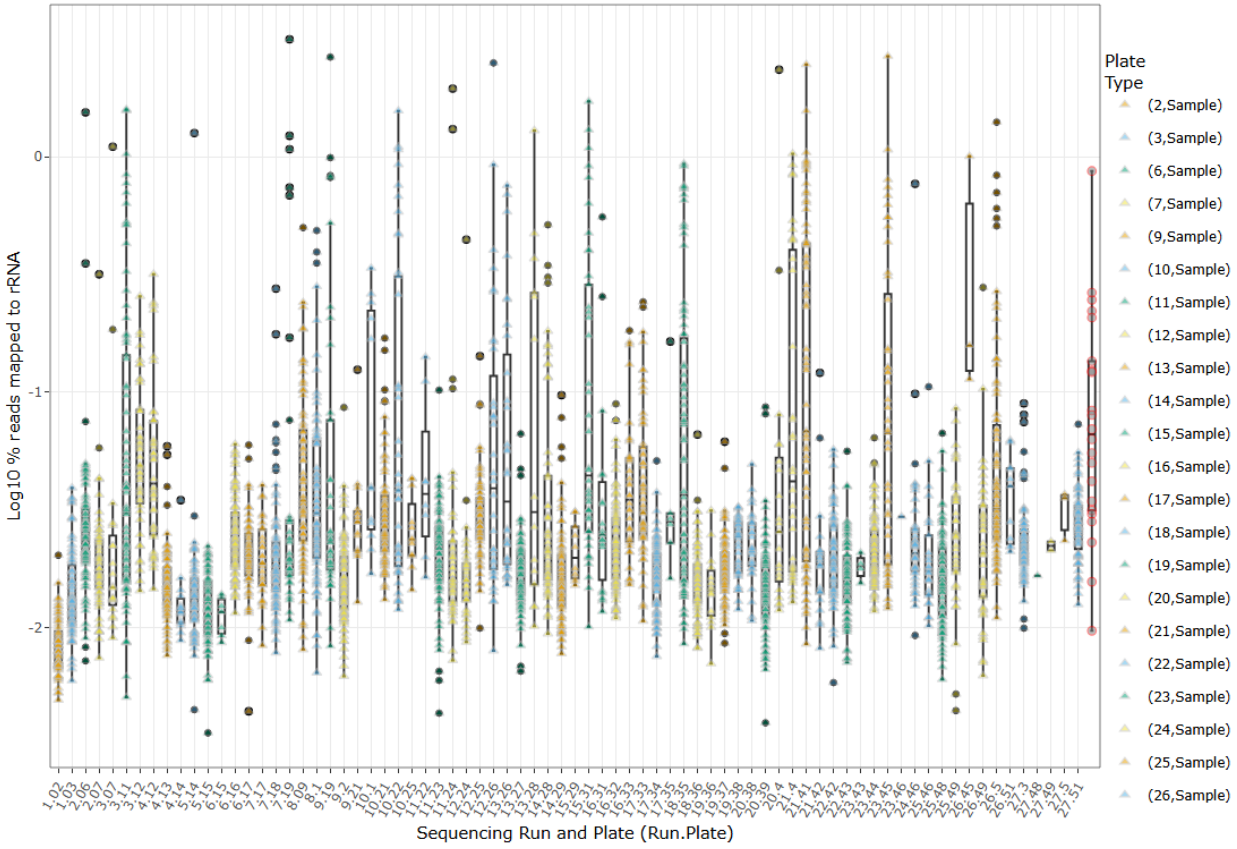
2. Intergenic Rate

During the QC analysis we found that 293 samples have an intergenic rate greater than 10%. These samples were not excluded but were flagged for possible DNA contamination.



3. Ribosomal RNA depletion

When looking at the percentage of reads mapped to ribosomal RNA (rRNA), we found that it was at a maximum of 3.14% and showed increased dispersion for certain plates. This indicates low level of ribosomal RNA contamination in these samples.



4. Expression profiling efficiency & estimated library complexity

During data generation, the University of Minnesota Genomics Center (UMGC) and the Advanced Research and Diagnostics Laboratory (ARDL) developed a novel metric to identify samples with poor quality RNASeq data.

This novel metric included simultaneously looking at both **Expression Profiling Efficiency** and **Estimated Library Complexity** to establish quality thresholds. Estimated Library Complexity is a metric that is created by the developers of Picard (and used in RNASeQC). This metric gave a better picture of quality than Total Reads or even Mapped Reads. The following is the description of the **Estimated Library Complexity** from the developers.

"Reads are sorted by the first N bases (5 by default) of the first read and then the first N bases of the second read of a pair. Read pairs are considered to be duplicates if they match each other with no gaps and an overall mismatch rate less than or equal to MAX_DIFF_RATE (0.03 by default). Reads of poor quality are filtered out to provide a more accurate estimate. The filtering removes reads with any poor quality bases as defined by a read's MIN_MEAN_QUALITY (20 is the default value) across either the first or second read. Unpaired reads are ignored in this computation.

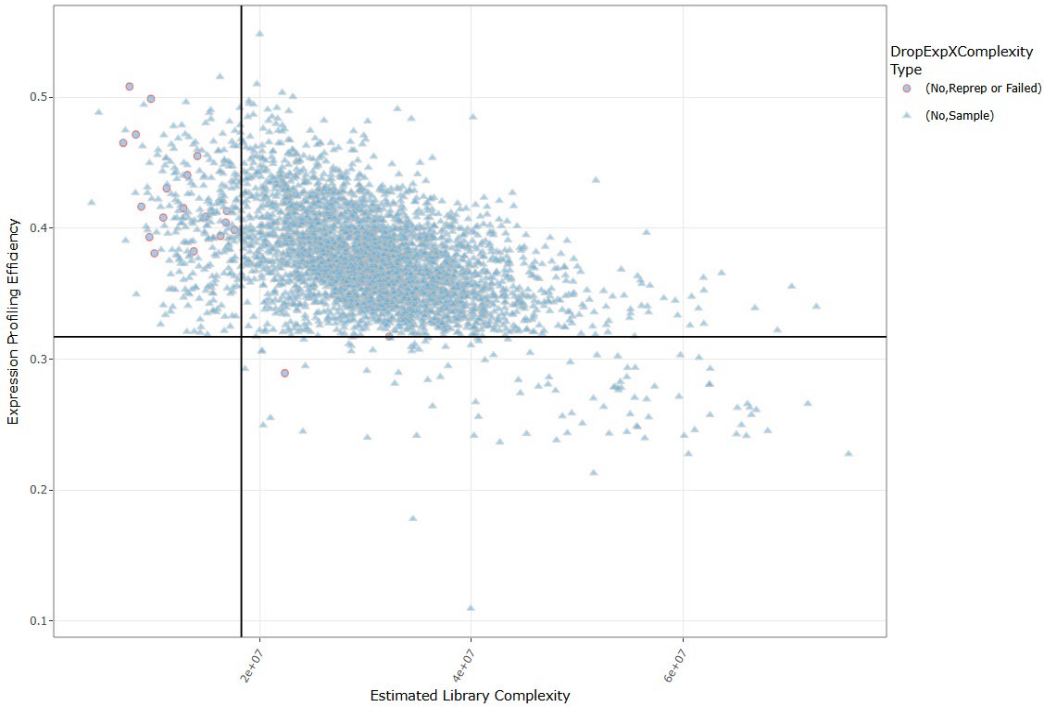
The algorithm attempts to detect optical duplicates separately from PCR duplicates and excludes these in the calculation of library size. Also, since there is no alignment information used in this algorithm, an additional filter is applied to the data as follows. After examining all reads, a histogram is built in which the number of reads in a duplicate set is compared with the number of of duplicate sets. All bins that contain exactly one duplicate set are then removed from the histogram as outliers prior to the library size estimation."

The **Expression profiling efficiency** is defined as the ratio of exon mapped reads to total mapped reads.

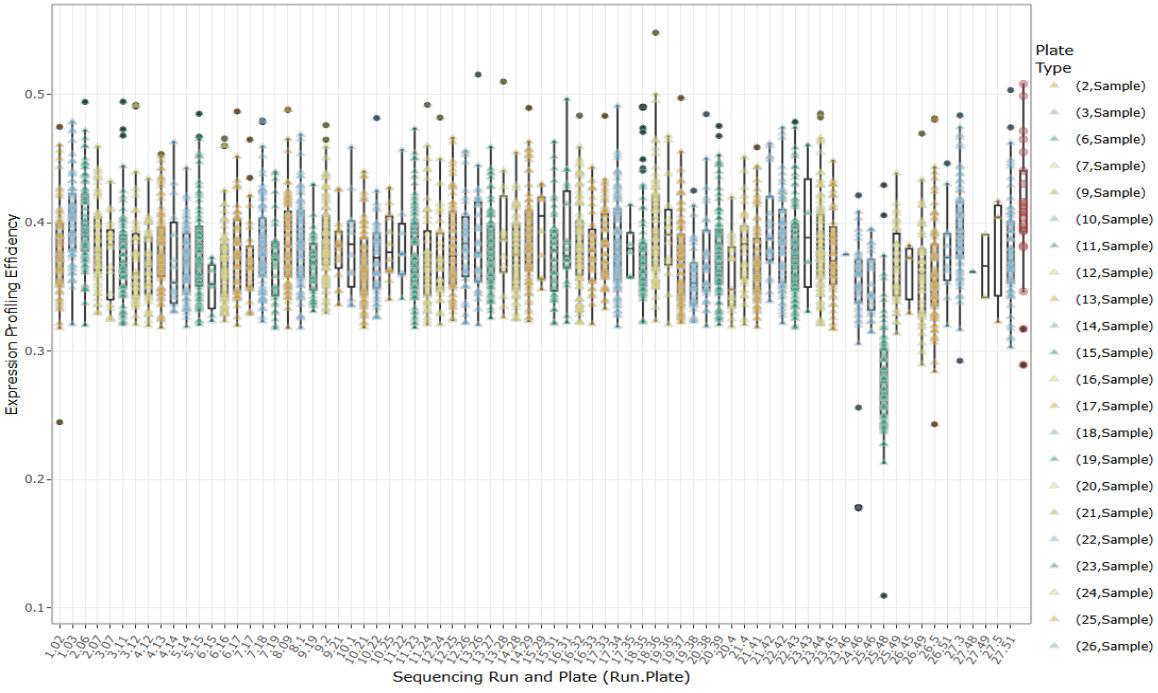
After plotting Expression Profiling Efficiency vs. Estimated Library Complexity, samples that were below the 10th percentile threshold for both metrices (0.317 for expression profiling efficiency and 18297436 for Estimated Library Complexity) were selected for repeat RNA extraction and sequencing. Applying these same thresholds to the final released sample set shows that no samples are below the 10^{th} percentile for both these metrices.

Samples below the 10^{th} percentile threshold for both metrices had new libraries prepared and resequenced to obtain new data that passed this QC metric. The final dataset had 22 samples that were resequenced to meet the minimum QC metrics and are represented in the red box blot (seq 27.51).

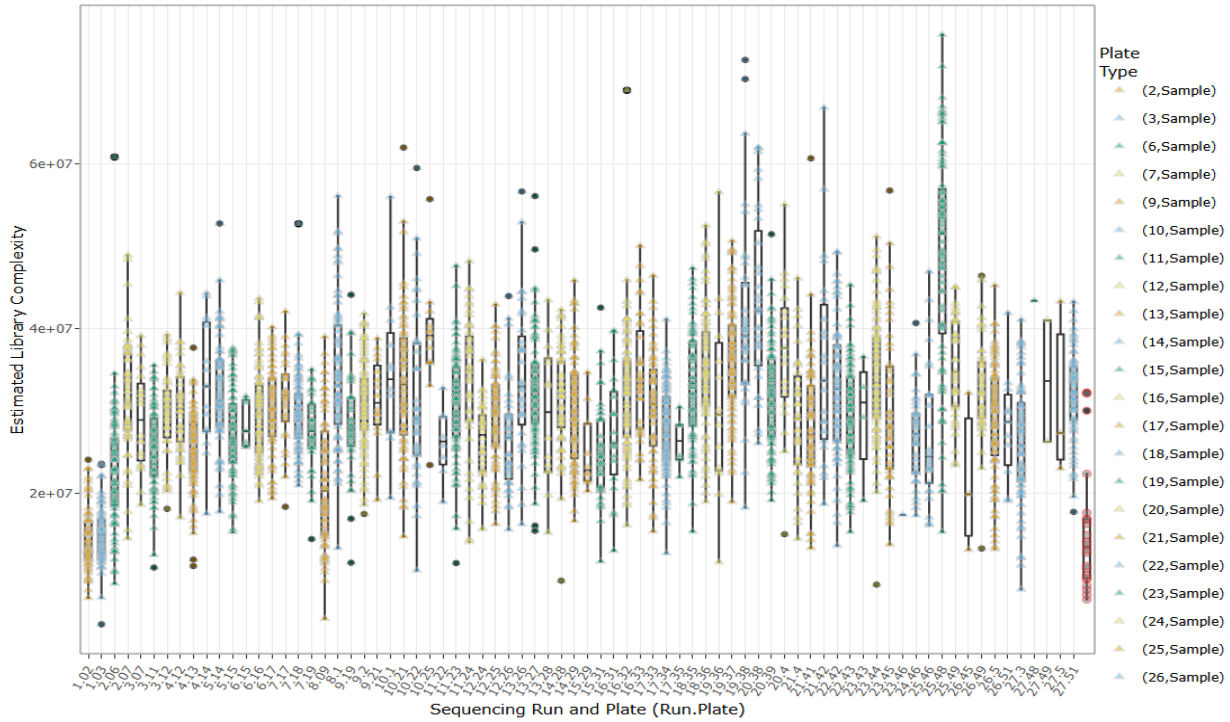
QC Plot



Expression Profiling Efficiency by Sequencing Run and Plate



Estimate Library Complexity by Sequencing Run and Plate



In summary, raw counts and log2cpm values for 3748 HRS participants who participated in the 2016 Venous Blood Study are included in this release. We did not do any additional filtration based on gene length and all genes were included in the final dataset.

Data Files

This release includes two .csv files: HRS_RNASeq_raw_releaseV1.csv and HRS_RNASeq_logcpm_releaseV1.csv. In each file, Column A contains the Gene IDs (in rows). HRS participant IDs (in columns) are represented by numeric HHIDs concatenated to PNs (e.g. HHID=012345 and PN=010 becomes 12345010). The cells contain the RNASeq raw counts or log2CPM normalized values.

Log2cpm Calculation

We processed RNASeq raw counts for 58,219 genes across 3,748 samples. To account for differences in sequencing depth and composition bias, we applied the **Relative Log Expression (RLE)** method to compute normalization factors for each sample. The data was then converted to **counts per million (CPM)** to adjust for library size and log2-transformed with a **pseudo-count of 2** for numerical stability. Finally, the transformed values were rounded to **six decimal places**, resulting in the **log2CPM normalized RNA-seq dataset**.

The log2CPM normalization is done using **edgeR** version **4.0.16**. We first converted raw counts into a **DGEList** object and then applied RLE normalization method using **calcNormFactors** function before passing them into **cpm** function for log2CPM transformation.

If You Need to Know More

This document is intended to serve as a brief overview to the RNASeq count data product. If you have questions or concerns that are not adequately covered here or on our Web site, or if you have any comments, please contact us. We will do our best to provide answers.

HRS Internet Site

Health and Retirement Study public release data and additional information about the study are available on the Internet. To access public data or to find out more about restricted data products and procedures, visit the HRS Web site.

Contact Information

If you need to contact us, you may do so by one of the methods listed below.

Internet: Help Desk at the HRS Web site (http://hrsonline.isr.umich.edu)

E-mail: hrsquestions@umich.edu

Postal Service:
Health and Retirement Study
The Institute for Social Research
426 Thompson Street
Ann Arbor, Michigan 48104

Citing this Document

Please include the following citation in any research reports, papers, or publications based on these data along with the citation for the reference epigenetic clock:

In text: "The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (NIA U01AG009740) and is conducted by the University of Michigan."

In references: "Seshadri G, Smith T, Klopack E, Crimmins EM, Thyagarajan BT, Faul JD. RNASeq Count Data from the 2016 Health and Retirement Study Venous Blood Study (VBS) – Release 1. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan; 2025."

References

Crimmins, E., Faul, J., Thyagarajan, B., & Weir, D. (2017). Venous blood collection and assay protocol in the 2016 Health and Retirement Study 2016 Venous Blood Study (VBS). In (pp. 1-73). Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.