**Assessing the capture efficiency and transcript representation fidelity of three RNA-Seq library preparation options using low-input *M. Musculus* samples**

**VIB Nucleomics**

Matthew Watson December 30, 2020

## Overview and Purpose

This project aims to assess the performance of three commercially available RNA-seq preparations using samples from the common mouse *Mus musculus*. Libraries that were prepared from each kit from low-input starting material were sequenced on the Illumina NextSeq500 platform using single read technology and assessed for basic fidelity using the following metrics and comparative statistics:

- Library read quality
- Alignment sequence quality
- RNA feature coverage (individual)
- Cumulative RNA feature coverage (total counts per feature)
- Relative gene expression levels and expression consistency
- Representative abundances of relevant transcripts
- Proportions of relevant transcripts that should be captured by the specific kit chemistry

Using a combination of the aforementioned metrics, an assessment will be made as to the best kit option for a particular type of RNA-seq/transcriptomic experiment, in order to inform future project design for obtaining the necessary high-quality RNA-seq data to answer a particular research question. For example, these findings may help to inform the best RNA prep option for future low input samples, as well as suggest a method of preparation that can optimize the retention of certain RNA species for both total RNA as well as targeted/mRNA pipelines.

## Sample List and Experimental Design

A total of 18 unique libraries were sequenced across 4 lanes on the NextSeq instrument. The sample designations and their corresponding preparations are as follows:

| Sample | Kit | Sequencing Cycles (bp) |
|---|---|---|
| 26-AR56a_S1 | Kit A | 151 |
| 28-AR61a_S2 | Kit A | 151 |
| 31-AR68a_S3 | Kit A | 151 |
| 36-ARE36_S4 | Kit A | 151 |
| 41-ARE41_S5 | Kit A | 151 |
| 44-ARE44_S6 | Kit A | 151 |
| AR56a_S17 | Kit B | 75 |
| AR61a_S19 | Kit B | 75 |
| AR68a_S22 | Kit B | 75 |
| ARE36_S27 | Kit B | 75 |
| ARE41_S32 | Kit B | 75 |
| ARE44_S35 | Kit B | 75 |
| AR56a_HT_S36 | Kit C | 75 |
| AR61a_HT_S37 | Kit C | 75 |
| AR68a_HT_S38 | Kit C | 75 |
| ARE36_HT_S39 | Kit C | 75 |
| ARE41_HT_S40 | Kit C | 75 |
| ARE44_HT_S41 | Kit C | 75 |

**Anonymous designation of kits prior to analysis**

For the experimental design stated above, the identity of each kit was intentionally kept anonymous prior to complete data analysis and quality control. One of the motivations for this particular project was to assess the various library control metrics outlined in the overview, and to be able to assign a particular library preparation kit to a set of samples given its transcriptomic profile and RNA capture patterns. The intention here is to analyze a combination of RNA annotation coverage as well as key performance indicators for library quality control, in order to successfully assign a kit description to each of the 3 kit possibilities. The three options for RNA library prep are described below.

**Data completeness and Distribution for FASTQ files**

It should be noted that for 6 of the 18 unique samples in the dataset, all of which correspond to type Kit A (26-AR56a_S1 to 44-ARE44_S6 in order of appearance in the table above), sequencing was conducted using paired end technology with Illumina 151-8-8-151 chemistry. The second read FASTQ files that are paired with the Read 1 FASTQ files used in this project were not available for analysis, and as such, these samples were treated as single end reads. Furthermore, the remaining 12 unique samples in the dataset were sequenced across two separate Illumina NextSeq sequencing runs.

**Descriptions of Kit Options/Preparations**

1. *Takara SMARTer Stranded Total RNA*- A total RNA option that aims to capture both protein coding and regulatory/non-coding RNA transcripts with an rRNA bead-based depletion kit. Library generation is achieved the standard way through cDNA synthesis and PCR amplification. Optimized for higher inputs of 10-100 ng of total RNA. Retains transcript strandedness to resolve transcript orientation from the genome.
2. *Takara SMARTSeq v4 mRNA Ultra Low Input*- Non-stranded mRNA option that uses oligo(dt) priming at the 5' end of the mRNA transcript, followed by traditional cDNA generation and amplification to retain coding transcripts. Optimized for low input RNA samples of 0 pg–1 ng total RNA.
3. *Takara SMARTSeq v4 mRNA High Throughput (HT)*- An automation-compatible version of the SMARTSeq v4 mRNA Ultra Low Input kit as described above. Optimized to include a one-step RT PCR for cDNA generation that varies from the normal throughput version of the kit. Input amounts are consistent with the previous kit description.

**Base-calling Quality Control**

FastQC v0.11.9 and MultiQC v. 1.9 were used to compile and assess the basic statistics of each library prior to alignment. Primary library statistics were compiled using FastQC, and visualization of primary statistics for all libraries in the dataset was achieved using MultiQC. Below is a summary table of the basic FASTQ file statistics that are compiled by FastQC. It is important to note that for this experiment, each sample was run across 4 different sequencing lanes for a total of 72 individual FASTQ files. After alignment, each alignment file was merged across the 4 lanes for a total of 18 merged BAM files for annotation (See Alignment). Libraries here are reported individually in order to identify any lane-specific trends in sequencing quality.

**Basic Statistics Summary Table**

| Sample Name | % Dups | % GC | Length | M Seqs |
|---|---|---|---|---|
| **26-AR56a_S1_L001_R1_001** | 45.3% | 58% | 115 bp | 4.7 |
| **26-AR56a_S1_L002_R1_001** | 45.9% | 58% | 115 bp | 4.6 |

| | | | | |
|---|---|---|---|---|
| **26-AR56a_S1_L003_R1_001** | 46.0% | 58% | 115 bp | 4.8 |
| **26-AR56a_S1_L004_R1_001** | 45.6% | 58% | 115 bp | 4.6 |
| **28-AR61a_S2_L001_R1_001** | 50.6% | 61% | 116 bp | 4.8 |
| **28-AR61a_S2_L002_R1_001** | 51.9% | 61% | 117 bp | 4.8 |
| **28-AR61a_S2_L003_R1_001** | 51.9% | 61% | 117 bp | 4.9 |
| **28-AR61a_S2_L004_R1_001** | 51.4% | 61% | 117 bp | 4.6 |
| **31-AR68a_S3_L001_R1_001** | 51.9% | 61% | 118 bp | 4.5 |
| **31-AR68a_S3_L002_R1_001** | 52.4% | 61% | 118 bp | 4.5 |
| **31-AR68a_S3_L003_R1_001** | 52.5% | 61% | 118 bp | 4.6 |
| **31-AR68a_S3_L004_R1_001** | 52.0% | 61% | 118 bp | 4.4 |
| **36-ARE36_S4_L001_R1_001** | 47.9% | 59% | 120 bp | 4.9 |
| **36-ARE36_S4_L002_R1_001** | 48.7% | 59% | 120 bp | 4.8 |
| **36-ARE36_S4_L003_R1_001** | 48.8% | 59% | 120 bp | 5.0 |
| **36-ARE36_S4_L004_R1_001** | 48.2% | 59% | 120 bp | 4.7 |
| **41-ARE41_S5_L001_R1_001** | 50.8% | 61% | 119 bp | 4.8 |
| **41-ARE41_S5_L002_R1_001** | 51.7% | 60% | 119 bp | 4.7 |
| **41-ARE41_S5_L003_R1_001** | 51.5% | 61% | 119 bp | 4.8 |
| **41-ARE41_S5_L004_R1_001** | 51.1% | 61% | 119 bp | 4.6 |
| **44-ARE44_S6_L001_R1_001** | 50.0% | 61% | 119 bp | 4.8 |
| **44-ARE44_S6_L002_R1_001** | 50.7% | 60% | 119 bp | 4.7 |
| **44-ARE44_S6_L003_R1_001** | 50.7% | 61% | 119 bp | 4.8 |

| | | | | |
|---|---|---|---|---|
| **44-ARE44_S6_L004_R1_001** | 50.2% | 60% | 119 bp | 4.6 |
| **AR56a_HT_S36_L001_R1_001** | 32.8% | 50% | 76 bp | 2.1 |
| **AR56a_HT_S36_L002_R1_001** | 32.7% | 50% | 76 bp | 2.1 |
| **AR56a_HT_S36_L003_R1_001** | 32.6% | 50% | 76 bp | 2.1 |
| **AR56a_HT_S36_L004_R1_001** | 32.3% | 50% | 76 bp | 2.1 |
| **AR56a_S17_L001_R1_001** | 32.4% | 51% | 76 bp | 2.5 |
| **AR56a_S17_L002_R1_001** | 32.5% | 50% | 76 bp | 2.6 |
| **AR56a_S17_L003_R1_001** | 32.2% | 51% | 76 bp | 2.5 |
| **AR56a_S17_L004_R1_001** | 32.0% | 51% | 76 bp | 2.5 |
| **AR61a_HT_S37_L001_R1_001** | 32.5% | 50% | 76 bp | 2.3 |
| **AR61a_HT_S37_L002_R1_001** | 32.6% | 50% | 76 bp | 2.3 |
| **AR61a_HT_S37_L003_R1_001** | 32.4% | 50% | 76 bp | 2.3 |
| **AR61a_HT_S37_L004_R1_001** | 32.2% | 50% | 76 bp | 2.3 |
| **AR61a_S19_L001_R1_001** | 32.7% | 51% | 76 bp | 2.7 |
| **AR61a_S19_L002_R1_001** | 32.5% | 51% | 76 bp | 2.8 |
| **AR61a_S19_L003_R1_001** | 32.4% | 51% | 76 bp | 2.7 |
| **AR61a_S19_L004_R1_001** | 32.0% | 51% | 76 bp | 2.7 |
| **AR68a_HT_S38_L001_R1_001** | 29.5% | 50% | 76 bp | 2.2 |
| **AR68a_HT_S38_L002_R1_001** | 29.6% | 50% | 76 bp | 2.2 |
| **AR68a_HT_S38_L003_R1_001** | 29.2% | 50% | 76 bp | 2.2 |
| **AR68a_HT_S38_L004_R1_001** | 29.1% | 50% | 76 bp | 2.2 |

| | | | | |
|---|---|---|---|---|
| **AR68a_S22_L001_R1_001** | 28.3% | 50% | 76 bp | 2.3 |
| **AR68a_S22_L002_R1_001** | 28.2% | 50% | 76 bp | 2.4 |
| **AR68a_S22_L003_R1_001** | 27.9% | 50% | 76 bp | 2.3 |
| **AR68a_S22_L004_R1_001** | 27.7% | 50% | 76 bp | 2.3 |
| **ARE36_HT_S39_L001_R1_001** | 31.4% | 50% | 76 bp | 2.1 |
| **ARE36_HT_S39_L002_R1_001** | 31.5% | 50% | 76 bp | 2.1 |
| **ARE36_HT_S39_L003_R1_001** | 31.0% | 50% | 76 bp | 2.1 |
| **ARE36_HT_S39_L004_R1_001** | 30.9% | 50% | 76 bp | 2.1 |
| **ARE36_S27_L001_R1_001** | 28.1% | 50% | 76 bp | 2.1 |
| **ARE36_S27_L002_R1_001** | 28.0% | 50% | 76 bp | 2.2 |
| **ARE36_S27_L003_R1_001** | 27.9% | 50% | 76 bp | 2.1 |
| **ARE36_S27_L004_R1_001** | 27.8% | 50% | 76 bp | 2.1 |
| **ARE41_HT_S40_L001_R1_001** | 31.6% | 50% | 76 bp | 2.0 |
| **ARE41_HT_S40_L002_R1_001** | 31.6% | 50% | 76 bp | 2.0 |
| **ARE41_HT_S40_L003_R1_001** | 31.4% | 50% | 76 bp | 2.0 |
| **ARE41_HT_S40_L004_R1_001** | 31.0% | 50% | 76 bp | 2.0 |
| **ARE41_S32_L001_R1_001** | 31.3% | 50% | 76 bp | 2.7 |
| **ARE41_S32_L002_R1_001** | 31.3% | 50% | 76 bp | 2.7 |
| **ARE41_S32_L003_R1_001** | 31.0% | 50% | 76 bp | 2.7 |
| **ARE41_S32_L004_R1_001** | 30.8% | 50% | 76 bp | 2.6 |
| **ARE44_HT_S41_L001_R1_001** | 32.2% | 50% | 76 bp | 2.1 |

| | | | | |
|---|---|---|---|---|
| **ARE44_HT_S41_L002_R1_001** | 32.0% | 50% | 76 bp | 2.1 |
| **ARE44_HT_S41_L003_R1_001** | 31.9% | 50% | 76 bp | 2.1 |
| **ARE44_HT_S41_L004_R1_001** | 31.9% | 50% | 76 bp | 2.1 |
| **ARE44_S35_L001_R1_001** | 30.8% | 50% | 76 bp | 2.7 |
| **ARE44_S35_L002_R1_001** | 31.0% | 50% | 76 bp | 2.7 |
| **ARE44_S35_L003_R1_001** | 30.7% | 50% | 76 bp | 2.7 |
| **ARE44_S35_L004_R1_001** | 30.5% | 50% | 76 bp | 2.7 |

***Explanation of table metrics***
*% Dups*- The percentage of reads that are determined to be duplicated by having the identical read sequence to another read in the same FASTQ file.
*% GC*- The percentage of all nucleotides within all of the reads in the FASTQ that are either a G or C. GC content can often cause analytical bias in downstream analysis tools for RNA-seq, so its calculation can be important for subsequent filtering and normalization (See Library GC Content).
*Length*- the average length of the reads of the library after sequencing. Note that due to read quality scores and sequencing chemistry, the average read length may be shorter than the number of cycles provided by an Illumina sequencing kit, and may differ by library.
*M seqs*- The number, in millions, of raw reads in the FASTQ.

This FASTQ summary table provides a good starting set of basic statistics that can be used to assess the preliminary quality of an NGS library. Any outlying or unpredictable values at this quality control step often need to be addressed through downstream filtering steps or troubleshooting of the assay where applicable. It should be noted that there is a high level of variability in the number of duplicated reads among the kit preparations. The ranges for read duplication rates are as follows:

Kit A: 45-52.5% Kit B: 27.9-32.7% Kit C: 29.1-32.8%

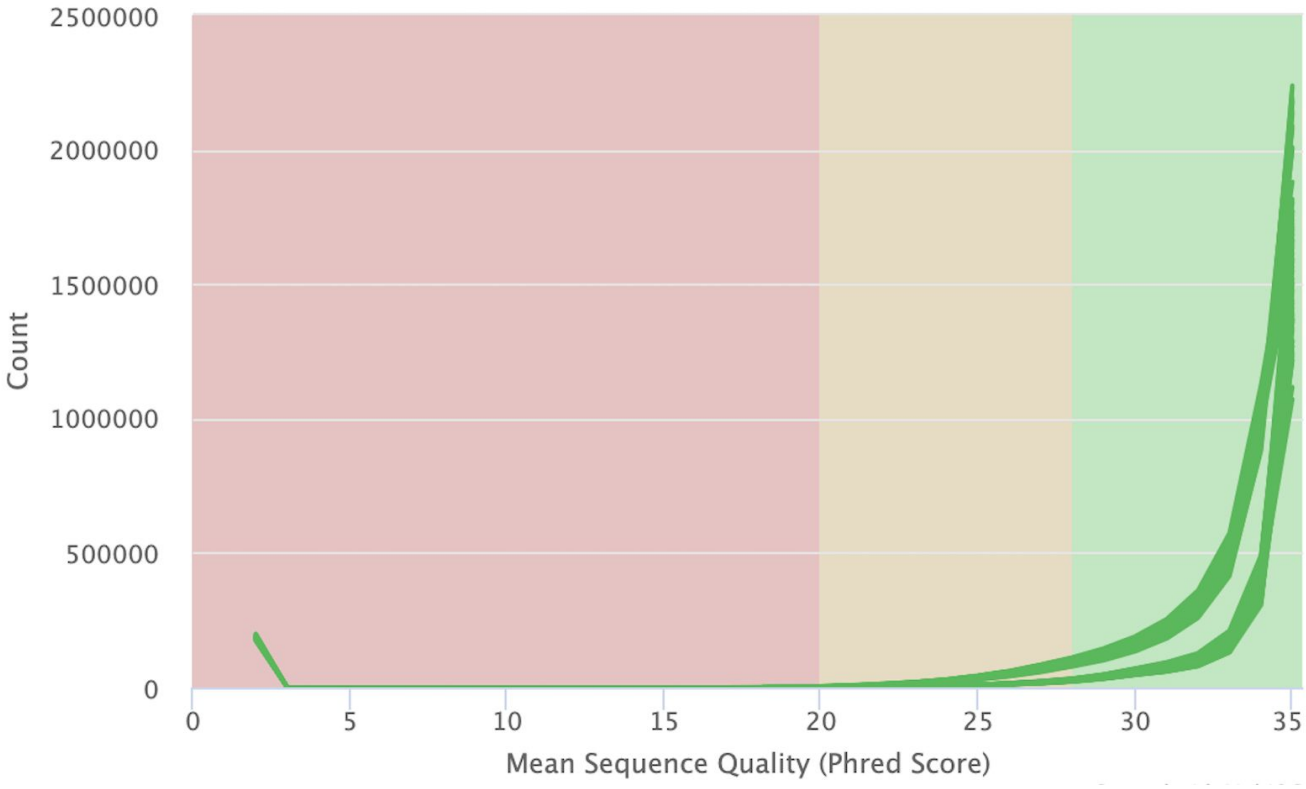
Duplicated reads can often confound RNA-seq analysis pipelines because it can be difficult to assess whether the duplicate read is a true biological transcript, or simply a by-product of the library preparation process. It is common for NGS libraries prepared from RNA to have a high level of over-represented and duplicated sequences, so observing a duplication rate of 45-52% of reads prepared using Kit A is not unusual, especially for libraries created from low-input material. It can be seen that the duplication rate for libraries prepared using either Kit B or C are much lower, ranging from approximately 28-33%; this is likely due to the specific vendor instructions for the preparation process as well as the specific kit chemistry. During the library prep process, it is possible for duplicate reads to be generated as artefacts from imperfect wet-lab procedures, including such steps as random fragmentation and priming during strand synthesis. Furthermore, with low-input libraries, it is common practice to increase the number of PCR cycles in order to obtain a sufficient final library concentration that can undergo bridge amplification and loading onto an Illumina flow cell, which may cause preferential amplification of certain sequences leading to read composition bias or “optical” PCR duplicate reads.

In certain analyses, duplicated reads are retained in order to identify their origin, but some analyses and workflows call for the marking of duplicate reads prior to alignment using a tool such as Picard.

FastQC was also used to assess the base-calling quality scores of raw FASTQ files. The following plot shows the distribution quality of the base-calling scores based on Phred quality scores (Q), which is calculated as the following:

$$Q = -10\log(P)$$ where P is the probability of a base being incorrectly called

A typical threshold of 30 is set to the Phred scores for retention of reads prior to downstream analyses. This value represents a base-calling error rate P of 1 in 1000 (0.001), which is considered to be the standard threshold for NGS workflows. With a successful Illumina sequencing run we would expect to see a very high proportion of reads that have an average Phred score of 30 or greater, based on the consistency and robustness of the sequencing chemistry.

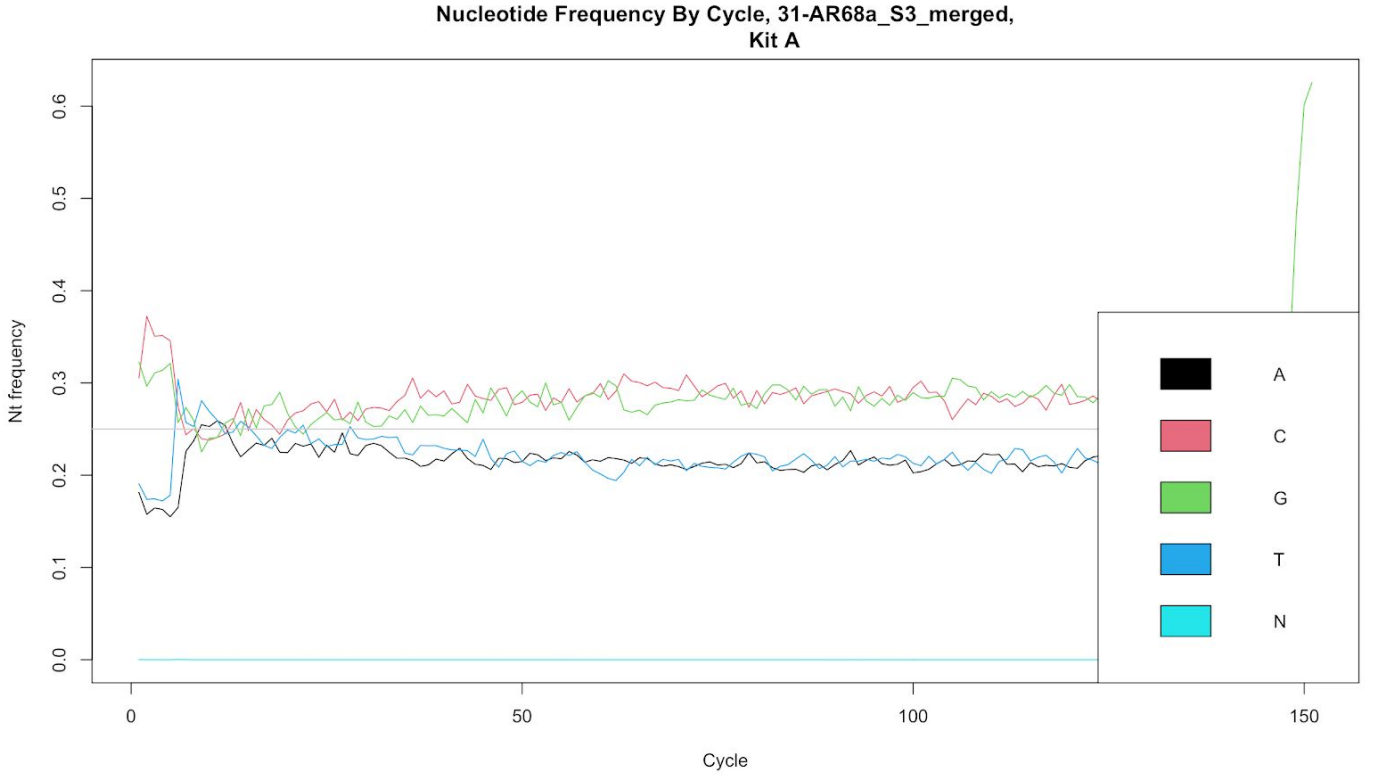


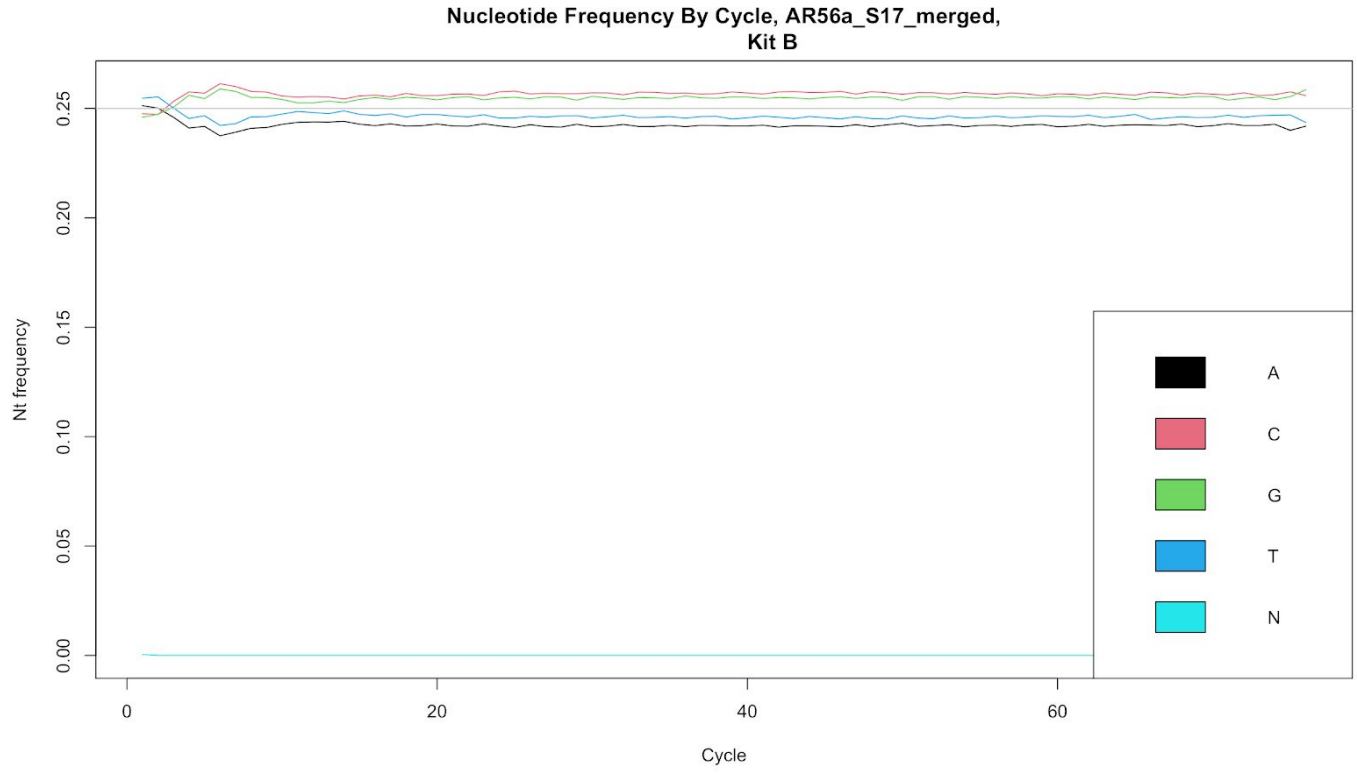

Without differentiating by sample, it is clear that all FASTQ files have predominantly high-quality read sequences, with the majority of read sequences having an average base Phred score of 30 or greater. This is reflected in the distribution to the right side of the graph, where there are visible, unimodal peaks between 30 and 35. A very small portion of read sequences will have a very poor average Phred score of 5 or less, visible in the bottom left portion of the graph; these read sequences will be filtered in the subsequent steps before alignment.

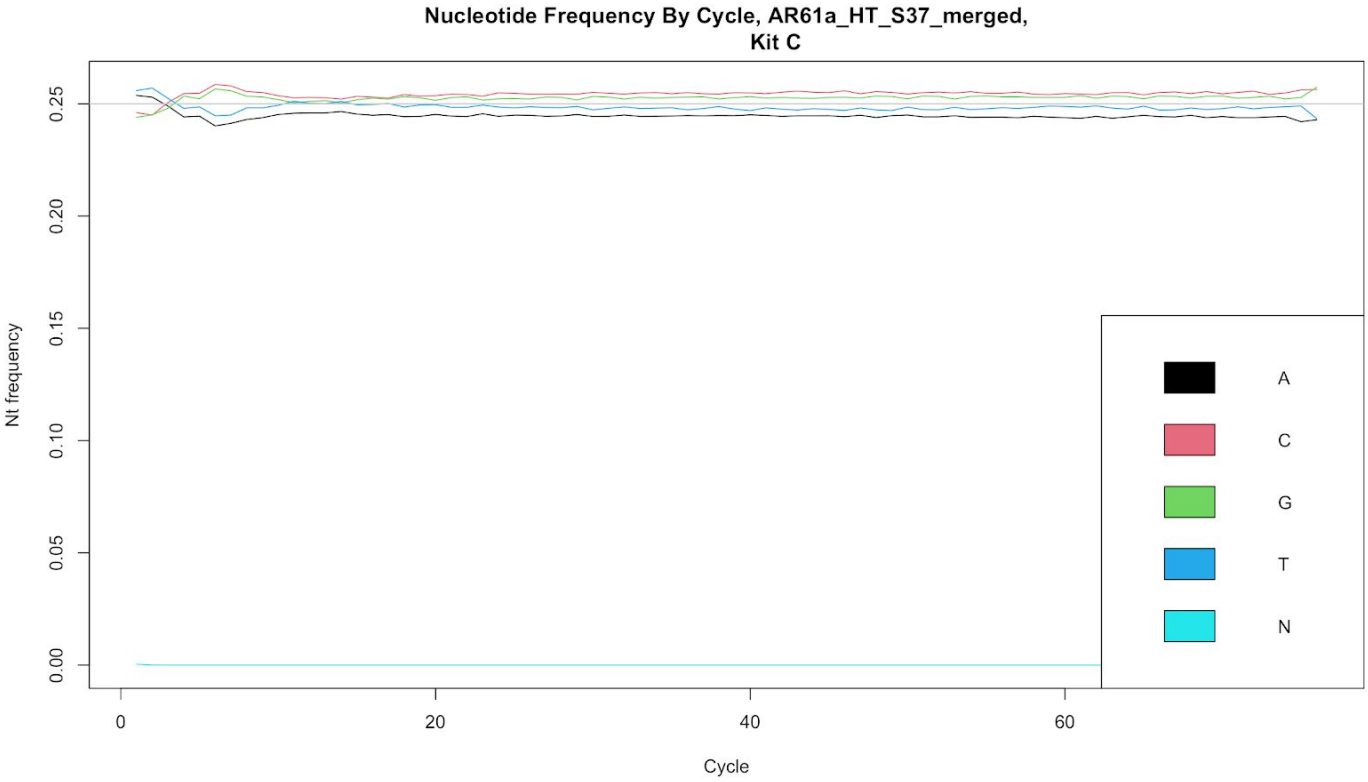
**Library GC Content**

The relative GC content of a library is believed to contribute to composition and counting bias during gene expression analysis. Therefore, it is important to ascertain and normalize for the relative GC content of a library during comparative analyses. From the FastQC reports and basic statistics summary table it can be seen that libraries from Kit A have a noticeable higher GC content (between 58-61% of nucleotides), whereas the range of GC nucleotides for SMARTSeq and SMARTSeq HT were either 50% or 51%. Using the EDAseq v. 2.22 package from Bio-conductor, this discrepancy in GC content can be viewed below. Since libraries prepared using Kit A are seen to have a much higher GC composition as compared to either SMARTSeq or SMARTSeq HT, normalization for gene expression levels will need to include a normalization step for GC content.
In addition to cataloging the mean GC content for each library (See Basic Statistics Summary Table), it is also beneficial to visualize the trend of nucleotide distributions along each of the cycles of Illumina sequencing, in

order to identify the characteristic nucleotide biases that occur both at the beginning and end of an Illumina sequencing run. An example of the GC composition for a library from each kit is shown below. Each library shown is highly representative of the other libraries from the same kit preparation, so for the purpose of keeping the report succinct, only one library from each kit preparation is shown. With these plots, we would expect to see a very even distribution of nucleotides for the majority of the library cycles, with the distributions for each nucleotide not fluctuating much above 0.25. For libraries with a noticeable GC bias, we observe a greater percentage of G and C nucleotides per cycle across the entire read length. This is most notable for Kit A preparations, as shown below i.e. we observe a greater gap between the G and C frequencies than both C and T.






A characteristic trend observed across all 18 libraries is the subtle presence of biased nucleotide compositions in the first several cycles of the sequencing run, starting from the left side of the figure. This can be seen for all kits as there is noticeable dispersion among the 4 nucleotide frequencies for approximately cycles 1-7. These artefacts are typical of Illumina sequencing; often the instrument takes the first few sequencing cycles to properly synchronize all of the flow cell clusters before beginning to call nucleotide bases accurately. As a result the nucleotides in these first few cycles often have composition bias and are usually of lower sequencing quality as compared to the middle bases of the read. This composition bias will be addressed in both the processing filtering steps at the read level, as well as downstream normalization procedures for nucleotide composition bias.

**Read Preprocessing and Filtering prior to Alignment**

Raw FASTQ files are often pre-processed before the alignment step in order to optimize the number of read sequences that align to the reference genome, as well as remove any technical sequencing artefacts or contaminating read sequences from the library. The following filtering and processing steps were completed for all FASTQ files prior to alignment:

***- Removal of adapter sequences-*** In Illumina sequencing, there is the possibility of having portions of the multiplexing adapter present in the raw read sequence. This often occurs when the read length is shorter than the number of sequencing cycles for the sequencing run, leading to portions of the adapter at the 3' end of the read. These sequences are not biologically relevant and can prevent the read from being properly aligned, so reads are trimmed of any adapter sequences. Cutadapt v. 1.18 was used to perform this operation with the following parameters:

--minimum-length 35 -O 10 -a *adapter_sequence*,
with -a providing the option for the specific Illumina adapter for each library

Using these parameters only reads that are longer than 35 bp after trimming are retained.

***- trimming of poly-A-tails-*** In certain RNA-seq library preparations, the capture of mature mRNA transcripts is achieved by enriching for transcripts that have a poly-A-tail at the 3' end of the transcript; these poly-A-tails are considered to be post-transcriptional modifications that do not appear in the original genomic sequence. These poly-A portions of read sequences are often removed to increase alignment scores to the reference genome. Cutadapt v. 1.18 was used with the following parameters:

-a 'TTTTTTTTTTTTTTT' -a 'AAAAAAAAAAAAAAA' -n 20 -m 35 -O 10

Using these parameters only reads that are longer than 35 bp after trimming are retained.

***- filtering using Phred quality scores-*** reads with lower quality scores are often removed in order to increase the confidence that reads are aligning uniquely to the genome. Using Cutadapt v. 1.18, low quality read sequences with Phred scores lower than 20 (see above) are removed using the following parameters:

-q 20 --minimum-length 35

Using these parameters only reads that are longer than 35 bp after trimming are retained.

**Summary Statistics from Read pre-processing and filtering**

The following table summarizes the effects on the raw FASTQ files for each kit after the preprocessing and filtering steps are applied. Values are presented as the percentage of total reads that are affected or modified by the particular filtering step, except for filtering based on read length which is expressed as the percentage of total bases with CutAdapt. Note that the values presented for the final column *% Reads Retained* signify the percentage of reads after all preprocessing steps and filtering have been completed, and indicate the percentage of the initial raw reads that will be carried to alignment. It is also very important to note that these values are not cumulative for the final retained read percentage. For example, a read could be trimmed based on a presence of contaminating adapter sequence or a poly-A tail, and subsequently be removed because the trimming step produced a final read that is shorter than 35 bp. For these reasons, there can be overlap in the percentages of filtering and processing steps that contribute to the final percentage of retained reads as shown in the table below.

| Kit Identifier | Adapter trimming (% of reads) | Poly-a-tail trimming (% of reads) | Filtering on quality score (% of reads) | Filtering on read length (% of bases) | % Reads Retained |
|---|---|---|---|---|---|
| Kit A | 0.4 ± 0.1 | 0.1 ± 0.05 | 2 ± 0.2 | 4 ± 0.3 | 96 ± 0.5 |
| Kit B | 0.1 ± 0.1 | 2.2 ± 0.1 | 0.4 ± 0.02 | 1.8 ± 0.1 | 97.5 ± 0.5 |
| Kit C | 0.2 ± 0.1 | 2.6 ± 0.2 | 0.4 ± 0.02 | 2 ± 0.1 | 97.4 ± 0.3 |

**Table 1.** Summary statistics for the effects for read trimming and pre-processing for all libraries by kit identifier, prior to alignment to a transcript reference.

The summary statistics compiled in Table 1 assist in highlighting the variation in filtering requirements and effects for each kit. Kit A requires slightly more 3' read trimming due to adapter contamination at the 3' end of the read, as well as increased filtering based on sequences that are too short or of low Phred quality scores. Conversely, both kits B and C have very similar profiles of read percentages affected by each filtering step as well as the percentage of reads retained after processing. In these two kits, there is more filtering required to remove polyadenylated sequences than in Kit A. This serves as one of the first indications of kit divergence in terms of quality and content of reads that are captured by each preparation. Overall, a high percentage of reads is retained for each kit for proceeding to alignment, with Kit A retaining a slightly smaller percentage of reads as compared

to the other two preparations, due to the level of adapter removal and short/low-quality read filtering that is applied to this kit.

**Alignment**

FASTQ files were individually aligned to the Mus_musculus.GRCm38.dna.primary_assembly reference genome obtained from Ensembl using STAR v. 2.7.5a. STAR is a splice-aware RNA-seq aligner that is optimized to align RNA transcripts across known splice junctions. It is considered to be one of the benchmark alignment tools for transcriptome bioinformatics pipelines. The following parameters were used to generate the reference genome in order to accommodate a maximum local memory capacity of 16GB:

--genomeSAsparseD 8

During individual FASTQ alignment, the following parameters are applied:

--readFilesCommand gunzip -c --outSAMtype BAM SortedByCoordinate

--alignIntronMin 50 --alignIntronMax 500000
-outSAMprimaryFlag OneBestScore --twopassMode Basic

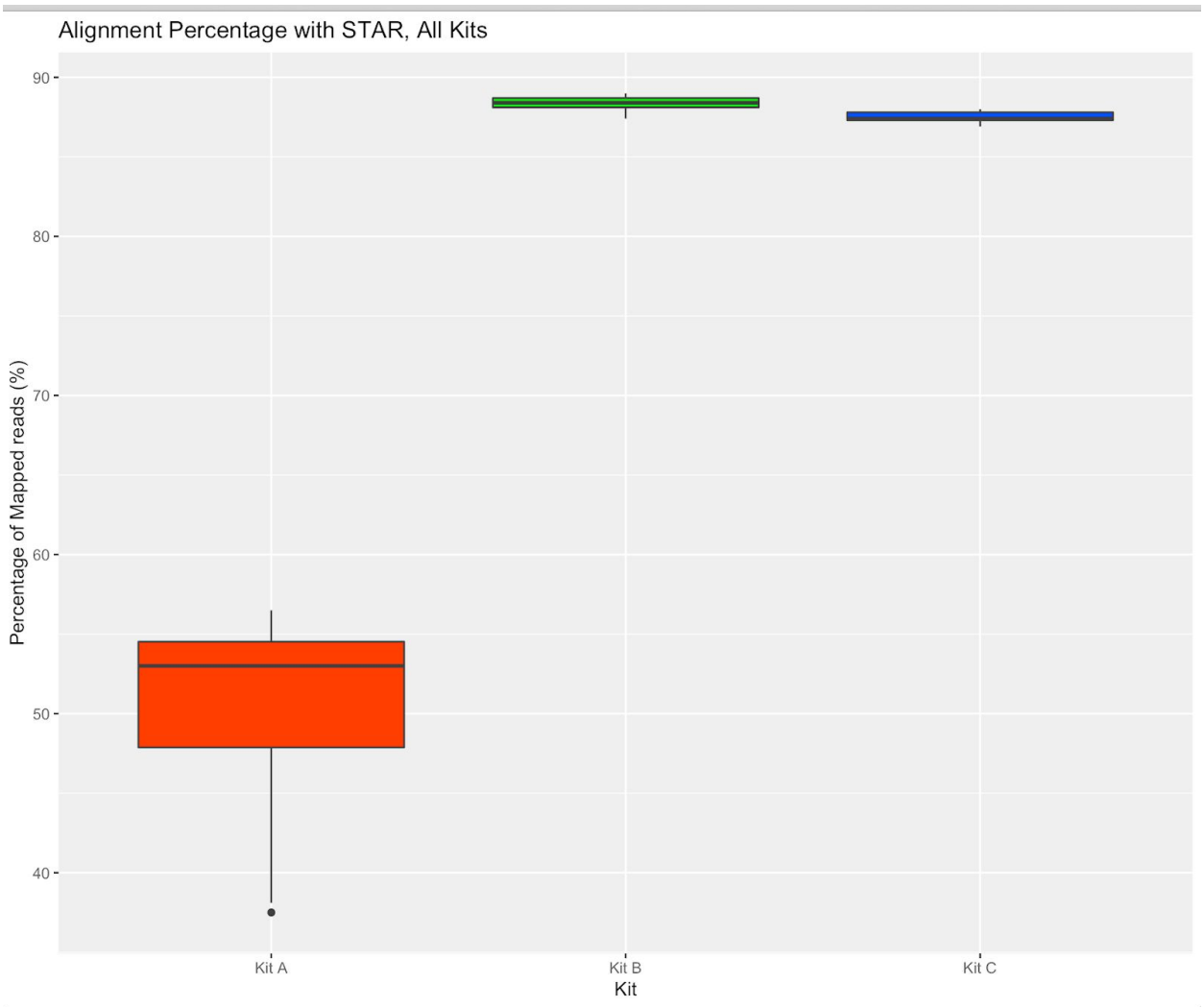
The two pass mode is an additional step that increases the total clock time for each alignment, but is overall beneficial for increasing the specificity for the aligner to detect novel junction splice sites within each library. This increased specificity can be useful for the detection of different transcript isoforms as well as novel junction splice sites.

**Alignment Statistics**

Basic alignment statistics for each library are shown below in the table, which covers the percentage of reads in each FASTQ that are successfully aligned to the reference genome, as well as the total counts of reads aligned expressed in millions of reads. The percentage of aligned reads per sample is very important to gauge the relative library quality and to observe any kit-specific patterns in read alignment percentages that may reflect on kit performance.

| Sample | Percent_Aligned (%) | Total_Seqs_Aligned | Kit |
|---|---|---|---|
| 26-AR56a_S1 | 37.975 | 6.8 | Kit A |
| 28-AR61a_S2 | 47.775 | 8.7 | Kit A |
| 31-AR68a_S3 | 52.825 | 9.1 | Kit A |
| 36-ARE36_S4 | 56.2 | 10.6 | Kit A |
| 41-ARE41_S5 | 54.475 | 9.8 | Kit A |
| 44-ARE44_S6 | 53.125 | 9.7 | Kit A |
| AR56a_S17 | 88.075 | 8.8 | Kit B |
| AR61a_S19 | 87.475 | 9.3 | Kit B |
| AR68a_S22 | 88.225 | 8 | Kit B |
| ARE36_S27 | 88.55 | 7.6 | Kit B |
| ARE41_S32 | 88.95 | 9.2 | Kit B |
| ARE44_S35 | 88.65 | 9.2 | Kit B |
| AR56a_HT_S36 | 87.45 | 7.2 | Kit C |
| AR61a_HT_S37 | 86.95 | 7.6 | Kit C |
| AR68a_HT_S38 | 87.975 | 7.6 | Kit C |
| ARE36_HT_S39 | 87.275 | 7.2 | Kit C |
| ARE41_HT_S40 | 87.75 | 6.8 | Kit C |
| ARE44_HT_S41 | 87.425 | 7.2 | Kit C |

A boxplot with the average percentage of mapped reads per kit is also effective in demonstrating the level of variance among the samples within each kit (shown below). One would expect a low level of deviation from the mean for the samples in a robust kit when measuring the percentage of successfully mapped reads. Deviations from the mean would indicate a larger degree of dispersion in library quality across a particular kit, and may indicate a deficiency in a particular kit being able to create representative libraries from a range of samples that have different biological complexities (measured as the total number of unique sequences). This hypothesis could be further investigated by creating and sequencing technical replicates for each sample type, which was not possible due to the particular set-up of this experiment.




Libraries prepared using Kit A show a noticeably lower percentage of successfully aligned reads as compared to either kits B or C. Kit C overall as compared to B has a very slightly smaller percentage of mapped reads across all samples, but at the sequencing depth provided in this experiment, this subtle difference is likely negligible. While read alignment percentages for both Kits B and C range from approximately 86-89%, indicating a highly successful alignment step, the range for Kit A read percentages shows a much larger range of deviation from the mean at 37.5-57%. Use of MultiQC in conjunction with the alignment logs compiled by STAR indicates that the majority of the unaligned reads in Kit A were too short to be successfully designated to a loci in the reference genome. Using stringent filtering steps at the read filtering and processing level, there was the expectation to retain only reads that are 35 bp long or greater, which was hypothesized to produce sufficiently long reads to align. However, it can be seen for these samples that there is a much higher percentage of reads that are shorter than the required length from STAR in order to perform the alignment. These values may be indicative of the RNA species types that are being retained (i.e. short RNA fragments), which may be reflective of the kit chemistry and mode of RNA capture (i.e. one kit is predicted to capture more non-coding regions, which may have shorter transcript lengths than full coding sequences). The percentage of mapped reads will be a key metric in deciding the fidelity of each kit as a high percentage of mapped reads should be a prerequisite for the selection

of a reproducible and consistent kit for a range of input samples; furthermore, unsuccessfully mapped reads cannot be attributed to a particular transcriptomic feature, and are therefore unusable in downstream RNA-seq expression analysis, so minimizing the unmapped percentage is an important element of an RNA library preparation that can generate high-quality transcriptomic data. If unmapped percentages remain high and variable for a particular kit, it is also important to characterize the nature of the unmapped reads so as to try to optimize the library preparation process or modify the procedures to suit the particular input sample.

**FeatureCounts Assignment Statistics**

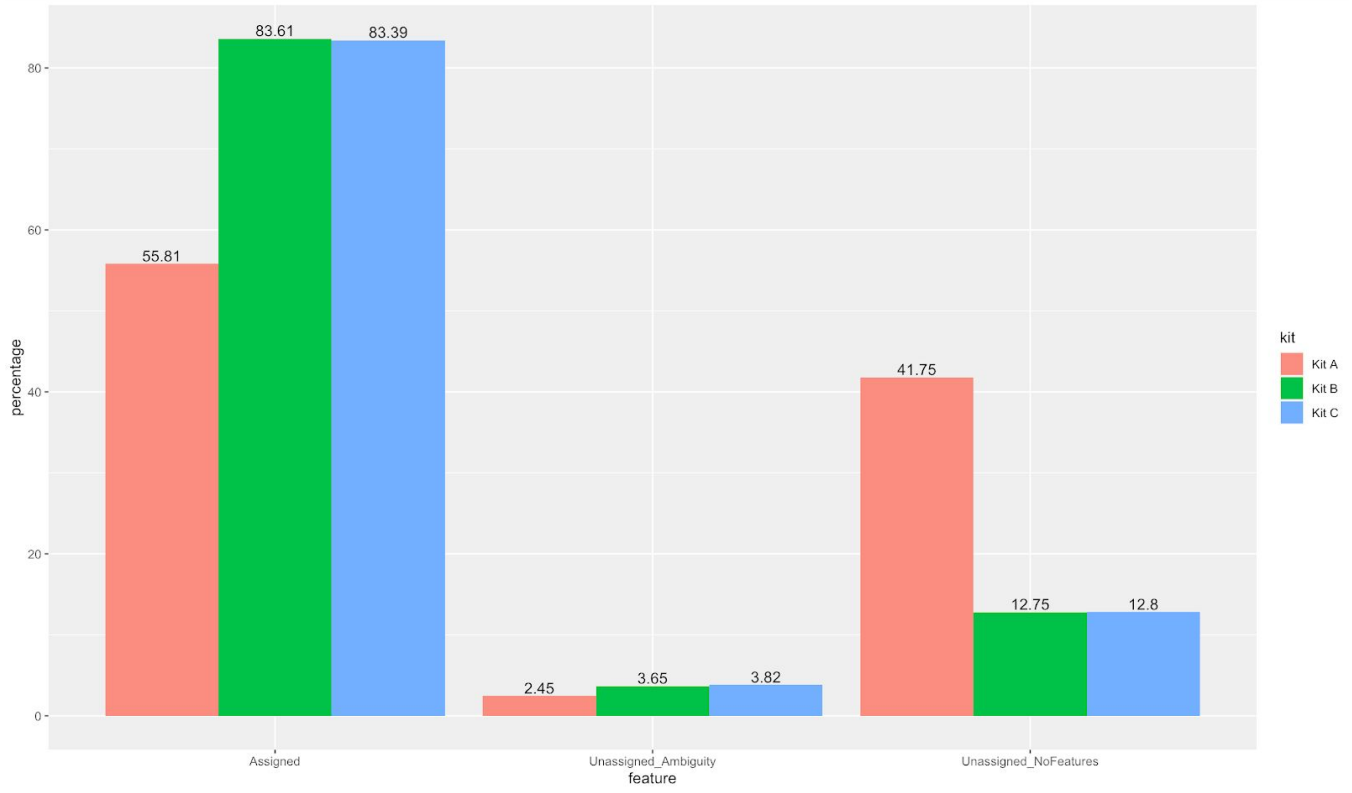
Within the featureCounts function from Rsubread, there are 14 categories for read assignment during the counting process, contained within the following list:

| |
|---|
| Assigned |
| Unassigned_Unmapped |
| Unassigned_Read_Type |
| Unassigned_Singleton |
| Unassigned_MappingQuality |
| Unassigned_Chimera |
| Unassigned_FragmentLength |
| Unassigned_Duplicate |
| Unassigned_MultiMapping |
| Unassigned_Secondary |
| Unassigned_NonSplit |
| Unassigned_NoFeatures |
| Unassigned_Overlapping_Length |
| Unassigned_Ambiguity |

The different classifications and proportions of unassigned reads are an important consideration during RNA-seq analysis in order to identify any specific wet lab or sequencing issues that may lead to reads being unassigned to a genomic feature. Reads may not be assigned to a feature for a variety of reasons, including sequence quality and length, a low degree of complexity leading to alignment in a repetitive portion of the genome, or a read that may map to multiple genomic loci. For the majority of the reads contained within the samples in this dataset, these unassigned categories will not be applicable, as our preprocessing tools were used to try to limit the number of unassigned features due to short fragment length and mapping quality.

The following plot describes the assignment of all the mapped reads from each library to a genomic feature using featureCounts. Note that any of the assignment categories from the table above that have 0 reads are not included in the graphic. Using this exclusion, we see that all 3 kits have reads assigned to only 3 categories: assigned, unassigned because of ambiguity, and unassigned because of a lack of feature in the annotation file used to perform the counting (unassigned_nofeatures).

During the annotation process, Kit A libraries were assigned a stranded option of 2 in featureCounts, corresponding to a reversely stranded/fr-first strand preparation. Libraries from both Kit B and C were not prepared using a stranded option, so their strand designation in featureCounts was 0. These designations are consistent with the vendor documentation for the strandedness of each kit.




The assignment profiles of both Kits B and C are very similar, and both exhibit a high percentage of assigned alignments over 80%. Kit B demonstrates noticeably a different profile, with a modest percentage of assigned alignments but a noticeable higher proportion of unassigned alignments that cannot be attributed to any gene or transcriptomic feature in the annotation file. Again, this higher proportion of unassigned reads may be due to the nature of the library prep chemistry, where it may be retaining certain transcriptomic features that have not been annotated, and are therefore not included in the annotation file used for assignment. The consistency and high degree of feature similarity between both B and C preparations suggests a robust technical reproducibility as well as accuracy during both the library prep and sequencing stages for these chemistries.

It should also be noted that for this analysis, read duplicates were kept unaltered for the counting and annotation steps. It is common during RNA-seq analysis to use Picard to mark the duplicate reads in an alignment file, by identifying reads that have identical genomic start and end coordinates and modifying the SAMtools flag for this metric. featureCounts allows the user to either ignore or use duplicate reads, and as such, marking duplicate reads for any of the FASTQ files would have produced a number of reads that would be assigned to the category of ‘unassigned_duplicate’. This proportion of the reads would roughly correspond to the proportion of duplicate reads found in the Basic Statistics Summary Table.

**Individual Feature Annotation Coverage (No Gene/Feature Filtering)**

Reads that were successfully aligned in the previous step were annotated and counted based on genome feature using featureCounts from the Rsubread v. 2.2.6 package. The GTF file used designated each transcriptomic feature (with a corresponding Ensembl gene ID) as an exon for the purposes of the featureCounts counting algorithm, and an accompanying gene information list in a text file provided the specific category of RNA feature for each of these possible annotations. A total of 46603 unique transcriptomic features were included in the annotation, with 43 unique RNA classifications that were possible. The distribution of the RNA types and their total possible counts is as follows:

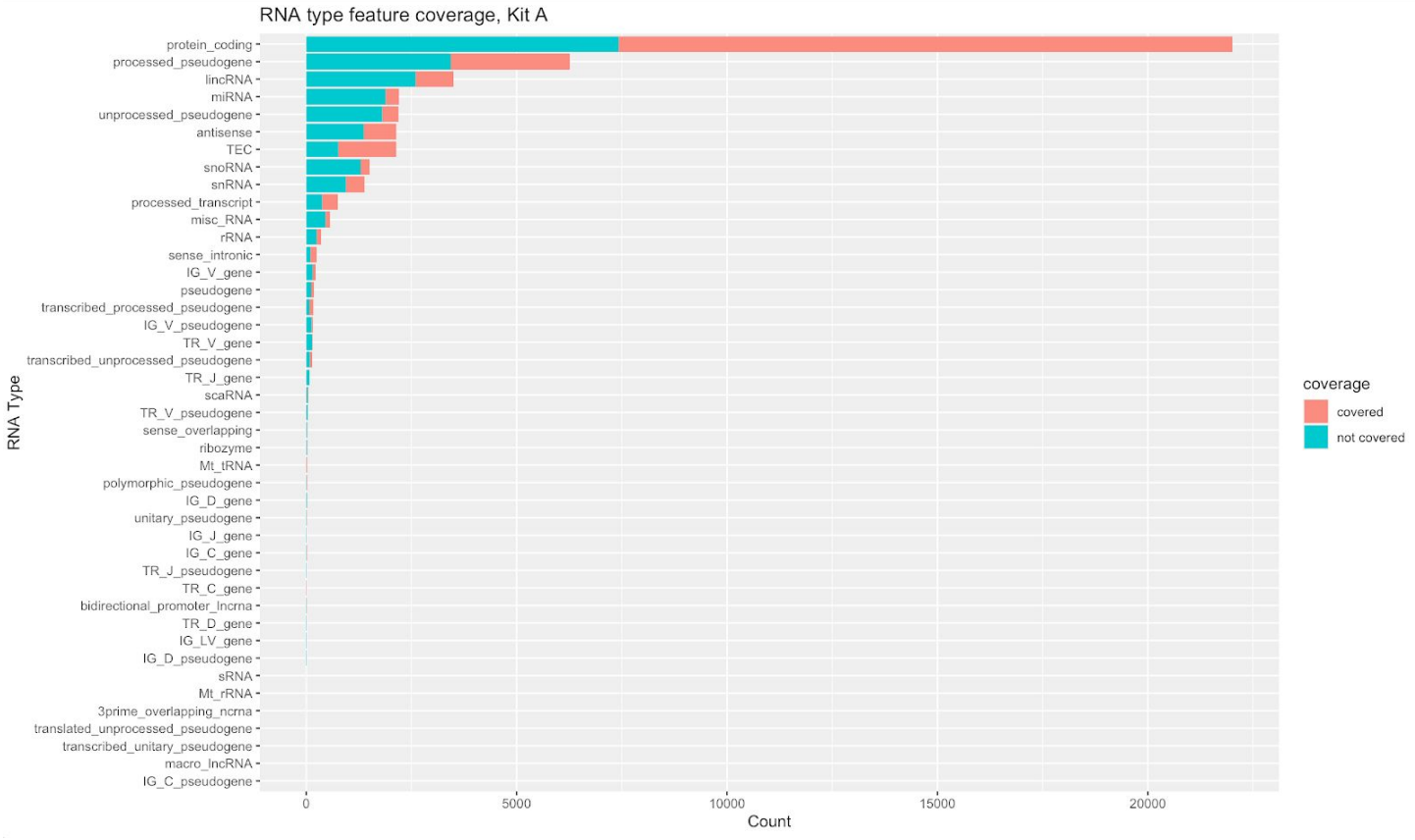
| | rna_type | total_features |
|---|---|---|
| 1 | 3prime_overlapping_ncrna | 2 |
| 2 | antisense | 2137 |
| 3 | bidirectional_promoter_lncrna | 8 |
| 4 | IG_C_gene | 13 |
| 5 | IG_C_pseudogene | 1 |
| 6 | IG_D_gene | 19 |
| 7 | IG_D_pseudogene | 4 |
| 8 | IG_J_gene | 14 |
| 9 | IG_LV_gene | 4 |
| 10 | IG_V_gene | 218 |
| 11 | IG_V_pseudogene | 156 |

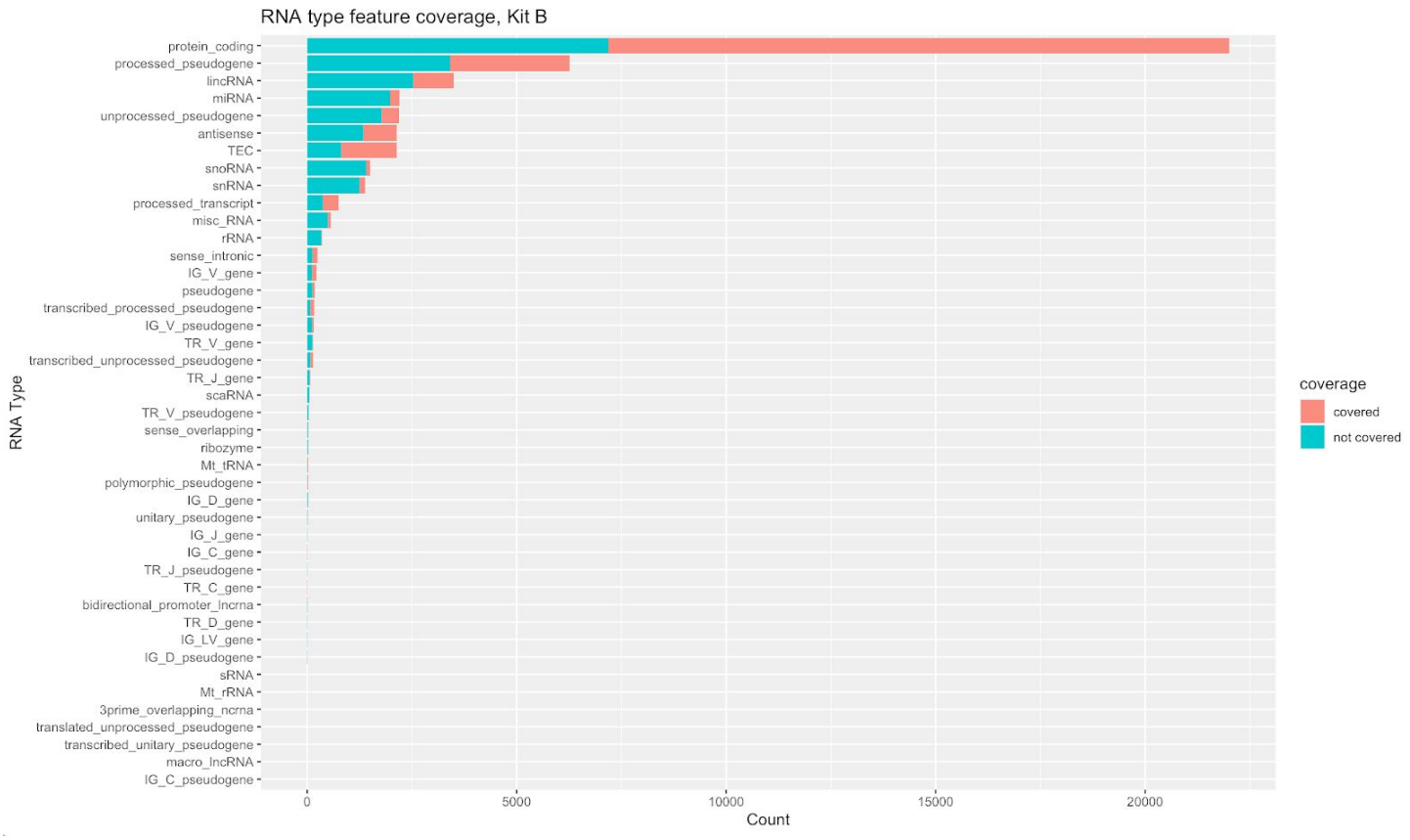
| | rna_type | total_features |
|---|---|---|
| 12 | lincRNA | 3495 |
| 13 | macro_lncRNA | 1 |
| 14 | miRNA | 2207 |
| 15 | misc_RNA | 564 |
| 16 | Mt_rRNA | 2 |
| 17 | Mt_tRNA | 22 |
| 18 | polymorphic_pseudogene | 21 |
| 19 | processed_pseudogene | 6258 |
| 20 | processed_transcript | 744 |
| 21 | protein_coding | 22013 |
| 22 | pseudogene | 178 |

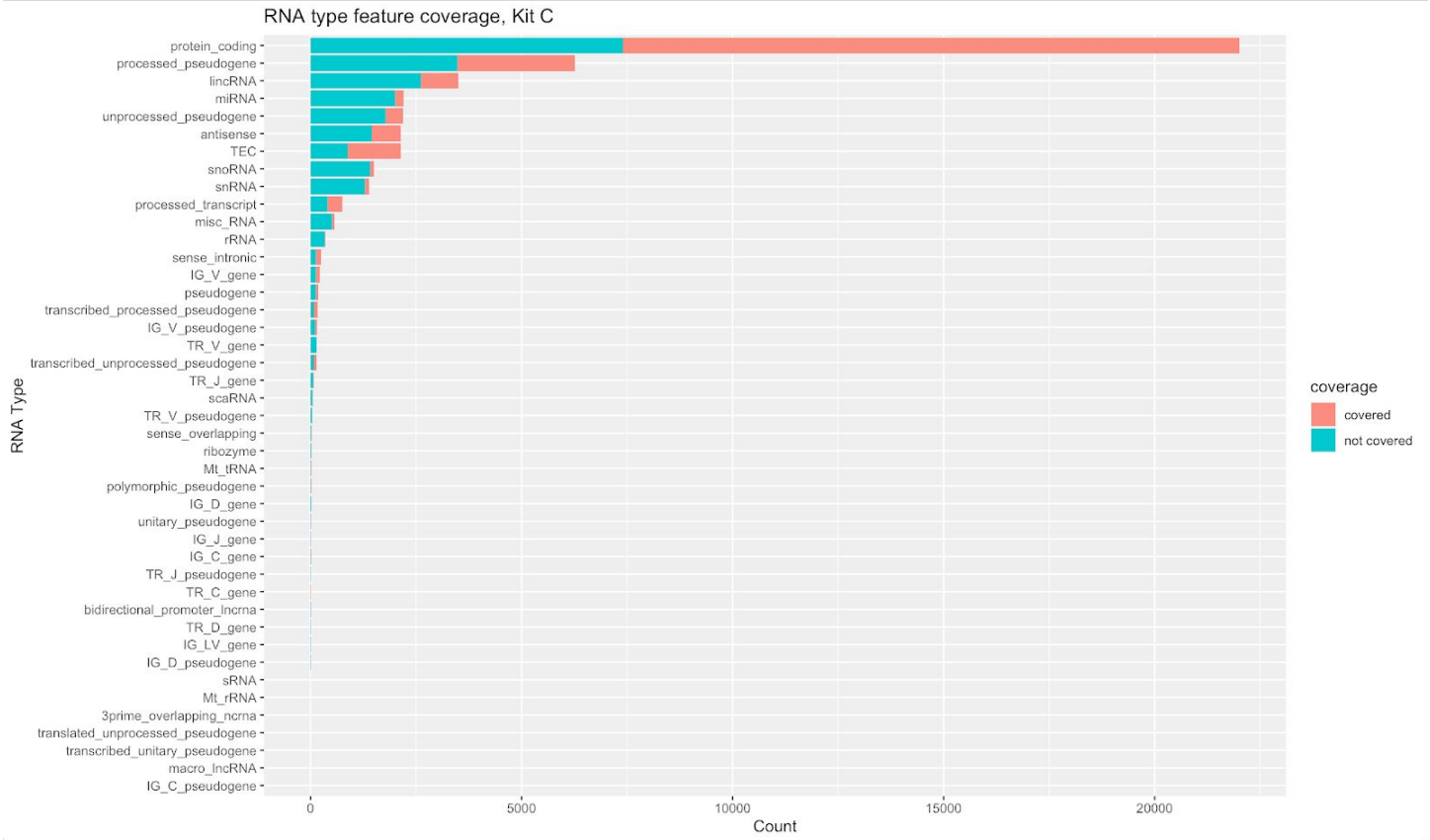
| | rna_type | total_features |
|---|---|---|
| 23 | ribozyme | 22 |
| 24 | rRNA | 354 |
| 25 | scaRNA | 51 |
| 26 | sense_intronic | 250 |
| 27 | sense_overlapping | 22 |
| 28 | snoRNA | 1508 |
| 29 | snRNA | 1384 |
| 30 | sRNA | 2 |
| 31 | TEC | 2134 |
| 32 | TR_C_gene | 8 |
| 33 | TR_D_gene | 4 |

| | rna_type | total_features |
|---|---|---|
| 34 | TR_J_gene | 70 |
| 35 | TR_J_pseudogene | 10 |
| 36 | TR_V_gene | 144 |
| 37 | TR_V_pseudogene | 34 |
| 38 | transcribed_processed_pseudogene | 173 |
| 39 | transcribed_unitary_pseudogene | 1 |
| 40 | transcribed_unprocessed_pseudogene | 142 |
| 41 | translated_unprocessed_pseudogene | 1 |
| 42 | unitary_pseudogene | 15 |
| 43 | unprocessed_pseudogene | 2193 |

During annotation, a kit was determined to have captured an RNA feature if at least one of its corresponding libraries obtained at least 1 count of its feature using featureCounts. The distribution of features that are both covered and not covered by each kit is shown below. It is important to appreciate that the tissue type for the mouse samples plays a significant role in the overall coverage of the RNA types. We do not expect to see perfect and uniform coverage of all RNA types in a particular tissue sample due to variation in tissue-specific gene expression, as well as any spatial-temporal differences in expression that are caused by biological processes. The comparison of the different individual annotation coverages serves as a useful indication of the ability for each kit to cover different RNA species, as the samples applied to each kit were originally derived from the same set of 6 unique biological mouse samples.


RNA type feature coverage, Kit A
protein_coding
processed_pseudogene
lincRNA
miRNA
unprocessed_pseudogene
antisense
TEC
snoRNA
snRNA
processed_transcript
misc_RNA
rRNA
sense_intronic
IG_V_gene
pseudogene
transcribed_processed_pseudogene
IG_V_pseudogene
TR_V_gene
transcribed_unprocessed_pseudogene
TR_J_gene
scaRNA
TR_V_pseudogene
sense_overlapping
ribozyme
Mt_tRNA
polymorphic_pseudogene
IG_D_gene
unitary_pseudogene
IG_J_gene
IG_C_gene
TR_J_pseudogene
TR_C_gene
bidirectional_promoter_lncrna
TR_D_gene
IG_LV_gene
IG_D_pseudogene
sRNA
Mt_rRNA
3prime_overlapping_ncrna
translated_unprocessed_pseudogene
transcribed_unitary_pseudogene
macro_lncRNA
IG_C_pseudogene
RNA Type
0
5000
10000
15000
20000
Count
coverage
covered
not covered

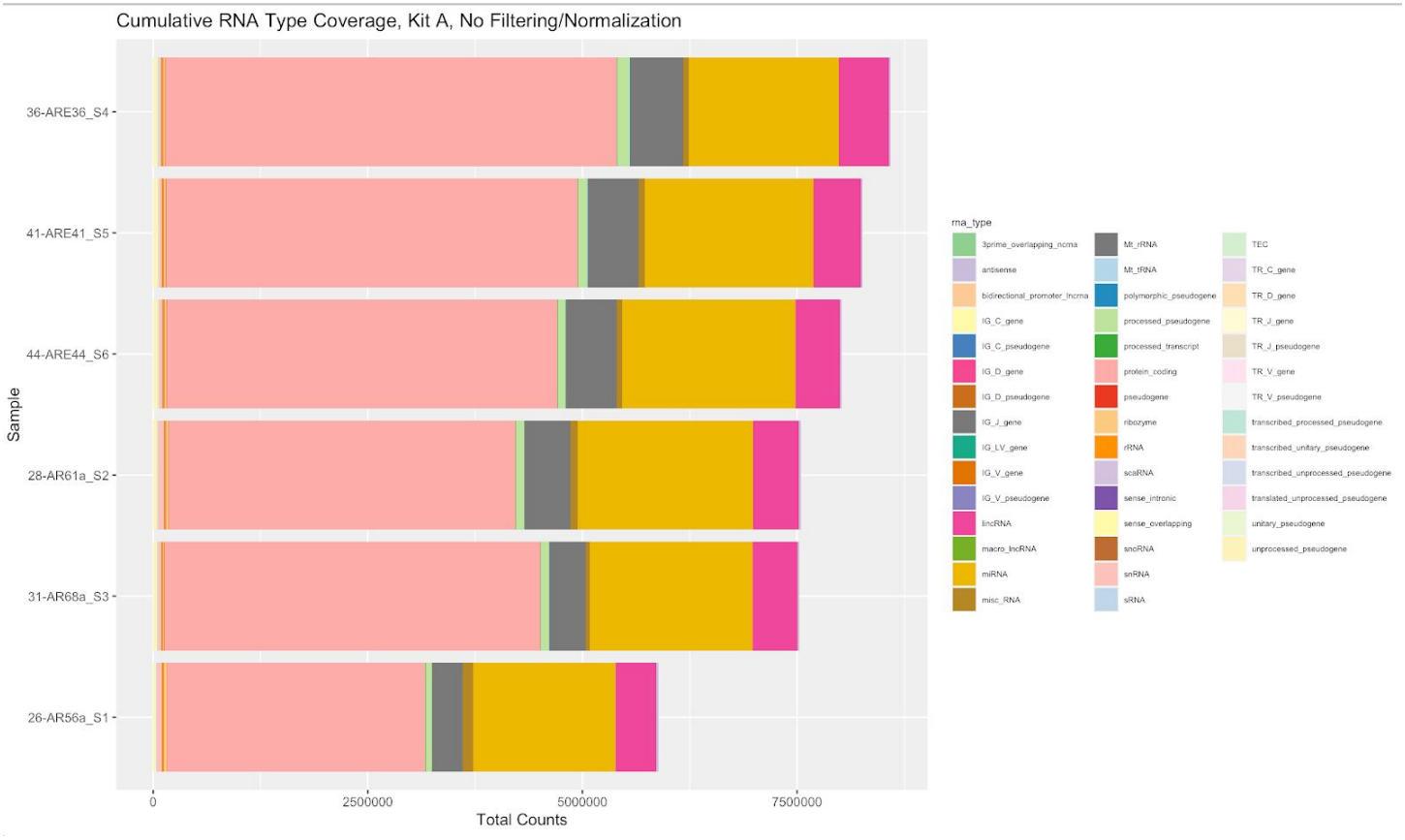

RNA type feature coverage, Kit B
protein_coding
processed_pseudogene
lincRNA
miRNA
unprocessed_pseudogene
antisense
TEC
snoRNA
snRNA
processed_transcript
misc_RNA
rRNA
sense_intronic
IG_V_gene
pseudogene
transcribed_processed_pseudogene
IG_V_pseudogene
TR_V_gene
transcribed_unprocessed_pseudogene
TR_J_gene
scaRNA
TR_V_pseudogene
sense_overlapping
ribozyme
Mt_tRNA
polymorphic_pseudogene
IG_D_gene
unitary_pseudogene
IG_J_gene
IG_C_gene
TR_J_pseudogene
TR_C_gene
bidirectional_promoter_lncrna
TR_D_gene
IG_LV_gene
IG_D_pseudogene
sRNA
Mt_rRNA
3prime_overlapping_ncrna
translated_unprocessed_pseudogene
transcribed_unitary_pseudogene
macro_lncRNA
IG_C_pseudogene
RNA Type
0
5000
10000
15000
20000
Count
coverage
covered
not covered

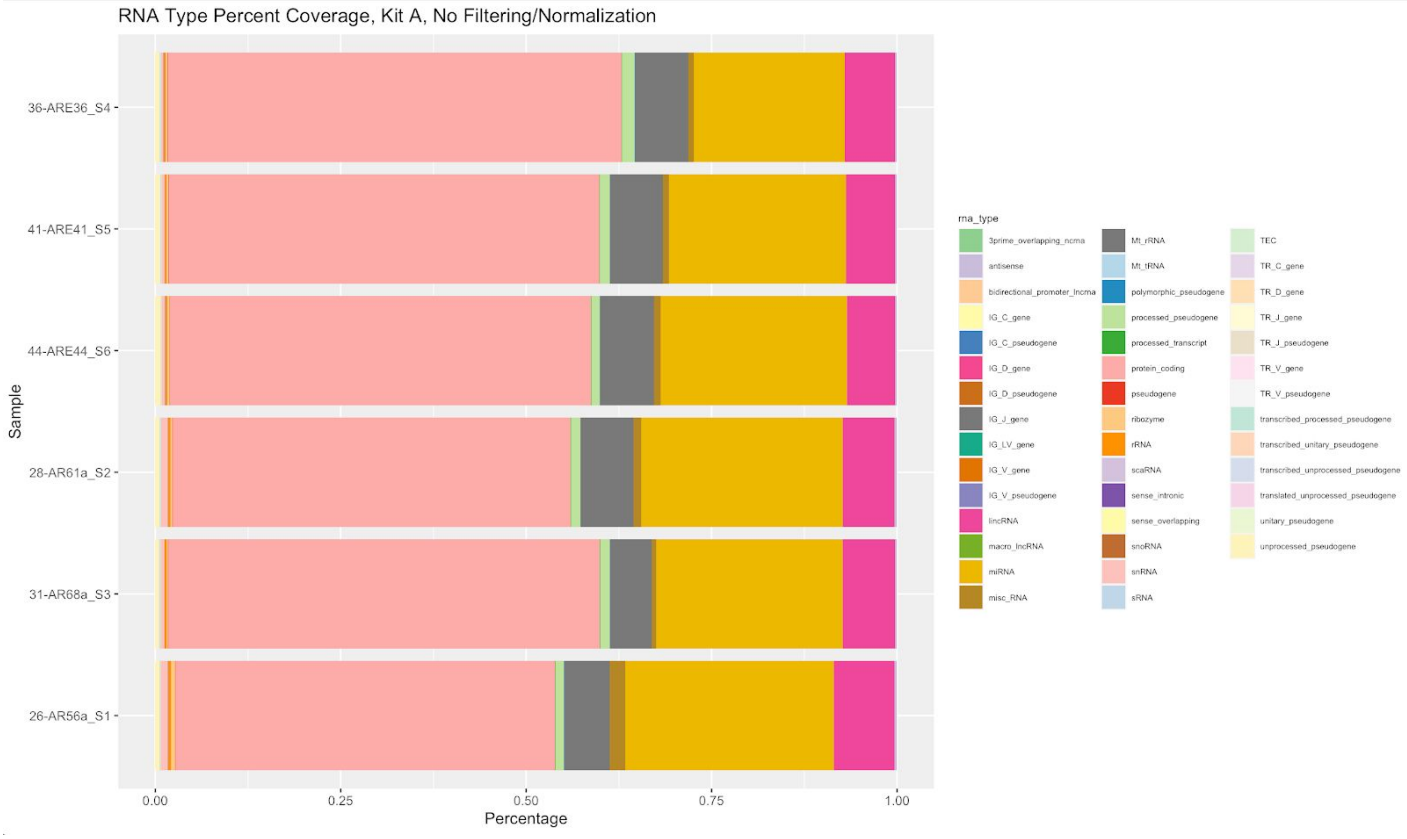


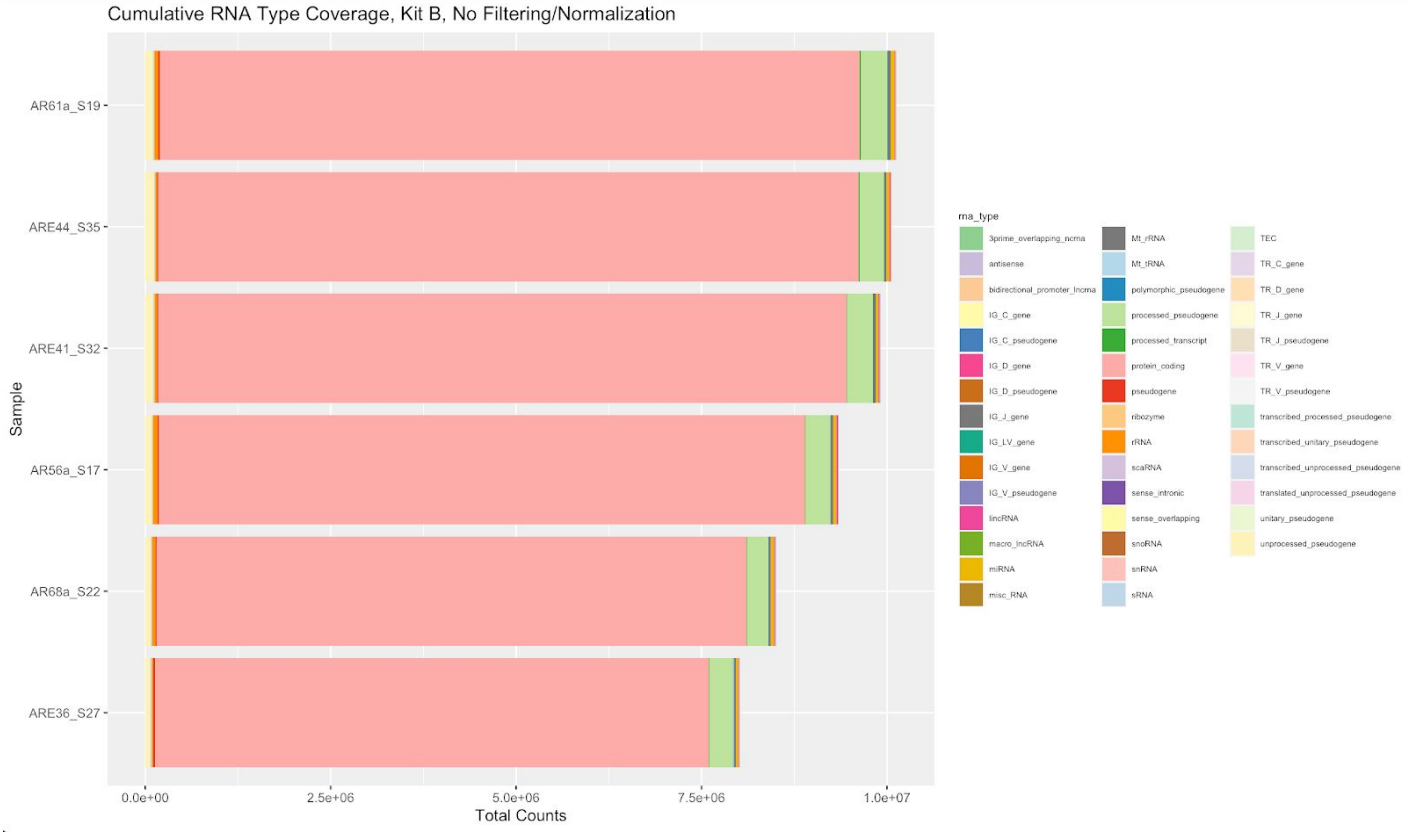

In general the three kits were able to capture and cover a similar pattern of RNA feature types when measured individually. It can be seen that Kit A was able to capture a slightly larger number of unique coding genes as compared to either B or C, whose individual RNA capture profiles are very similar to each other. However these graphs do not illustrate the overall capture rates of each kit in terms of the total number of each possible feature that are obtained. They need to be analyzed in conjunction with total feature counts to assess overall transcriptome coverage; additionally, we would need to verify that samples prepared using the same kit had similar patterns of RNA capture coverage, indicating an acceptable degree of technical reproducibility. Because of variations in the sequencing depth for each of the kits, higher annotation coverages for individual features may simply be a result of sampling from a greater number of reads, rather than observing a kit that is capable of capturing the majority of transcriptomic features across a wide range of input sample types. The individual annotation coverage may indicate the consistency of the gene expression profiles of the input samples that were provided to each kit.

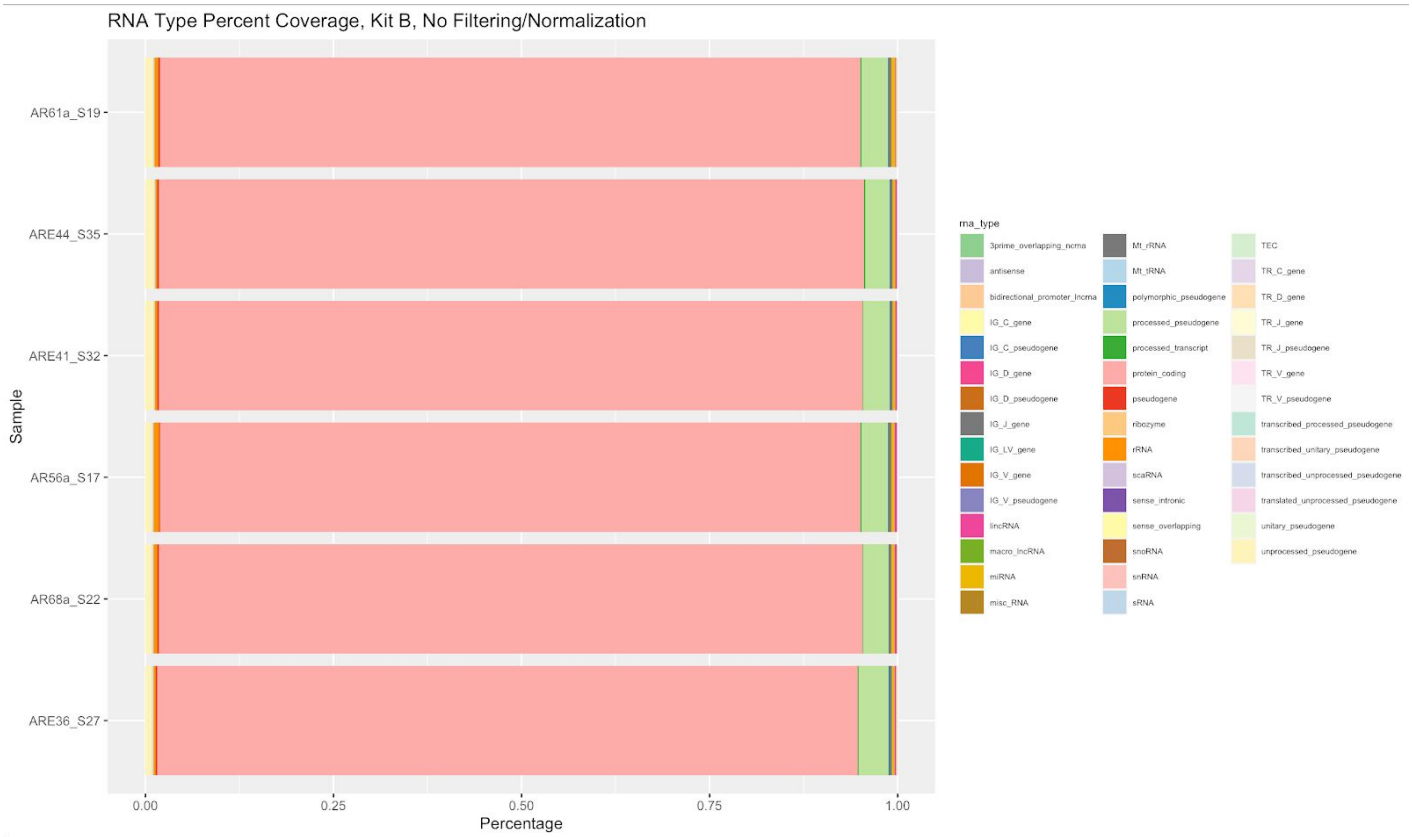
**Cumulative RNA type Capture Coverage (No Gene/Feature Normalization)**

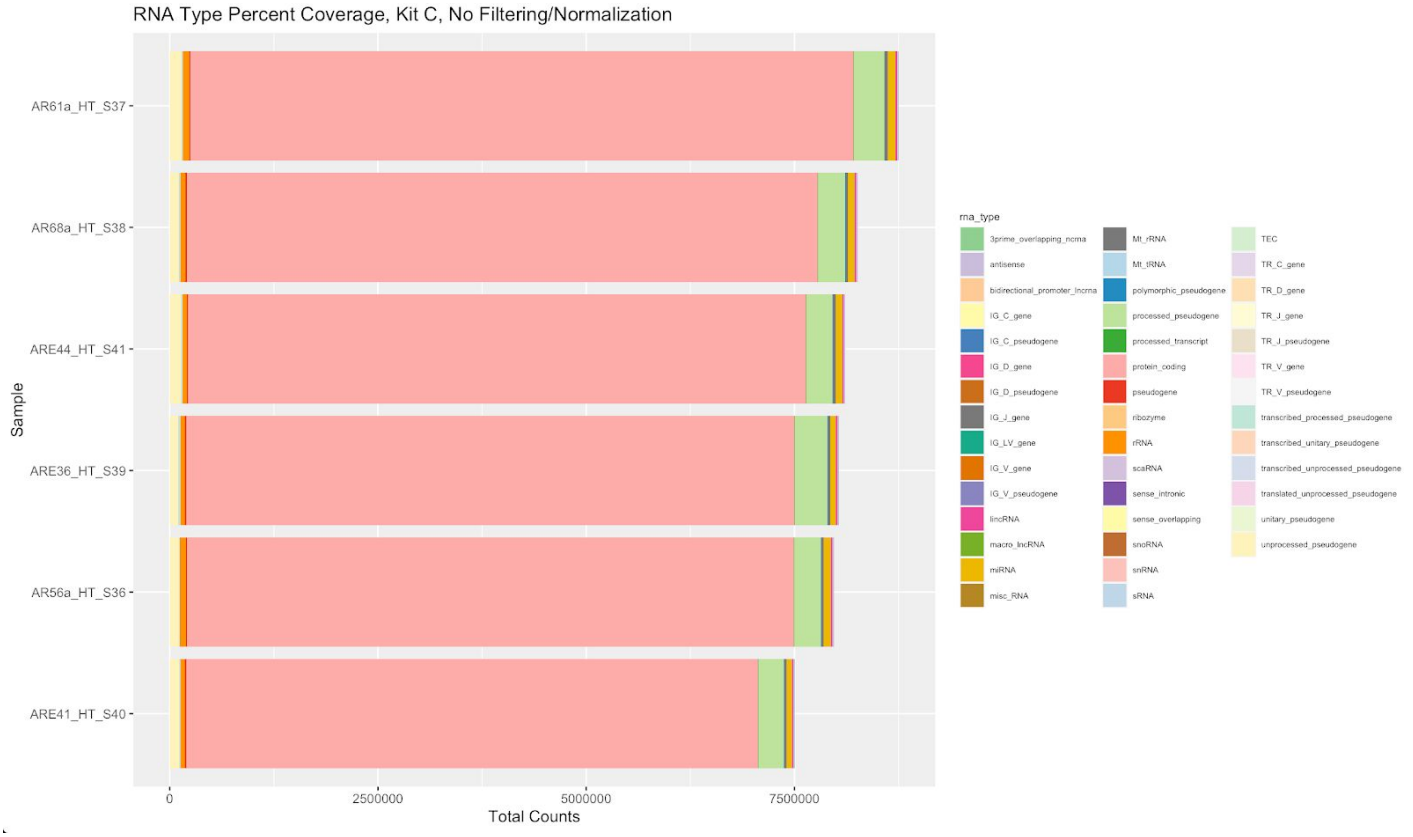
Total alignment counts for each annotated feature were also performed as a means of assessing the overall library coverage capabilities. The results from these analyses are shown below; for each sample a total feature count as well as percentage is shown. It is important to include the overall percentage of RNA type captured in each sample, as the libraries will naturally have slightly different sequencing depths, which will lead to variations in the total counts for each feature. In general we would need to see consistency across the percentages of RNA types for each kit to conclude that the preparations from a specific kit can be considered robust and reproducible. However, identifying truly reproducible libraries would require replicates for each biological sample applied to the kits, which would serve as a good extension of this particular experiment.

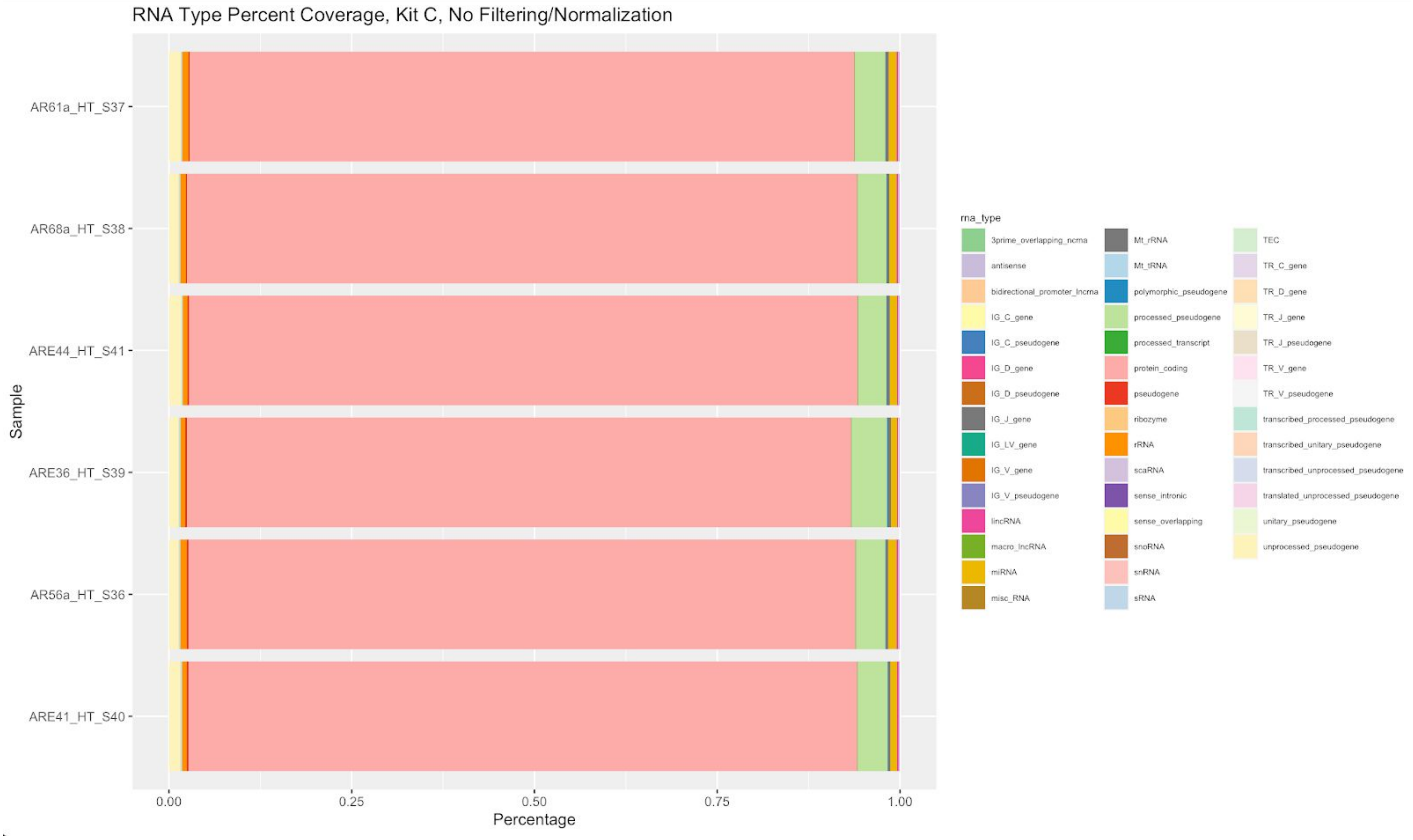

Cumulative RNA Type Coverage, Kit A, No Filtering/Normalization
36-ARE36_S4
41-ARE41_S5
44-ARE44_S6
28-AR61a_S2
31-AR68a_S3
26-AR56a_S1
Sample
0
2500000
5000000
7500000
Total Counts
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


RNA Type Percent Coverage, Kit A, No Filtering/Normalization
36-ARE36_S4
41-ARE41_S5
44-ARE44_S6
28-AR61a_S2
31-AR68a_S3
26-AR56a_S1
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


Cumulative RNA Type Coverage, Kit B, No Filtering/Normalization
AR61a_S19
ARE44_S35
ARE41_S32
AR56a_S17
AR68a_S22
ARE36_S27
Sample
0.0e+00
2.5e+06
5.0e+06
7.5e+06
1.0e+07
Total Counts
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


RNA Type Percent Coverage, Kit B, No Filtering/Normalization
AR61a_S19
ARE44_S35
ARE41_S32
AR56a_S17
AR68a_S22
ARE36_S27
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


RNA Type Percent Coverage, Kit C, No Filtering/Normalization
AR61a_HT_S37
AR68a_HT_S38
ARE44_HT_S41
ARE36_HT_S39
AR56a_HT_S36
ARE41_HT_S40
Sample
0
2500000
5000000
7500000
Total Counts
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


RNA Type Percent Coverage, Kit C, No Filtering/Normalization
AR61a_HT_S37
AR68a_HT_S38
ARE44_HT_S41
ARE36_HT_S39
AR56a_HT_S36
ARE41_HT_S40
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lncRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene

Assessment of the total feature counts as well as percentage of RNA features captured by each kit indicates that both B and C kit options capture a greater percentage of exonic/coding regions as compared to Kit A, with averages percentages of reads aligning to protein coding features being of 85-90%. The other RNA types that include various regulatory and non-coding RNAs are not strongly present in either of Kit B or C, with the next highest represent group, processed pseudogenes, being represented in approximately 5% of the reads in each library for these kits. Pseudogenes are non-functional "copies" of transcripts that resemble fully functional mRNA molecules; since they have poly-A-tails, their retention in these preparations confirms the poly-A-tail capture functionality and suggests an mRNA capture mechanism for Kits B and C. Conversely, we see a greater diversity of RNA types that are captured by Kit A at a percentage of 5% or higher. With Kit A we see a capture of protein coding regions in 50-60% of reads, with a larger proportion of annotations being found in miRNA, mtrRNA, and IG_D_gene categories as compared to B or C. In Kit A we see a smaller proportion of reads that correspond to processed pseudogenes (1-2%). Given this information, an initial hypothesis that Kit A is a total RNA capture kit is made.

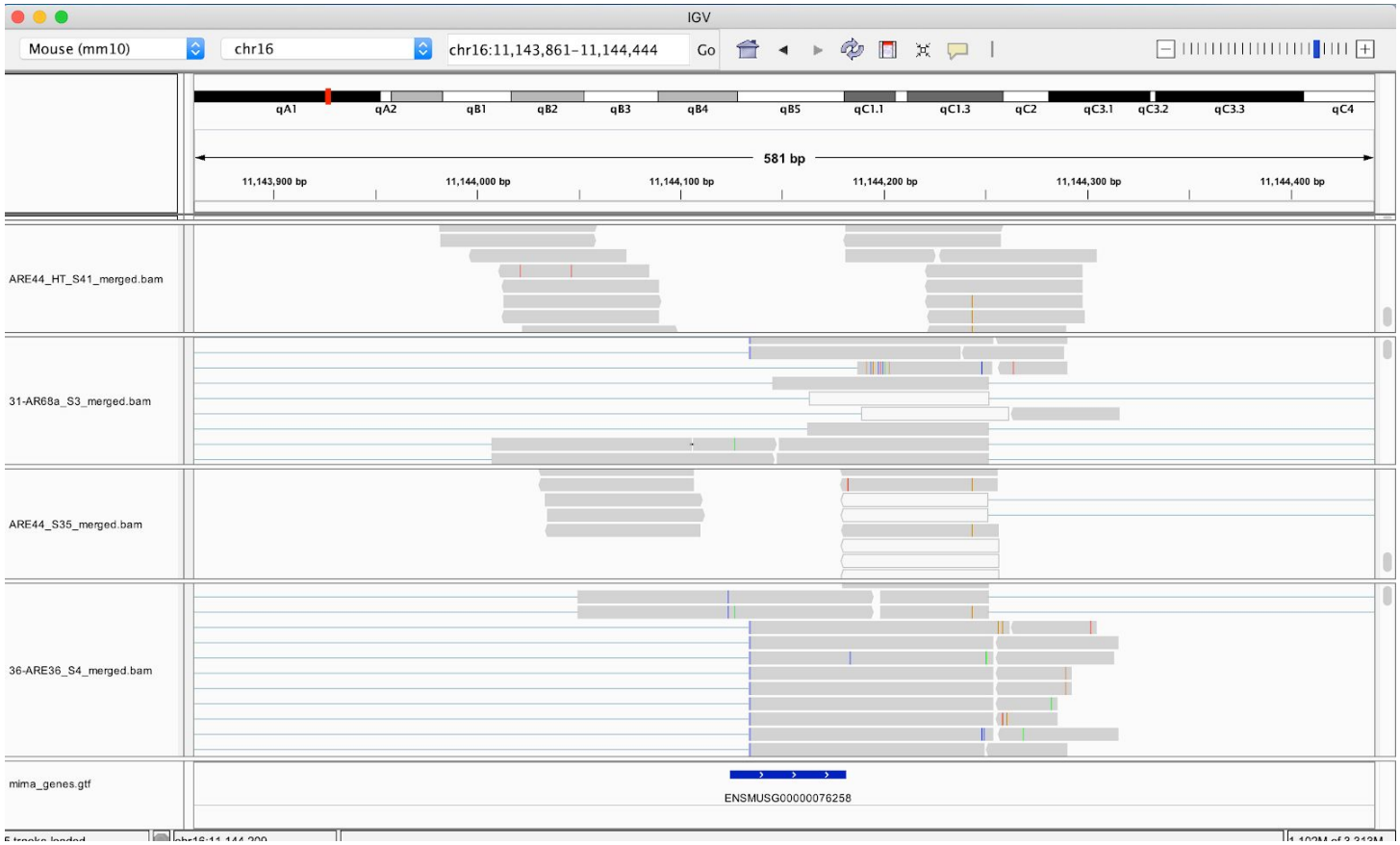
The relatively high capture rate of miRNA species with the Kit A preparation is of particular interest because miRNA mature transcripts are typically 21-25 nt in length; therefore, with stringent filtering options at the read level requiring read sequences to be at least 35 bp long before alignment, we would not expect to observe the extensive annotation of these short, non-coding regulatory RNAs in the analysis. By observing a subset of the miRNA genes that were annotated and collecting their transcript information from Ensembl, we can observe the transcript counts for the most expressed miRNA genes and their respective transcript lengths as given by Ensembl:

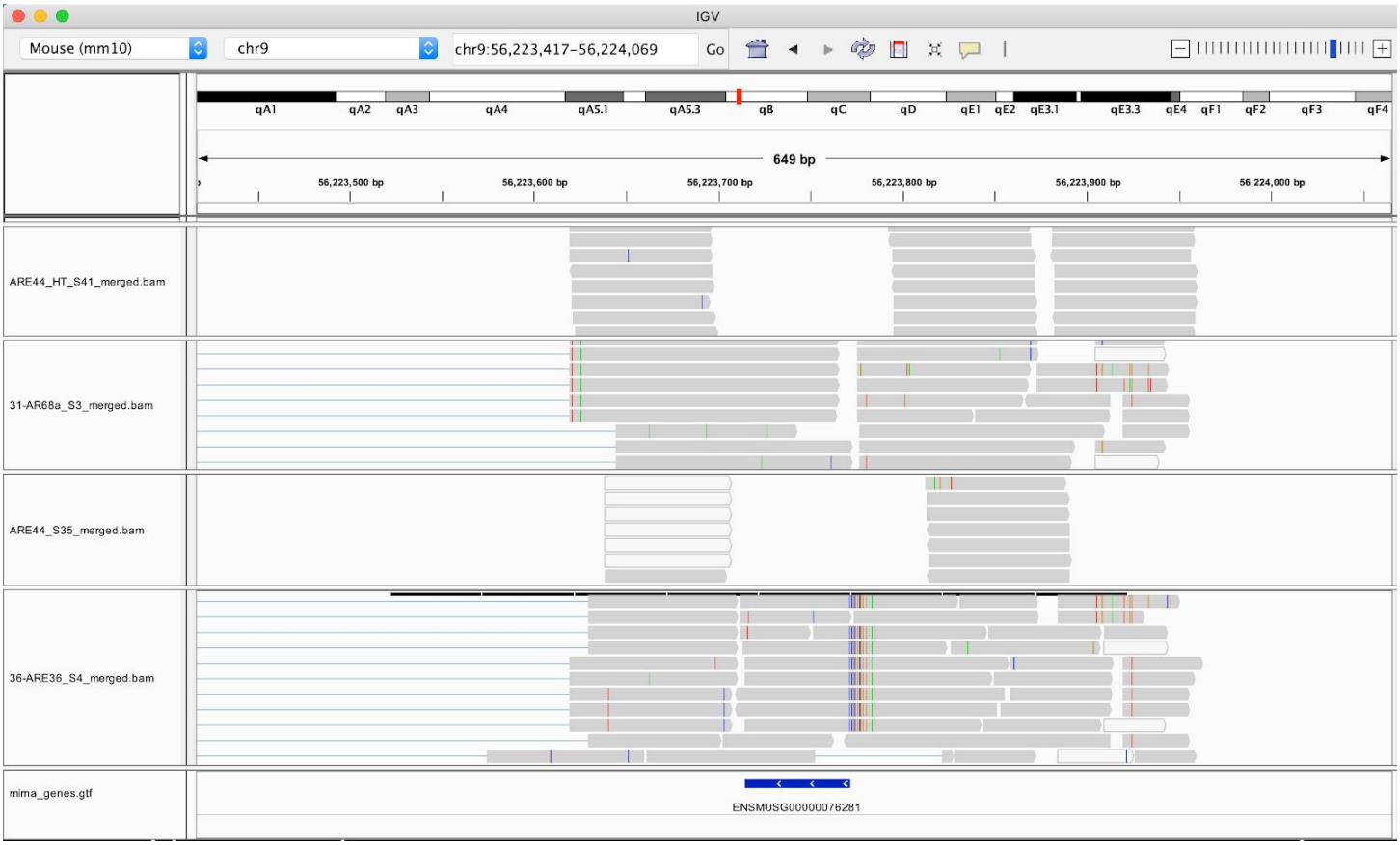
| | gene | sample | counts | length |
|---|---|---|---|---|
| *1895* | ENSMUSG00000076258 | 28-AR61a_S2 | 1494800 | 57 |
| *1893* | ENSMUSG00000076258 | 44-ARE44_S6 | 1488887 | 57 |
| *1894* | ENSMUSG00000076258 | 41-ARE41_S5 | 1469603 | 57 |
| *1896* | ENSMUSG00000076258 | 31-AR68a_S3 | 1388224 | 57 |
| *1891* | ENSMUSG00000076258 | 36-ARE36_S4 | 1304536 | 57 |
| *1892* | ENSMUSG00000076258 | 26-AR56a_S1 | 1154844 | 57 |
| *1937* | ENSMUSG00000076281 | 28-AR61a_S2 | 285666 | 57 |
| *1933* | ENSMUSG00000076281 | 44-ARE44_S6 | 282975 | 57 |
| *1938* | ENSMUSG00000076281 | 26-AR56a_S1 | 267824 | 57 |
| *1934* | ENSMUSG00000076281 | 41-ARE41_S5 | 263641 | 57 |
| *1935* | ENSMUSG00000076281 | 31-AR68a_S3 | 262955 | 57 |
| *1936* | ENSMUSG00000076281 | 36-ARE36_S4 | 221664 | 57 |
| *4052* | ENSMUSG00000088609 | 28-AR61a_S2 | 137335 | 57 |
| *4055* | ENSMUSG00000088609 | 44-ARE44_S6 | 133129 | 57 |
| *4051* | ENSMUSG00000088609 | 41-ARE41_S5 | 131369 | 57 |
| *4054* | ENSMUSG00000088609 | 31-AR68a_S3 | 127022 | 57 |
| *4056* | ENSMUSG00000088609 | 26-AR56a_S1 | 125467 | 57 |
| *4053* | ENSMUSG00000088609 | 36-ARE36_S4 | 113492 | 57 |
| *3965* | ENSMUSG00000088246 | 28-AR61a_S2 | 62100 | 56 |
| *3961* | ENSMUSG00000088246 | 26-AR56a_S1 | 60022 | 56 |
| *3963* | ENSMUSG00000088246 | 31-AR68a_S3 | 56909 | 56 |
| *3964* | ENSMUSG00000088246 | 44-ARE44_S6 | 54372 | 56 |
| *10372* | ENSMUSG00000098973 | 28-AR61a_S2 | 53214 | 123 |
| *3966* | ENSMUSG00000088246 | 41-ARE41_S5 | 49153 | 56 |
| *10371* | ENSMUSG00000098973 | 41-ARE41_S5 | 48609 | 123 |

It can be seen that the majority of gene counts that are attributed to miRNA features are accounted for by the first several miRNA genes as shown above. Furthermore, the length provided by Ensembl refers to the precursor, non-mature miRNA transcript from which the mature transcript is created. The typical length of the precursor transcript is 50-150 nt in length. This table therefore suggests that there are very high numbers of precursor miRNA transcripts that are being captured by the Kit A as a Total RNA kit, which may assist in explaining why this RNA feature category is so highly represented in the raw annotation coverage. Furthermore, when considering all of the miRNA genes that are found in Kit A libraries, the transcript length of the shortest counted feature is 39 nt, which again helps to explain why a cut-off of 35 bp for read lengths at the filtering step still allows for the retention of a large number of miRNA transcripts. However, the relative abundance of each of these putative miRNA transcripts is extremely high as compared to the protein coding sequences that were annotated, and may be due to some confounding factors during library prep or sequencing. It is entirely possible that some biased portion of the library prep process is causing preferential amplification of small read fragments in the library, leading to a skewed proportion of miRNA transcripts in the final representation. The values will need to be assessed after filtering and normalization to verify that these abundances are accurate.

Visualization of a subset of the alignments that correspond to miRNA loci can assist in resolving the alignment structures and quality for the libraries from each kit. Below are two screenshots taken from the Integrated Genomics Viewer (IGV) v.2.8.2, which provides a comprehensive and interactive tool for visualizing NGS alignments and various levels of resolution (i.e. chromosome or individual nucleotide).. Two libraries from Kit A are included, and one each from Kit B and C. The particular loci being viewed correspond to both ENSMUSG00000076258 and ENSMUSG00000076281, which are both predicted miRNA genes as annotated by Ensembl. These genes had the greatest and second greatest number of total annotations for miRNA genes from Kit A, as seen in the miRNA features table above. The genomic coordinates for the predicted transcripts from these miRNA genes in *M.Musculus* are as follows:

First IGV figure: ENSMUSG00000076258, Chromosome 16:11,144,125-11,144,181
Second IGV figure: ENSMUSG00000076281, Chromosome 9:56,223,715-56,223,771






In IGV, individual reads are shown as light grey blocks, and hovering over each particular alignment will grant information on its length, its read of origin from the FASTQ file, genomic location and CIGAR mapping summary, which indicates the number of matches, insertions, deletions, or gaps between alignments. For split reads that map to multiple genomic loci, the gap between alignments is shown as a light blue line, and the read information indicates the direction where the remainder of the read has mapped, as well as the total number of bases that spans the split alignment. The screenshots from IGV for these particular miRNA loci reveal a difference among kit samples for both the read depth and nature of the alignments. For samples from Kit A (31-AR68a_S3 and 36-ARE36_S4), there is a much higher proportion of split/chimeric reads where portions of the read that map to multiple loci in the genome that have a very high degree of variability in genomic distances; this is shown with reads that are connected to a light blue line. These gaps range in size from 10,000 to over 500,000 bp for these particular genes and are mapped to both the forward and reverse reference strand. For each of the other samples that originate from Kit B or C, we see a very small number of split reads, with the overwhelming majority being reads that fully align to one location within the gene. The majority of reads that cover these loci in Kit B and C are not split, and map almost completely to the locus shown in the viewer with very few insertions, deletions, or substitutions. Furthermore, the total number of reads that align to the miRNA loci is much smaller for both Kit B and C than the aligned read counts for Kit A.

These observations from IGV suggest a noticeable difference in mapping behaviour for certain miRNA genes in Kit A as opposed to the other two kits. The reads corresponding to miRNA genes that appear to be abundantly expressed in Kit A are chimeric and split across large genomic regions, which may suggest artefacts of an incomplete library preparation where portions of short reads from distant loci became fused, possible during amplification, leading to a high degree of split alignments. It is possible that since both kits B and C do not suffer from the same read fusion events leading to split read artefacts in the alignment step, that we do not observe the same proportion of chimeric alignments in miRNA loci as is seen with libraries from Kit A.

**Filtering and Normalization of Data for Counting and Annotation**

Filtering and normalization steps for the alignment data are essential for the removal of lowly expressed genes from the dataset that may interfere with proper annotation coverage and the statistical power of downstream differential expression tests. Additionally, we need to correct for RNA composition bias and sequencing depth to make the raw counts comparable across experimental runs. Therefore the following filtering and normalization steps are applied to the data before annotation and counting:

***- Removal of gene features with very low counts:*** It is often unreliable to call the expression of a gene that produces very few counts in an RNA sample. Filtering for these genes is often necessary to increase the reliability and statistical significance of differential expression patterns across treatments. This is achieved using a simple filtering by counts/reads per million, or RPM. Using this calculation, gene features with an RPM less than 1 for any of the samples in a kit are removed.

Of the 46603 possible gene annotations, the following gene numbers are retained for each kit after the RPM filtering:

| | |
|---|---|
| Kit A: | 11796 |
| Kit B: | 12722 |
| Kit C: | 12296 |

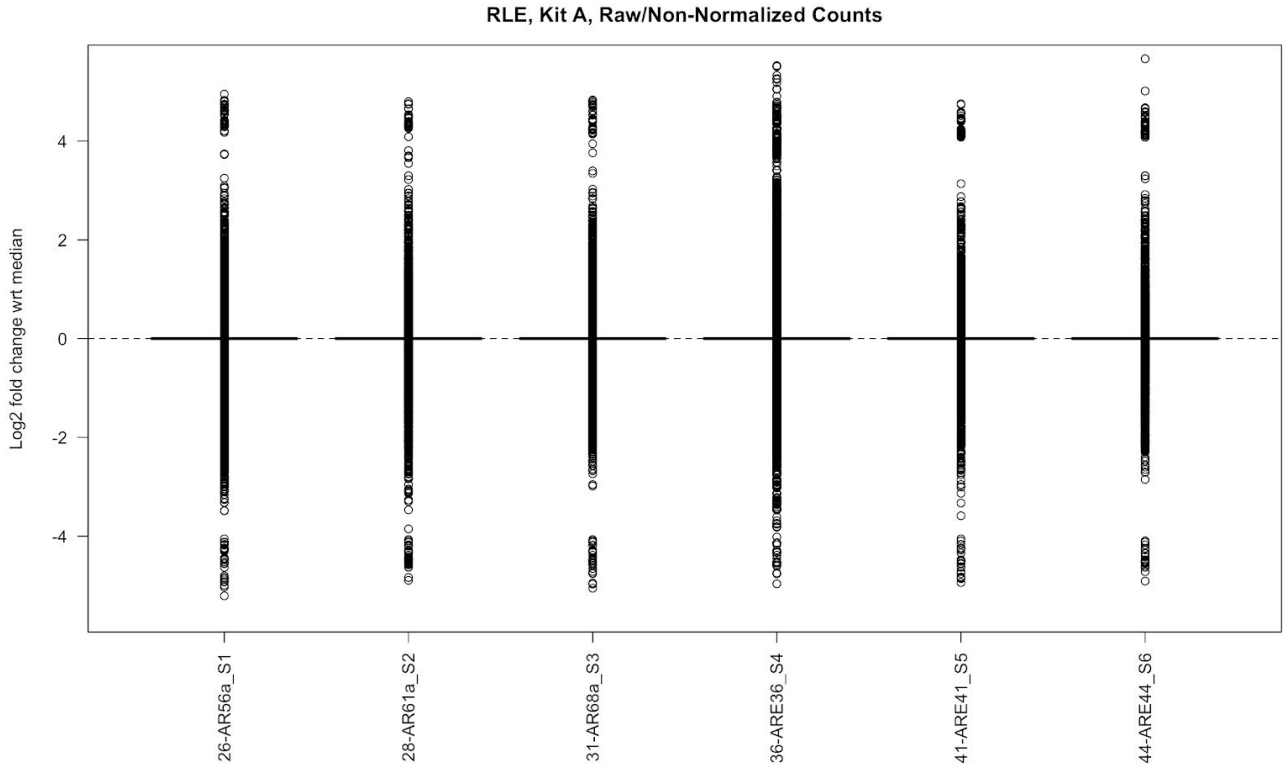
***- Normalization for the relative GC content of each library,*** conducted using the EDAseq package and the "within lane normalization" function using full quantile normalization.

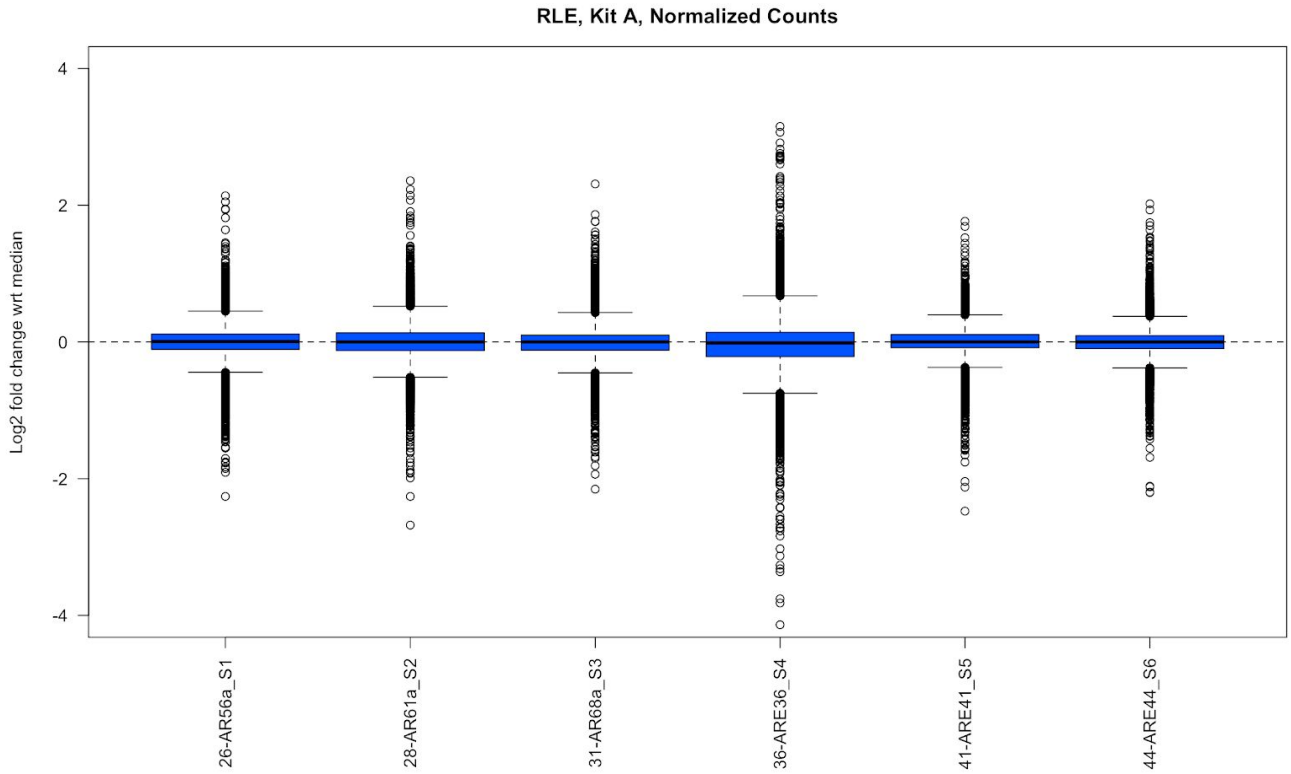
***- Normalization for sequencing depth.*** Libraries that have a greater number of total reads will naturally have more raw counts for certain features due to sequencing depth. Gene expression assays need to use normalized counts for sequencing depth, again achieved using EDAseq and the "between lane normalization" function using full quantile normalization.

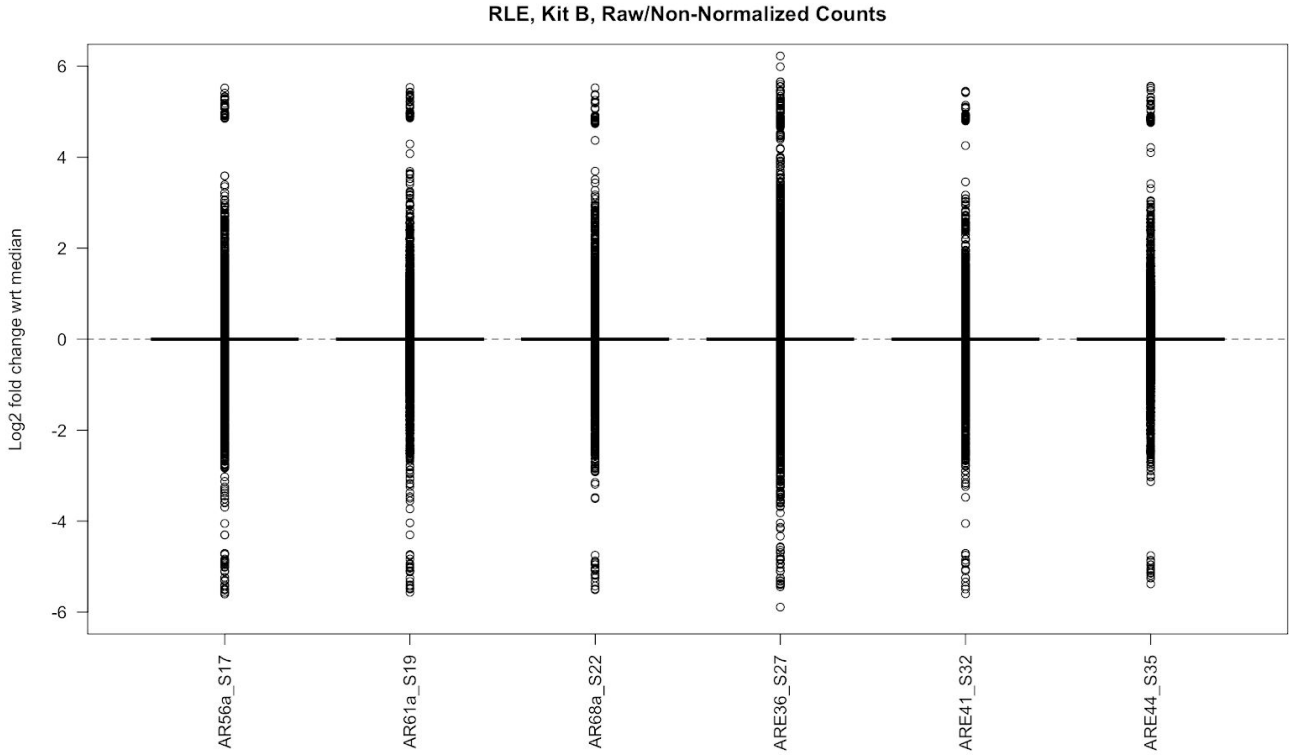
For both of the procedures listed above, all libraries belonging to the same preparation were normalized together, and separately from the other kits, in order to correct for batch effects and kit-specific biases as much as possible.

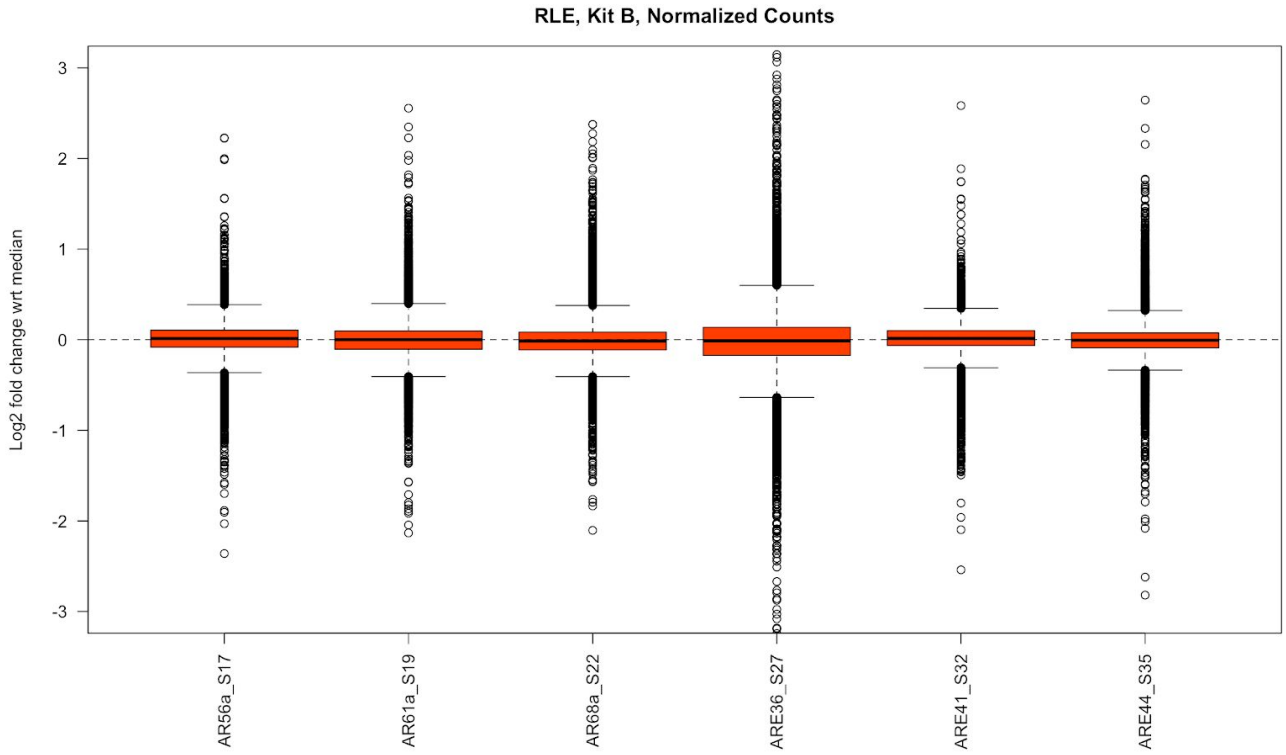
**Summary of Relative Transcript Expression Levels (Normalized)**

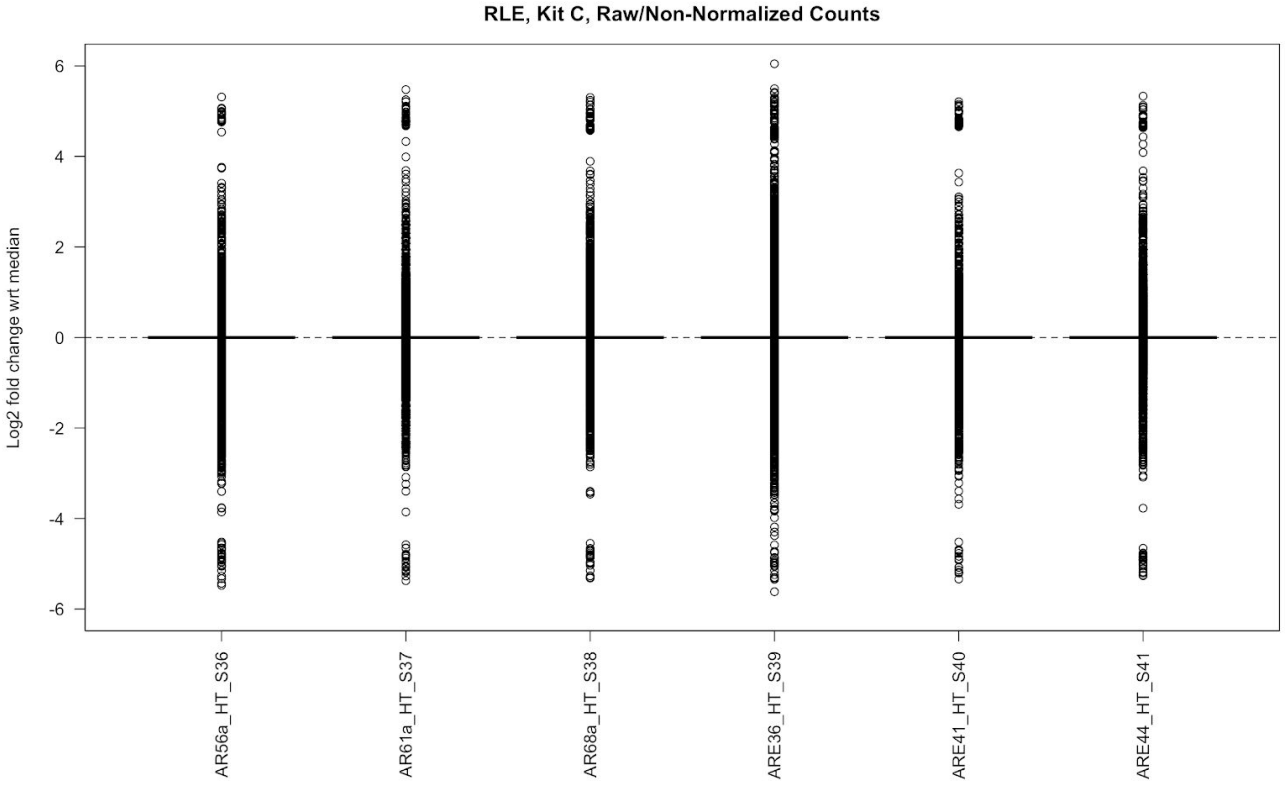
A characteristic goal of RNA-seq analysis and differential gene expression analysis is the identification and quantification of both differentially expressed genes, as well as any outlier genes that demonstrate unwanted technical variation due to experimental noise. Relative log expression (RLE) box plots can be used to view the distributions of read counts per gene relative to the median gene expression within the sample, providing an idea as to the degree of highly and lowly expressed genes in the dataset. This observation is important because it is assumed under most conditions that the majority of genes are not differentially expressed, and furthermore, that many genes have a low baseline level of expression in a particular biological condition. For this reason, we expect to see the majority of genes cluster around zero within each boxplot, with a tight distribution. Any deviations from the centre line of zero may indicate outlying genomic features or features that were not corrected using the normalization procedures, as described in Filtering and Normalization of Data for Counting and Annotation. The RLE boxplots for non-normalized data are also included to demonstrate the effects of the full normalization process on the distribution of expression values.

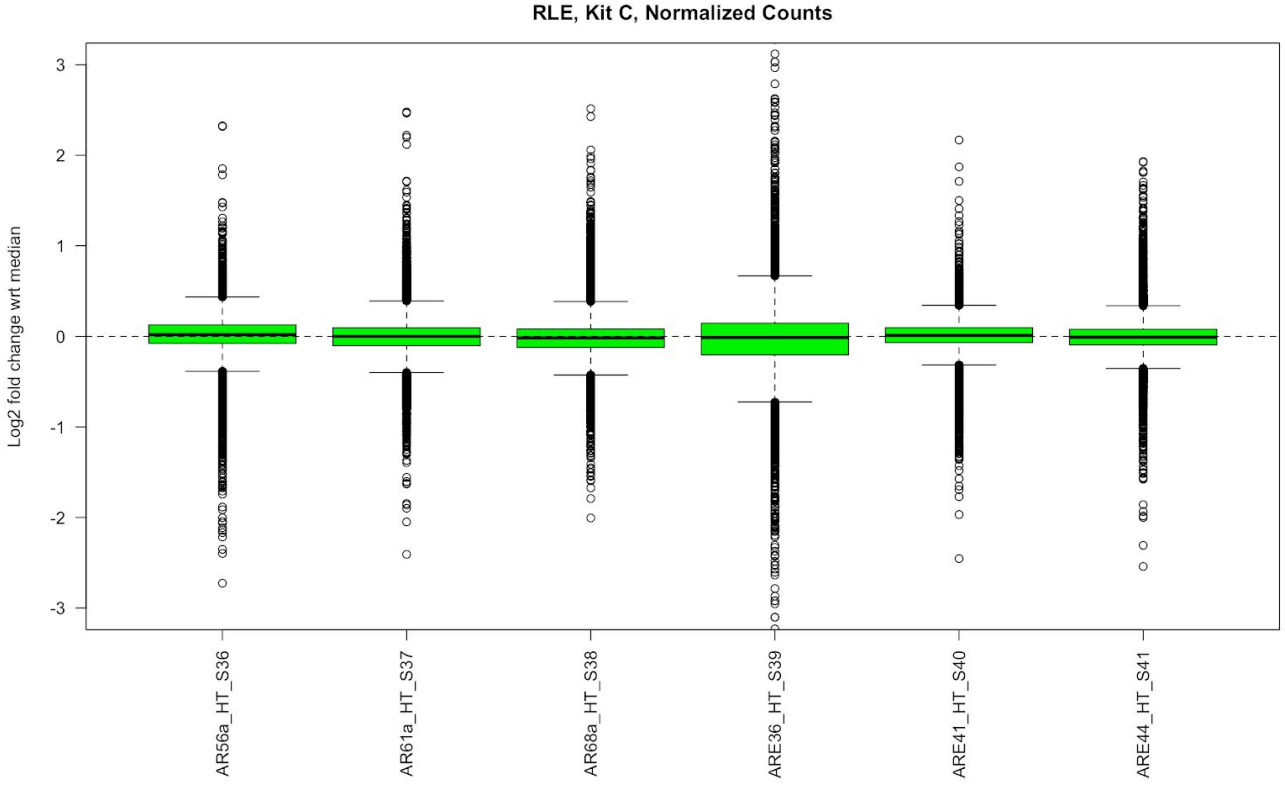

RLE, Kit A, Raw/Non-Normalized Counts
Log2 fold change wrt median
4
2
0
-2
-4
26-AR56a_S1
28-AR61a_S2
31-AR68a_S3
36-ARE36_S4
41-ARE41_S5
44-ARE44_S6


RLE, Kit A, Normalized Counts
Log2 fold change wrt median
4
2
0
-2
-4
26-AR56a_S1
28-AR61a_S2
31-AR68a_S3
36-ARE36_S4
41-ARE41_S5
44-ARE44_S6


RLE, Kit B, Raw/Non-Normalized Counts
Log2 fold change wrt median
6
4
2
0
-2
-4
-6
AR56a_S17
AR61a_S19
AR68a_S22
ARE36_S27
ARE41_S32
ARE44_S35


RLE, Kit B, Normalized Counts
Log2 fold change wrt median
3
2
1
0
-1
-2
-3
AR56a_S17
AR61a_S19
AR68a_S22
ARE36_S27
ARE41_S32
ARE44_S35

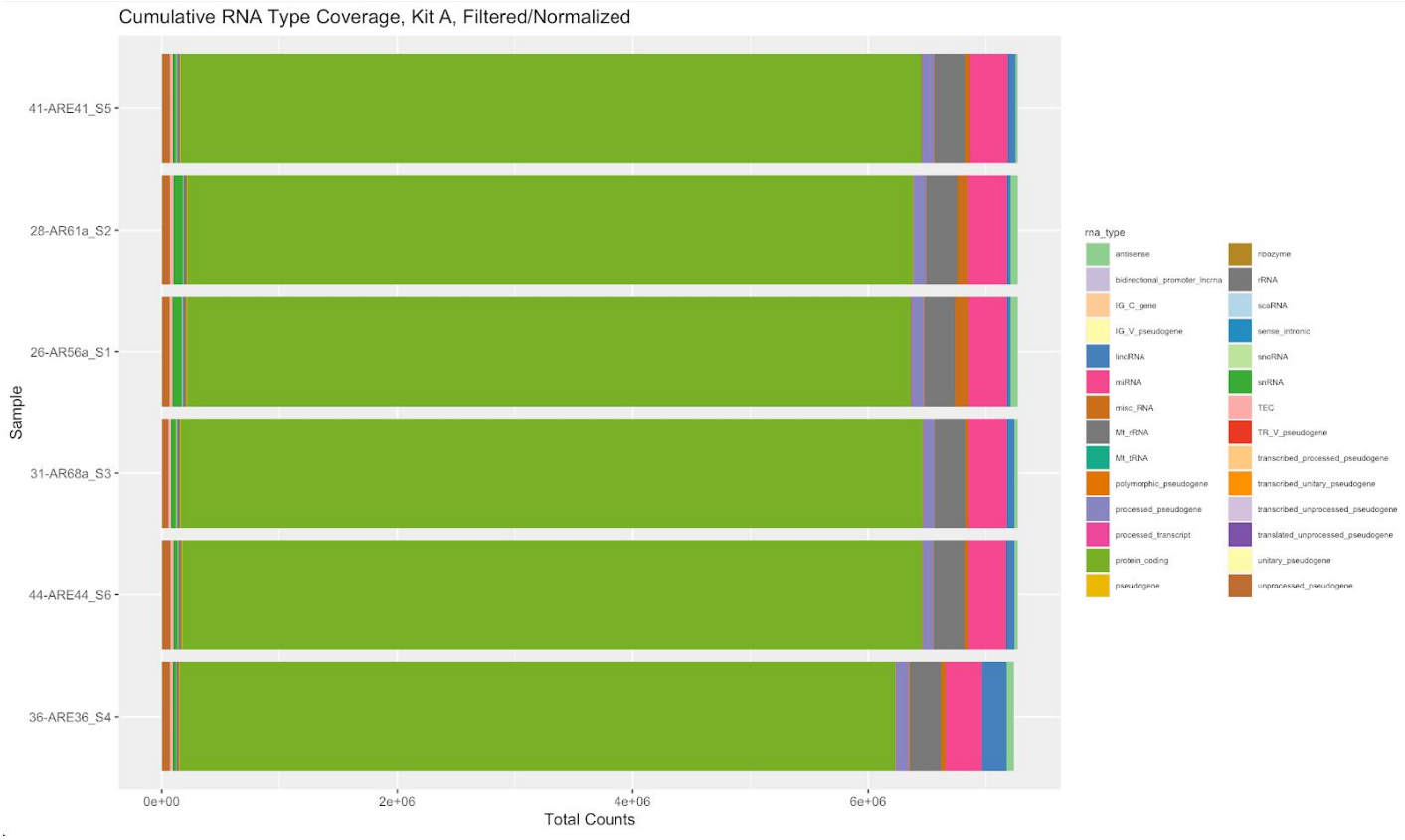

RLE, Kit C, Raw/Non-Normalized Counts
Log2 fold change wrt median
6
4
2
0
-2
-4
-6
AR56a_HT_S36
AR61a_HT_S37
AR68a_HT_S38
ARE36_HT_S39
ARE41_HT_S40
ARE44_HT_S41

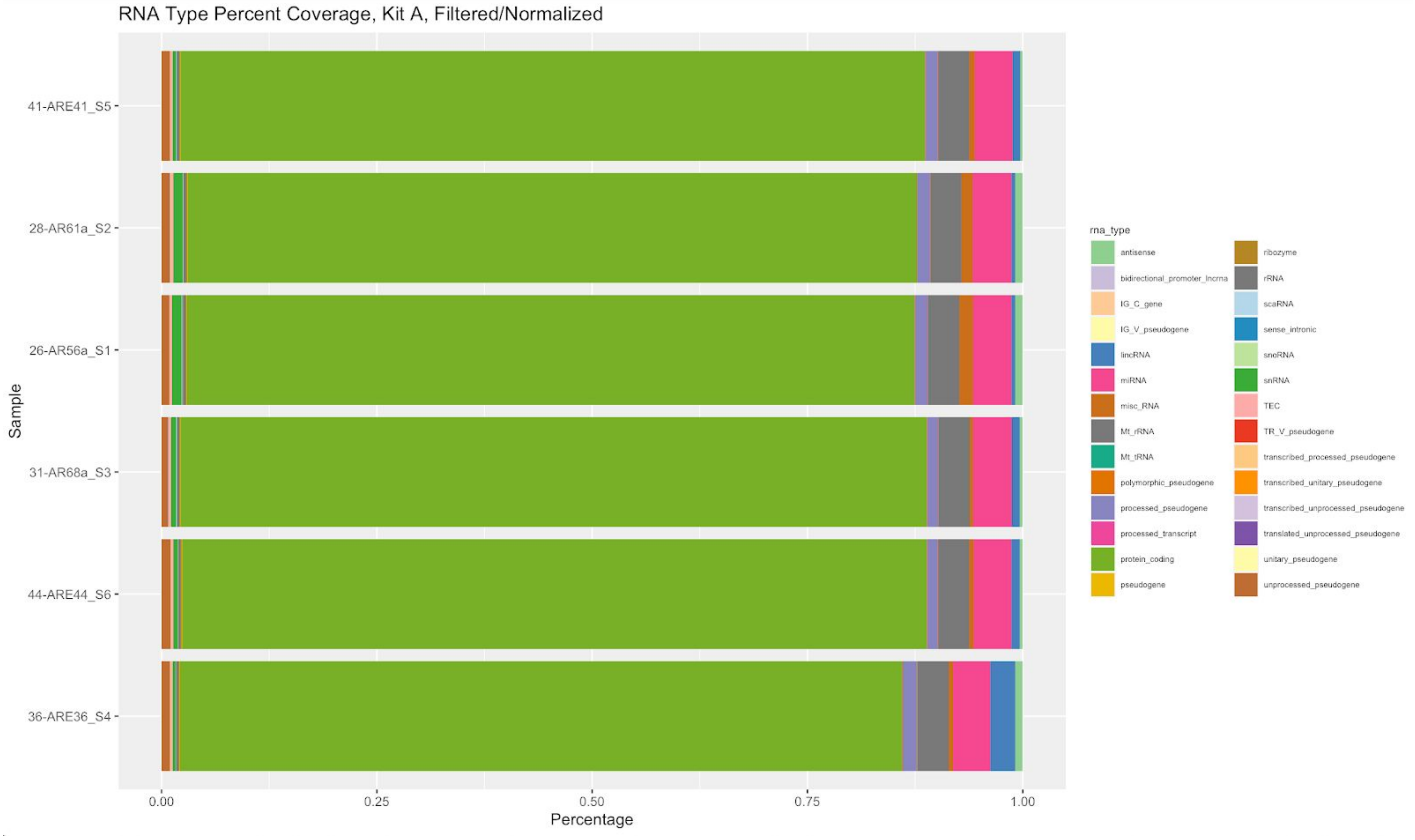

RLE, Kit C, Normalized Counts
Log2 fold change wrt median
3
2
1
0
-1
-2
-3
AR56a_HT_S36
AR61a_HT_S37
AR68a_HT_S38
ARE36_HT_S39
ARE41_HT_S40
ARE44_HT_S41

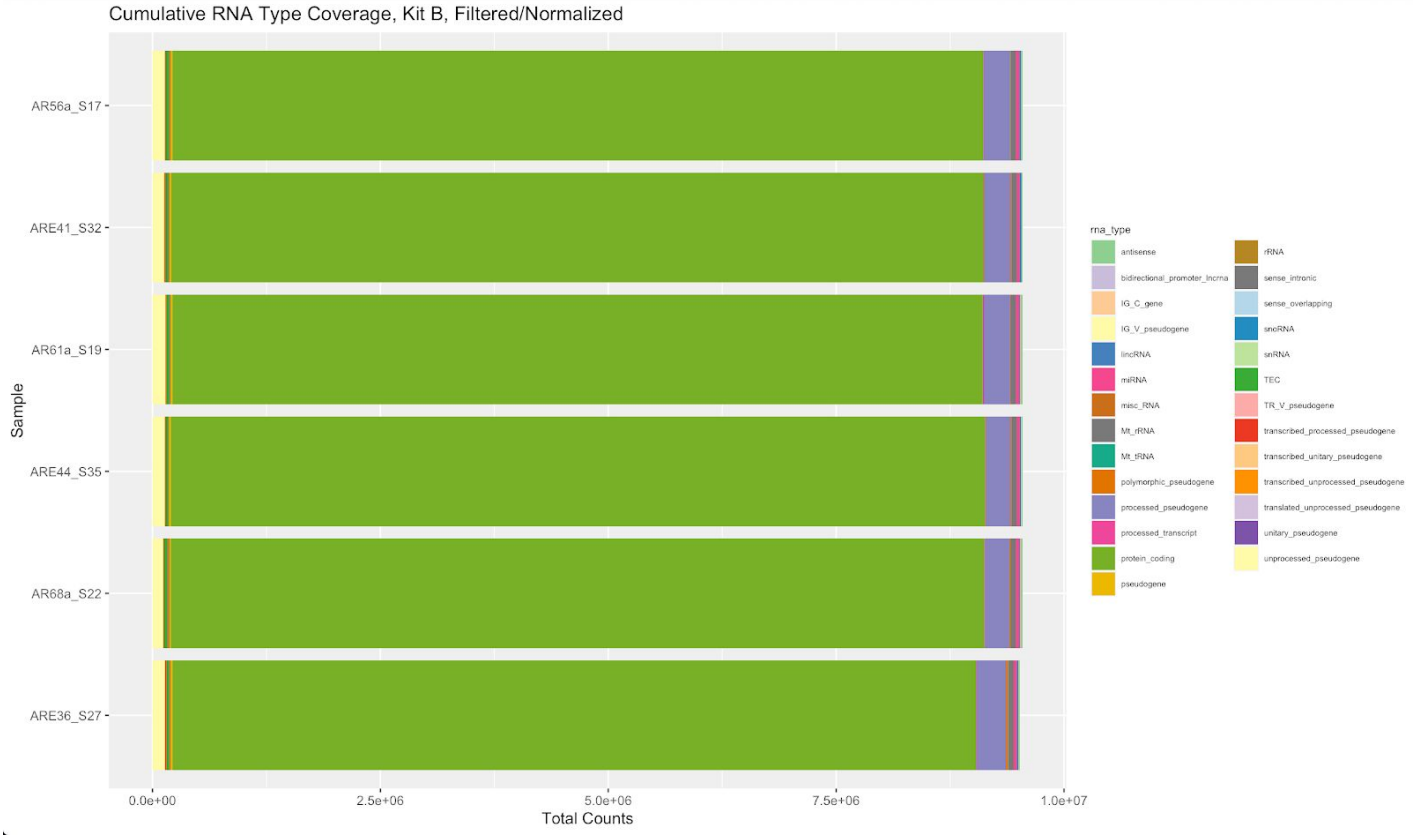
For each kit it can be observed that the majority of transcriptomic features cluster very close to and around the zero center line after normalization, with a minority of features having either a greater or smaller degree of expression as compared with the median for the kit. Prior to normalization we observe very few genes that cluster towards the median zero line, and a much larger log transformed range of expression corresponding to both highly and lowly expressed features. It can be seen that all three kits benefitted from normalization and have similarly strong RLE profiles across all of the samples within each kit. There is some within-kit variability in the RLE for certain samples; for example, for all three kit preparations it can be seen that samples corresponding to mouse type ARE36 exhibits the greatest degree of variation in gene expression values from the median, with gene features that are both more highly and lowly expressed as compared to the median and the largest range between the upper and lower quartiles. This may be due to the specific biological properties of the tissue type from which this sample was derived. Overall, viewing tight distributions for the majority of the samples within each kit suggests that the normalization and preprocessing steps for the raw count data were very effective at filtering out outlying absent genes and correcting for confounding features. It may be possible to attribute the slight variation in relative gene expression to the natural biological variability across the tissue types that were prepared with each kit.

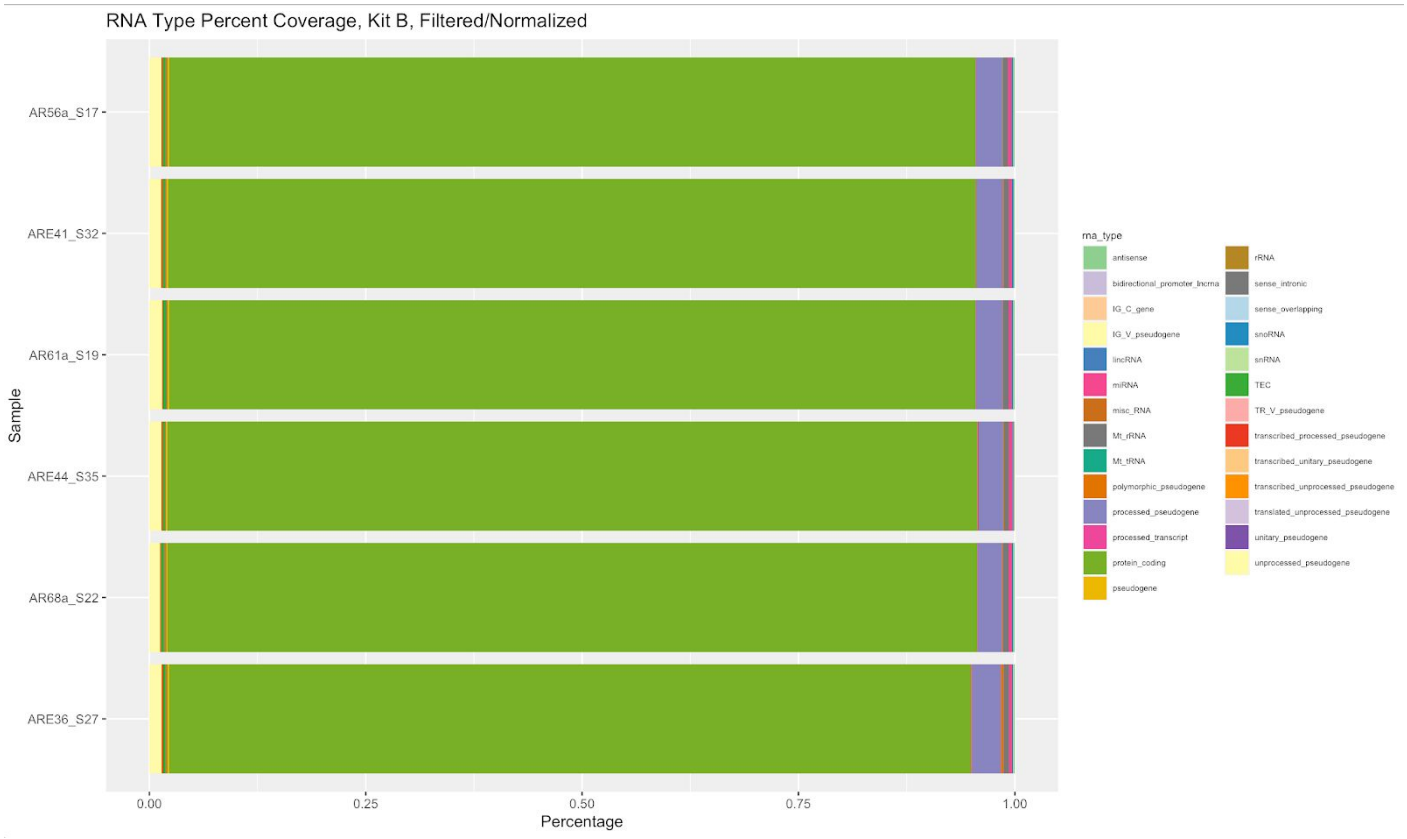
## Cumulative RNA Annotation Coverage After Normalization

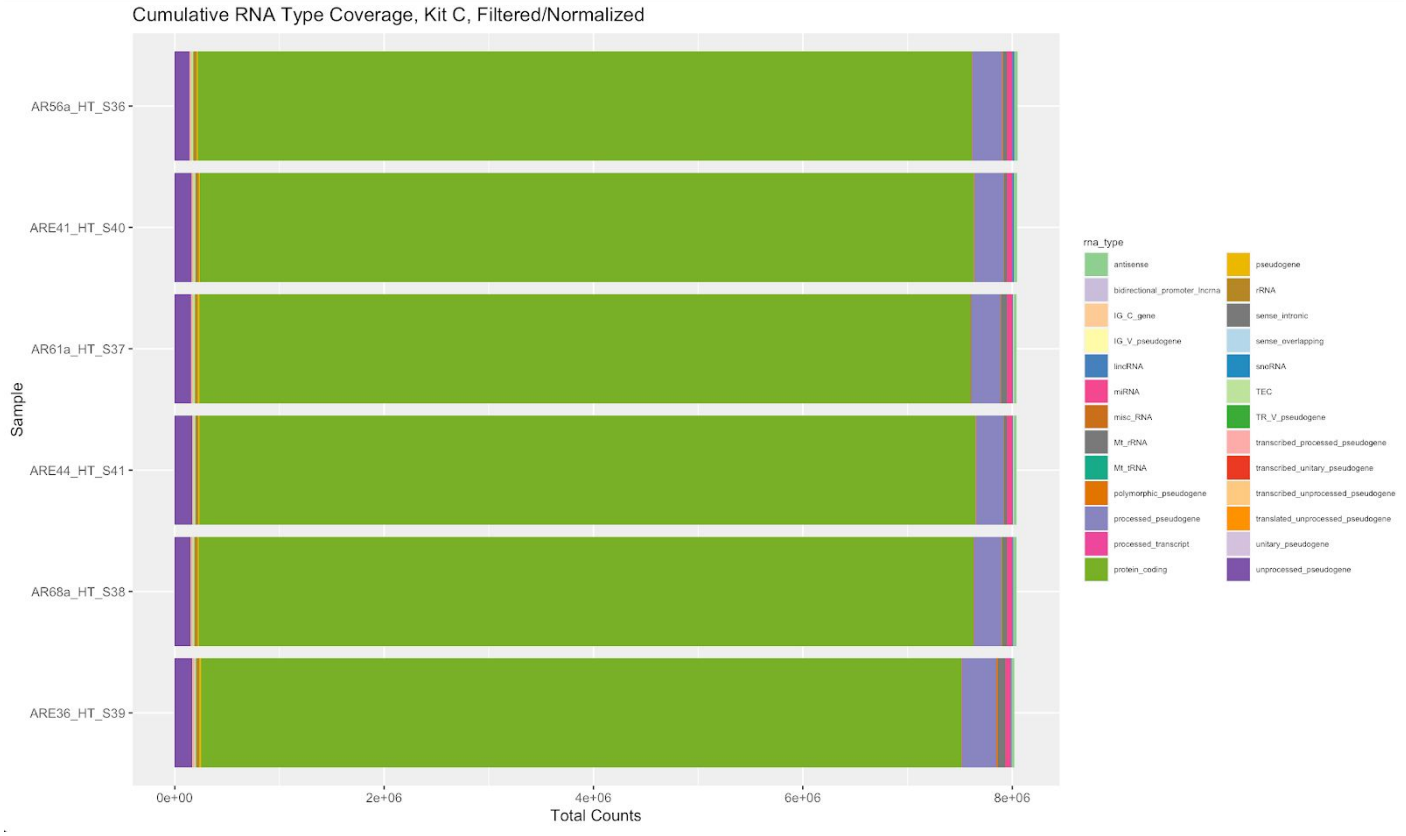
Annotation coverage of the cumulative totals of RNA species is conducted again after the normalization and filtering steps that were described above. We would expect to see correction for the relative abundance of different annotation types once GC content of various genes is taken into consideration, as well as the removal of lowly expressed genes and a correction for the sequencing depth for each library. The normalization annotation totals are shown below; similarly to the raw counts, both the total counts as well as percentage for each sample are included in order to correct for slight variations in the total read counts after normalization.

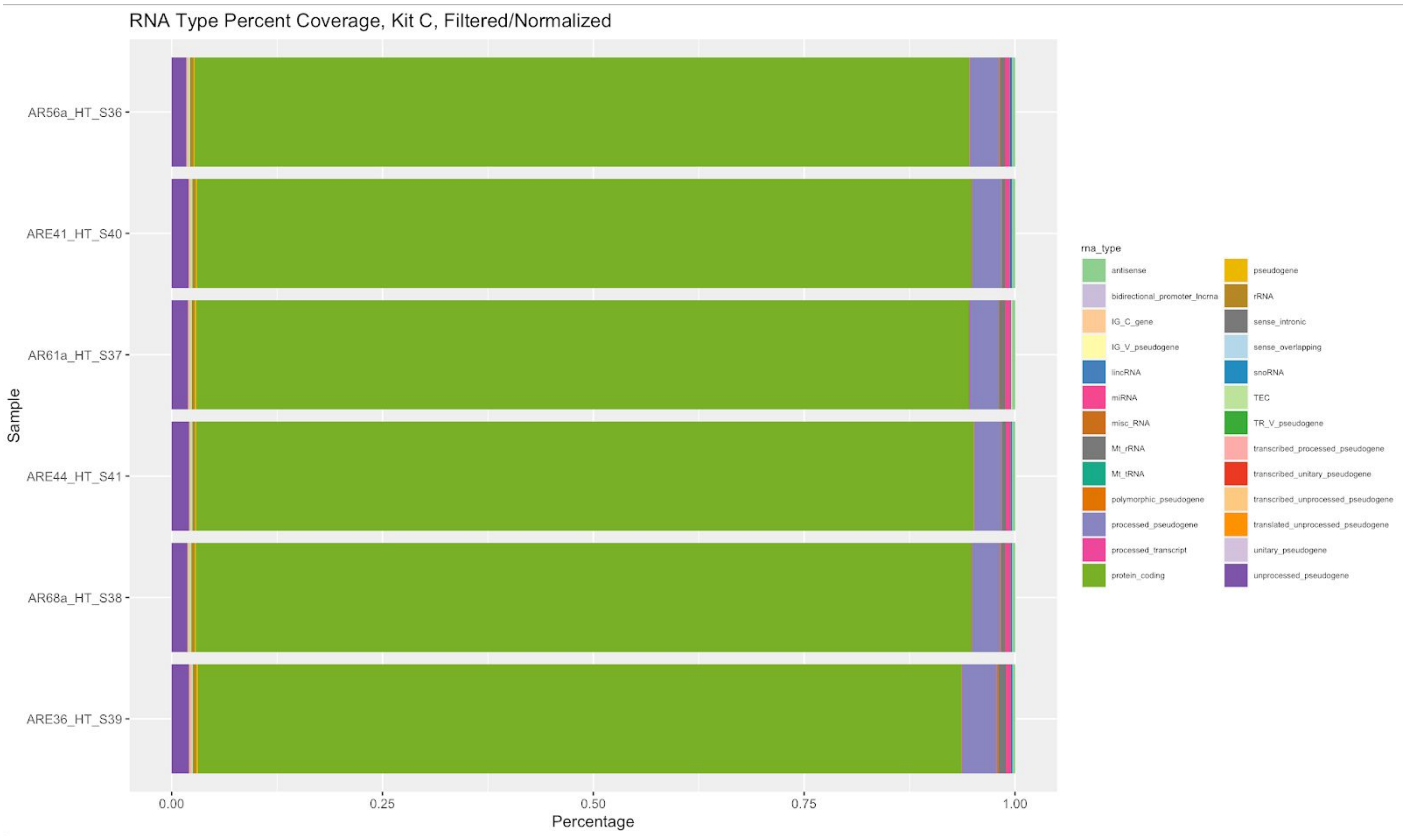



RNA Type Percent Coverage, Kit A, Filtered/Normalized
41-ARE41_S5
28-AR61a_S2
26-AR56a_S1
31-AR68a_S3
44-ARE44_S6
36-ARE36_S4
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
snoRNA
snRNA
TEC
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


Cumulative RNA Type Coverage, Kit B, Filtered/Normalized
AR56a_S17
ARE41_S32
AR61a_S19
ARE44_S35
AR68a_S22
ARE36_S27
Sample
0.0e+00
2.5e+06
5.0e+06
7.5e+06
1.0e+07
Total Counts
rna_type
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
TEC
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene


RNA Type Percent Coverage, Kit B, Filtered/Normalized
AR56a_S17
ARE41_S32
AR61a_S19
AR44_S35
AR68a_S22
ARE36_S27
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
TEC
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene

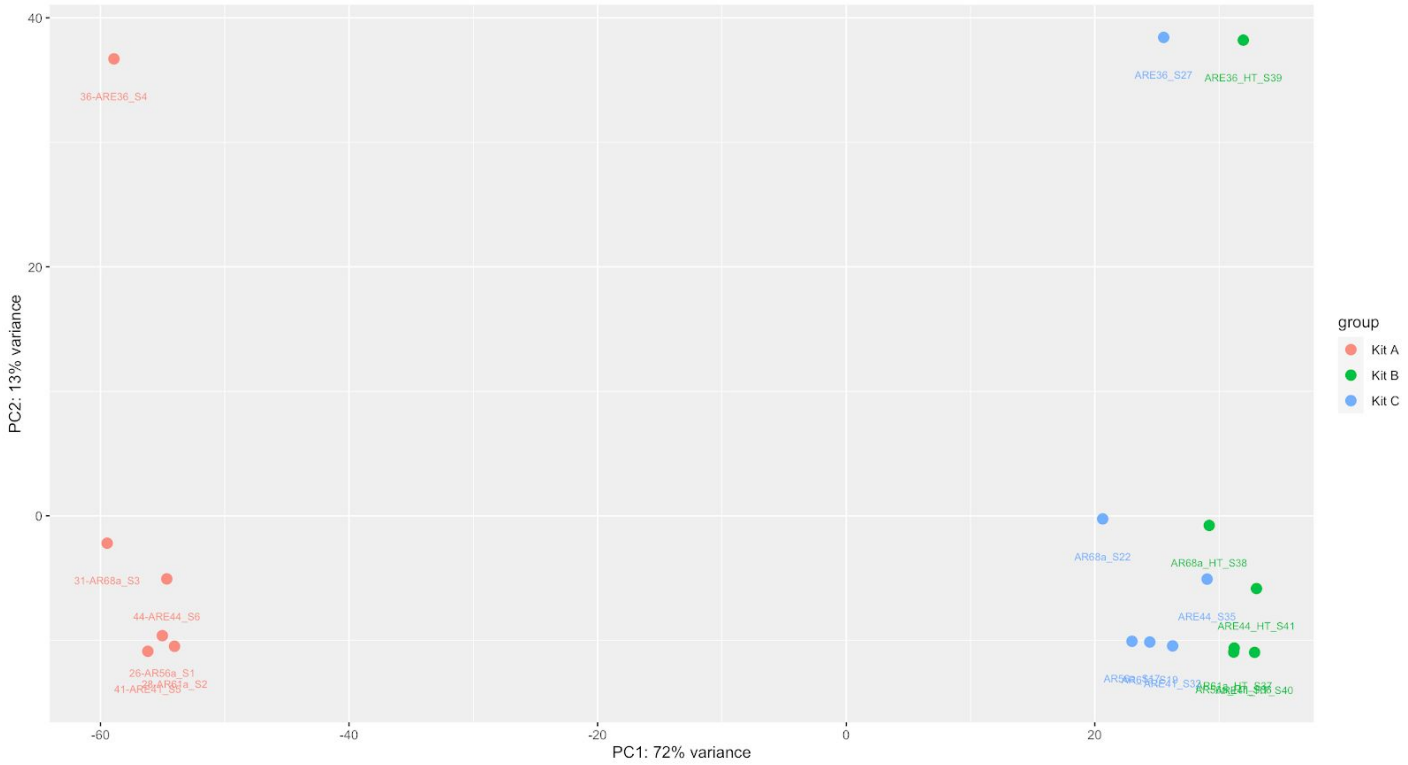

Cumulative RNA Type Coverage, Kit C, Filtered/Normalized
AR56a_HT_S36
ARE41_HT_S40
AR61a_HT_S37
AR44_HT_S41
AR68a_HT_S38
ARE36_HT_S39
Sample
0e+00
2e+06
4e+06
6e+06
8e+06
Total Counts
rna_type
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snoRNA
TEC
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene




Several observations can be made from the normalization annotation coverage graphs. The normalization procedure corrected for variations in library size/sequencing depth, so all annotation features are now expressed approximately as a proportion of the same number of reads. Furthermore, we can see a significant correction in the relative RNA abundances for Kit A libraries. Whereas prior to normalization, protein coding elements constituted approximately 50% of the total Kit A annotations, after count filtering and normalization we see that this proportion is closer to 75-80% for all samples in this category. We still see the presence and representation of the other types of regulatory RNAs that were identified in the raw counts (most notably miRNA, lncRNA and MtrRNA), but these abundances have been corrected to account for a smaller proportion of the final library contents. The normalization results suggest that there was an initial biased representation of certain non-coding RNA species in the Kit A libraries, most notably miRNA transcripts, that were likely due to artifacts of the library preparation stage that confounded the counting process. Both B and C library preparations do not demonstrate the same significant shift in relative RNA abundance after normalization as compared with Kit A. The protein coding percentages remained at approximately 85-90%, and the proportions of the next most abundant RNAs such as unprocessed and processed pseudogenes were also consistent. While their coverage profiles remain very similar, it is noteworthy that Kit C seems to capture a slightly greater number of miRNA genes as compared to its normal throughput counterpart, which becomes slightly more pronounced when looking at the normalized coverage.

**Principal Component analysis of gene expression profiles**

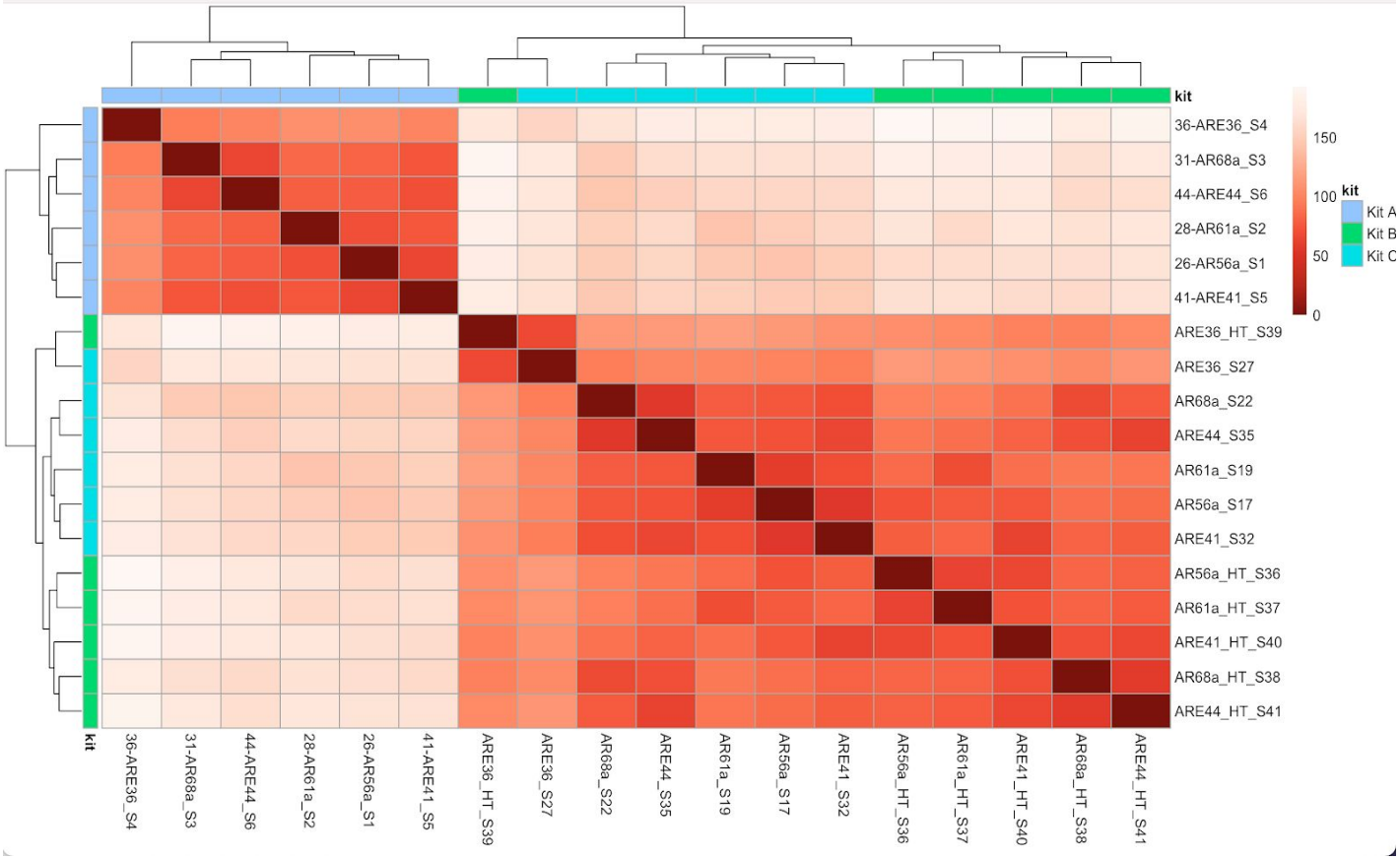
Principal component analysis (PCA) is a common dimensionality reduction technique and exploratory data analysis tool that aims to find the linear combinations of factors that explain the greatest amount of variance within a dataset. The principal components of the dataset are formed from combinations of the genes that explain the greatest gene expression variation among the samples, and are used to create an orthogonal projection of the dataset into a new vector space where all principal components are uncorrelated and orthogonal to one another. This method proves very useful in bioinformatics where you often have many more components and features (i.e. genes) than samples in a dataset. Using this technique it is possible to find the genes that contribute the greatest amount of variance in your dataset in just the first few components. A principal component plot for all 18 unique samples is shown below. This plot represents the samples plotted in relation to how much of each component 1 and 2 contribute to their expression profiles using the normalized data.




The PCA plot reveals that there exist two components composed of linear combinations of certain genes that can account for 85% of the variance of the count data for all of the samples. This value represents a significant portion of the variance and proves to highlight the exploratory power of PCA to reduce data dimensionality for visualization. By plotting each sample by kit, we can easily identify clustering patterns of samples that have similar gene expression profiles based on the genes included in the top 2 components. It becomes obvious that Kit A libraries cluster close together with characteristic values for both components 1 and 2, with the exception of one outlier. Furthermore, both SMARTSeq kit options possess fairly similar gene profiles, again with one outlier per sample. Overlaying the sample names reveals that the sample that demonstrated the greatest gene expression variance as measured previously by RLE, ARE36, is the outlier in PCA. This further reinforces the hypothesis that this sample has a greater degree of biological variability in gene expression as compared to the other 17 samples.

**Euclidean heatmap distance for gene expression profiles**

Another method of demonstrating the similarity among samples based on gene expression is to visualize the Euclidean distances between pairs of samples, measured as the square root of the sum of differences for all of the gene count features between two libraries. Using this method, a heatmap of the pairwise distances for all 18 samples can be generated, and a simple clustering of similar gene profiles can be made, with similar samples having a smaller Euclidean distance from each other. In order to generate an effective heatmap, a log transformation of the normalized counts is created before computing the distances, as a way of ensuring that the heatmap measurements are not dominated by a few highly expressed genes. This ensures that each of the gene features contributes roughly equal proportions to the Euclidean distance measurements. The heatmap was generated using the DESeq2 package v. 1.28.1 from Bioconductor after a DeSeq2 object is created using the normalized counts produced from EDAseq as described above. The kit categorization for each sample is used for annotation for both the rows and columns (which are identical designations).




Results from the heat map are consistent with the general clustering patterns that were seen with PCA. As expected, libraries produced with Kit A cluster together with similar expression patterns, and both Kit B and C form another larger, more general cluster with overlapping similarity in gene expression profiles. It is interesting to note that while most libraries within each category (Kit B or C) are more similar to each other than to the other libraries in the B-C category (i.e. all of the libraries in Kit B are more similar to each other than each of them is to a library in Kit C), one biological sample forms an outlier, in which a library each from Kit B and C are more similar to each other than to the other libraries in their kit group. This sample is ARE36, which is reflected in the clustering patterns in the heatmap; ARE36_HT_S39 and ARE36_S27 form a similarity grouping before each of them groups with the other libraries in the B-C group. This heat map further confirms the hypothesis that this particular sample has a gene expression profile that deviates more significantly from the other samples in this dataset. Using clustering gene expression data visualization techniques such as PCA and Euclidean distance, we can readily identify trends in gene expression both within kit groupings, as well as outside of them in terms of unique biological variability at the sample level.

## Matching the anonymous kits to a library preparation

Based on the cumulative analysis conducted in the report, including the various quality control checks and annotation coverage/gene expression profiles that were created, the following kit descriptions were designated to each anonymous kit, with a brief explanation as to the rationale for the designation given the result of the aforementioned analyses.

**Kit A -> *Takara SMARTer Stranded Total RNA***

- Annotation coverage for Kit A revealed the capture of several categories of regulatory and non-coding RNA species such as lincRNA, miRNA, and MtrRNA, which is typical of a total RNA preparation. These species would not be captured in such abundances using an mRNA oligo-based approach.

- The stranded option during featureCounts of 2 corresponds to a reversely stranded library preparation. The SMARTSeq total RNA option was the only kit included in this project that retained information on the strand orientation of the transcripts during library prep.
- SMARTer libraries are not optimized for extremely low input starting amounts below 10ng. As such one would expect to see a greater degree of read duplication from inputs below 10ng in order to produce sufficient library material for flow cell loading for sequencing. This level of read duplication is much higher than either Kit B or C was confirmed with FastQC
- Quality control and trimming of the reads indicated a greater number of reads that were too short for passing to alignment, as well as a greater proportion of reads that contain contaminating adapter sequences and the 3' end. These read characteristics are typical of libraries that are not optimized to convert low starting RNA amounts into libraries with fragments of a sufficient length

**Kit B -> *Takara SMARTSeq v4 mRNA***

- Annotation coverage both before and after normalization for Kit B indicates a very high proportion of coding genes/regions that is indicative of mRNA capture for coding transcripts. Furthermore the next most highly represented transcript categories are the various pseudogenes (processed and unprocessed), which are poly-adenylated transcripts. This helps to confirm the method of RNA capture for Kit B through an oligo-dt binding step to poly-A-tails.
- Levels of read duplication are moderate, which is consistent with the kit chemistry for mRNA Ultra Low Input to optimize performance for very small amounts of starting material
- A featureCounts stranded option of 0 is selected for this kit, corresponding to a kit that does not retain strand information. This corresponds to the chemistry of the mRNA capture used in SMARTSeq.
- Read trimming and filtering steps revealed a greater degree of read trimming due to polyadenylated sequences. This suggests a higher proportion of reads with poly-a-tails which is again reflective of oligo-dt bead capture kits that target poly-a-tails for enrichment.

**Kit C -> *Takara SMARTSeq v4 mRNA High Throughput (HT)***

- The rationale for this kit designation is very similar to the observations made for Kit B in terms of annotation coverage, read duplication rates, kit strandedness, and read trimming. The kit chemistry corresponding to the high throughput version of SMARTSeq mRNA is almost identical to the manual throughput version, so these findings were expected once the identity of Kit B was established
- Differentiation between Kit C and C was made possible by the inclusion of an "HT" designation for the high throughput library names in the dataset

## Discussion and Interpretation of Kit Functionality

The determination of the functionality and fidelity of each of the kits described in this project needs to include a comprehensive assessment of all of the metrics described in this report. Due to the flexible nature of transcriptomic studies and library prep chemistries, a single RNA preparation kit by itself is likely not going to be effective for all possible RNA-seq experiments; it is more likely that a specific preparation will provide high-quality data for a project given a specific context, but perhaps not for others. Therefore the consideration of the particular research question that aims to be addressed by a particular RNA-seq experiment should be used in order to guide the selection of a particular kit. For example, it can be seen that libraries prepared using SMARTer Total RNA are able to retain moderate numbers of non-coding and regulatory RNA species such as lincRNAs and miRNA transcripts. Both SMARTSeq mRNA preps do not retain these same species, due to their particular chemistry of targeting polyadenylated mature mRNA transcripts and ignoring those without this feature. In the event that a researcher needs to be able to characterize both protein coding and non-coding regulatory genomic features, use of either SMARTSeq mRNA option may result in read sequences that are not of biological relevance to the research question at hand. It may be more effective to use a total RNA option, and while the SMARTer libraries do exhibit some deficiencies in terms of high read duplication rates and outlying/biased genomic features, this kit should probably be recommended over an mRNA prep in this instance. It is also important to

appreciate that mRNA capture methods do not typically work well with degraded samples because of the difficulty in capturing degraded poly-A-tails. In the event that a researcher wishes to produce robust libraries from degraded samples, the TruSeq Stranded kit option is likely the only preparation that can achieve this result

Conversely, many transcriptomic studies focus on canonical differential gene expression analysis (DGE), where measurements of protein coding features are compared across different biological states and conditions. In an instance such as this one, the use of either SMARTSeq mRNA option becomes encouraged, as these kits seem to provide consistent, high-quality protein coding annotations that can be useful for DGE; furthermore, these preparations have acceptable levels read duplication and even GC content distribution, both of which should contribute minimal bias to the results. The researcher may not be particularly interested in the presence and/or abundance of regulatory non-coding elements, and as such, SMARTSeq mRNA is optimized to deliver information on protein coding/exonic regions over other RNA species, particularly for very small starting amounts of RNA.

A very important consideration for any RNA-seq experiment is the ability for a particular preparation to produce a representative transcriptome landscape; that is, to generate a set of transcript abundances that is reflective of the actual biological components within the sample at any given point. This consideration may often be overlooked during NGS data analysis, as a researcher may be interested only in high and differential abundances of specific RNA species relevant to his/her research question, and may therefore neglect to appreciate any biases that arose to produce those results. The conversion of RNA molecules into a library that is compatible with Illumina sequencing will inherently generate artefacts and biases that cause the library to stray from the true RNA content of the cell. It is therefore essential to appreciate and try to mitigate these sources of bias as much as possible, in particular during the library preparation stages.

For these reasons it should be noted that libraries produced from low-input material may especially suffer from artificial biases during library prep, and therefore may require extensive processing and filtering steps to reduce these biases. This was seen with the SMARTer total RNA libraries, as normalization was required in order to produce an expression profile that was consistent with the kit chemistry and comparable across different library groups. The need for normalization for this preparation may indicate the limitations in this kit in being able to convert representative abundances of total RNA into a robust cDNA library from very small input amounts. Therefore, in the event that non-coding RNAs need to be assessed from low-input material, the appropriate read depth should be adjusted for this kit and data analysis should include the effects of normalization on the relative abundances of these RNA species of interest. Additionally, high levels of read duplication, as well as proportions of the alignment that remained unassigned to genomic features, need to be considered in terms of “bang for your buck” sequencing; the percentage of final reads that have some biological relevance to your study. Ideally, a researcher should be able to select a preparation kit that provides a high degree of usable transcript data from robust, reproducible libraries that target the specific RNA species that are to be examined in the experiment.

## Conclusions

Through a combination of canonical NGS quality control tests, annotation coverage and gene expression analysis, it is possible to ascertain the identity of an RNA library preparation kit by evaluating anonymous FASTQ files. The difference in RNA species retained by certain kits provides very good evidence to highlight the difference between total RNA and targeted/mRNA preparations, and reveals a list of considerations that should be taken when planning an RNA-seq experiment with low input samples. Furthermore, the various quality control metrics that are collected at both the read and alignment level can be very informative, and can serve to confirm the unique characteristics of libraries produced from different RNA capture kits. Taken together, this project highlights an RNA-seq data analysis pipeline that is comprehensive and complete in its ability to stratify RNA libraries by kit preparation.

## Appendix A: Improving Alignment Rates for Kit A

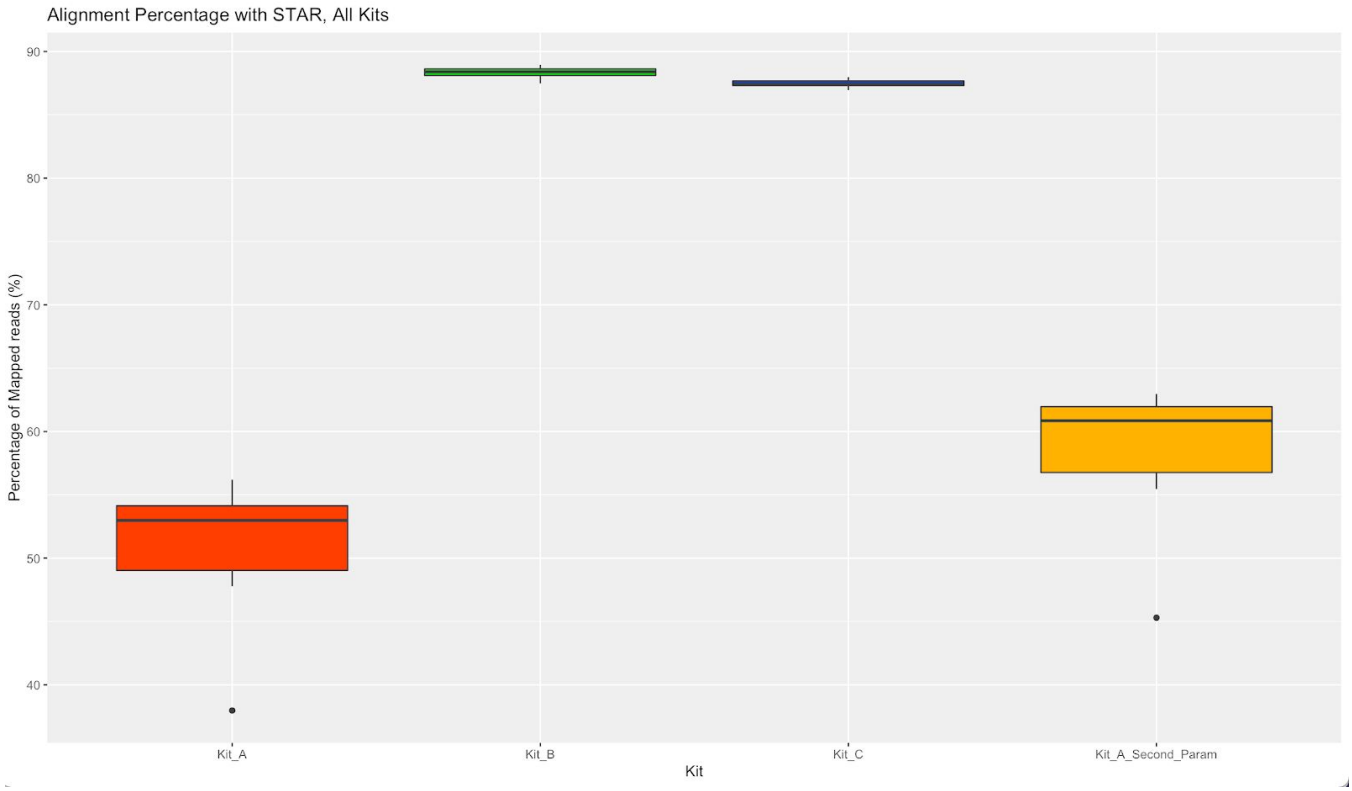
The percentages of successfully aligned reads using STAR were particularly low for Kit A libraries as compared with the other library preparations. Therefore, an adjustment to the hyperparameters for STAR was considered in an attempt to increase this percentage and derive more usable read data from Kit A libraries.

In addition to the parameters that are specified in the section titled *Alignment*, the following parameters were modified for he filtered and processed FASTQ samples from Kit A:

```
--outReadsUnmapped Fastx \
--seedSearchStartLmax 15  \
--outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0  --outFilterMatchNmin 50
```

The modification of the seedSearchStartLmax parameter from the default value of 50 to 15 allows for smaller portions of spliced reads to be aligned to the reference genome using STAR. Since our previous steps permitted the shortest read length retained after read processing and trimming to be 35 bp, we allow for the approximate even splitting of these reads to produce mappable fragments between 15-17 bp long. The previous default parameter of 50 excluded any read fragments that after splicing, were shorter than 50 bp, and discarded these reads at the alignment step. It is expected that since Kit A retains many RNA species that have shorter read lengths, as well as produce read fragments from degraded and low-input material that have shorter average lengths than 100bp, the adjustment of this parameter will allow for shorter spliced read fragments to successfully map when they wouldn't under the default parameters. Repeat alignment of Kit A libraries was conducted under these parameters, and the annotation procedure was repeated to see if the unique mapping percentage could increase while simultaneously maintaining or increasing the percentage of features counted using featureCounts.

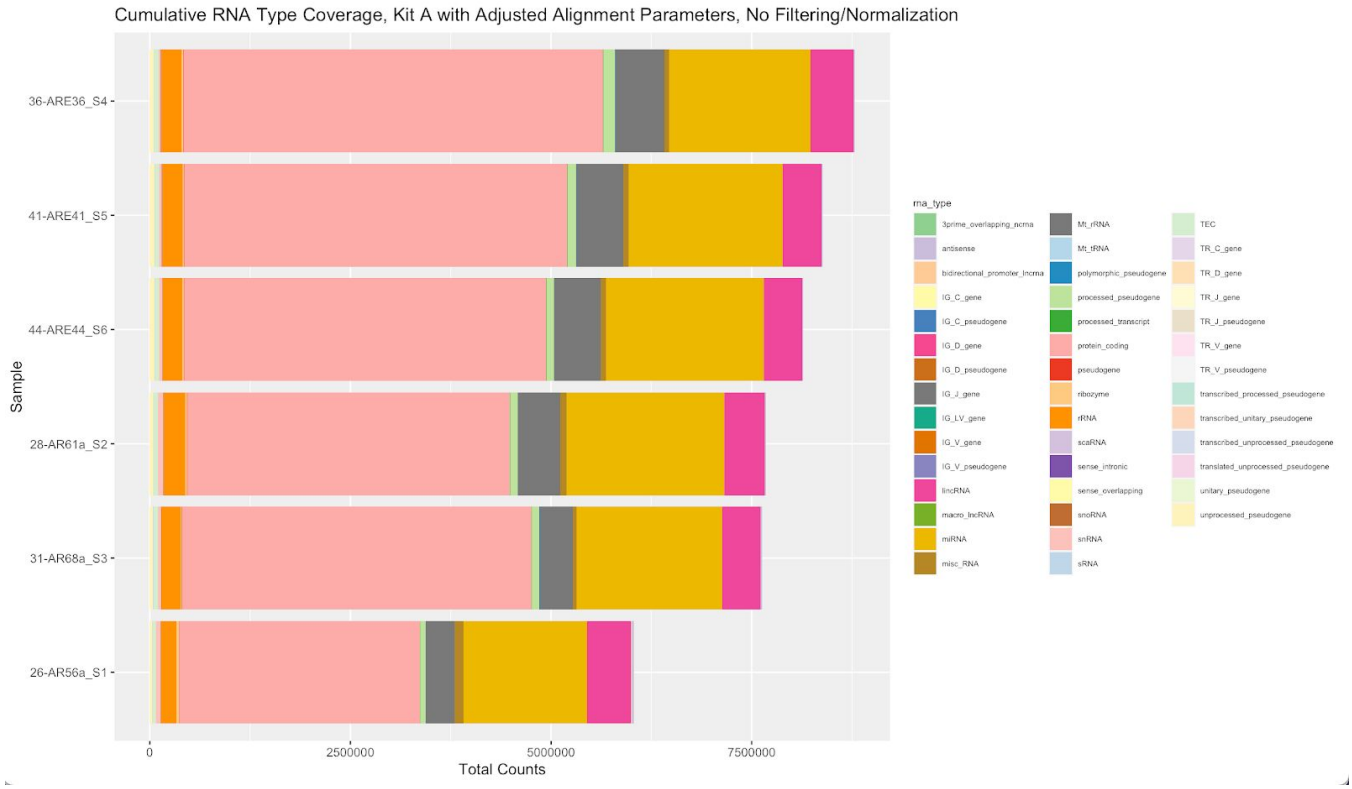
An updated boxplot of the alignment rates for the 3 kits, including all of the samples in Kit A with the newly adjusted parameters (designated as *Kit_A_Second_Param)*, is shown below.

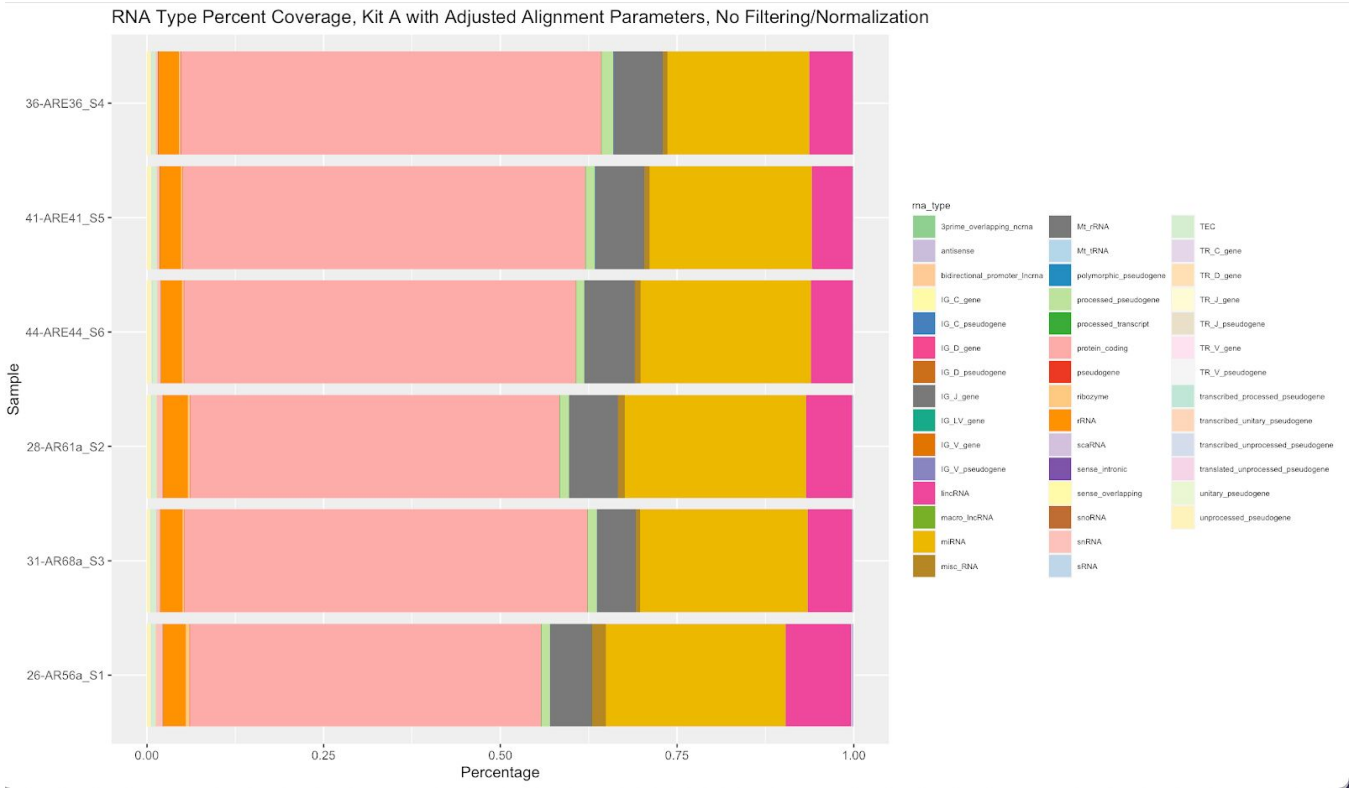


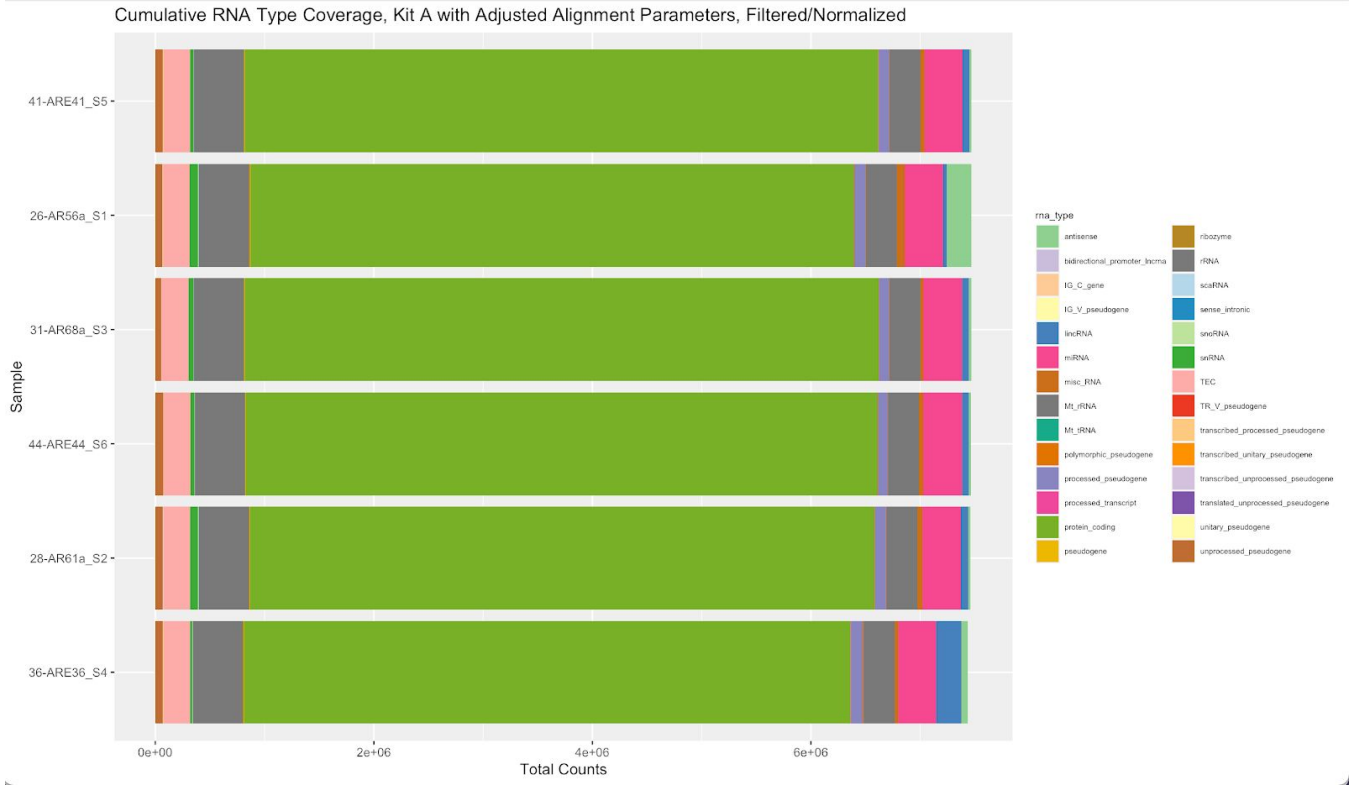
It can be seen from the updated boxplot that the adjustment of the STAR alignment parameters for Kit A increased the overall mapping rate for the 6 samples by 8-9% on average. While these percentages were not drastically increased to a mapping success rate that is comparable to either Kit B or C (> 87%), the increase in the percentage of successfully aligned reads may have significant implications for the ability to cover and annotate the various RNA species that can be captured with Kit A. Therefore the individual and cumulative annotation coverage was repeated for these libraries after re-alignment, as the proportions of different RNA species counted using featureCounts may have changed noticeably.

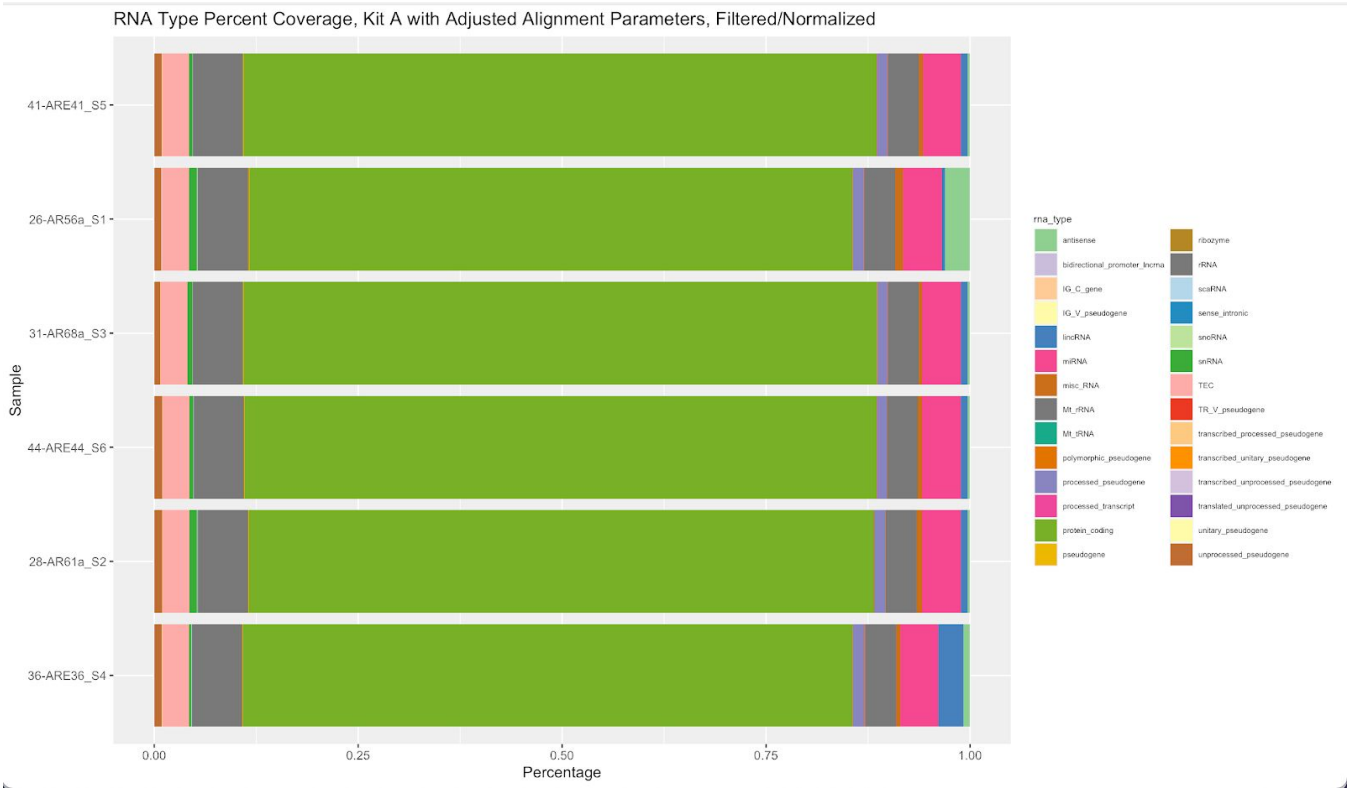
It should be noted that, in addition to increasing the unique mapping rate, the percentage of multi-mapping reads with the adjusted parameters also increased substantially to approximately 10% per sample (not shown in the figure). This is a predictable outcome when the adjustment of the parameters permits the alignment of smaller spliced read portions; when smaller read portions are permitted to align, they are more likely to correspond to multiple genomic locations because of the reduction in complexity and increase in alignment ambiguity with smaller fragments. It is hypothesized that permitting smaller read fragments to map in a splice-aware manner allows for mapping to regions of the genome that are more repetitive and therefore have a lower degree of genomic complexity.

Barplots of the cumulative and coverage percentage of the transcriptomic elements were again generated for the re-aligned Kit A samples in a manner consistent with the previous figures in the section titled *Cumulative RNA type Capture Coverage (No Gene/Feature Normalization)* and *Cumulative RNA Annotation Coverage After Normalization* respectively. Both the raw and normalized feature counts are shown below. It should be noted that boxplots for the relative expression of all mapped features for both the raw and normalized counts, similar to the figures produced in *Summary of Relative Transcript Expression Levels (Normalized),* were not reproduced, as these figures essentially look identical to the plots generated with the initial STAR parameters, and any changes to the specific identity of the RNA species captured in the new alignment procedure cannot be visualized using RLE plots.




RNA Type Percent Coverage, Kit A with Adjusted Alignment Parameters, No Filtering/Normalization
36-ARE36_S4
41-ARE41_S5
44-ARE44_S6
28-AR61a_S2
31-AR68a_S3
26-AR56a_S1
Sample
0.00
0.25
0.50
0.75
1.00
Percentage
rna_type
3prime_overlapping_ncrna
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_D_pseudogene
IG_J_gene
IG_LV_gene
IG_V_gene
IG_V_pseudogene
lincRNA
macro_lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
sRNA
TEC
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene

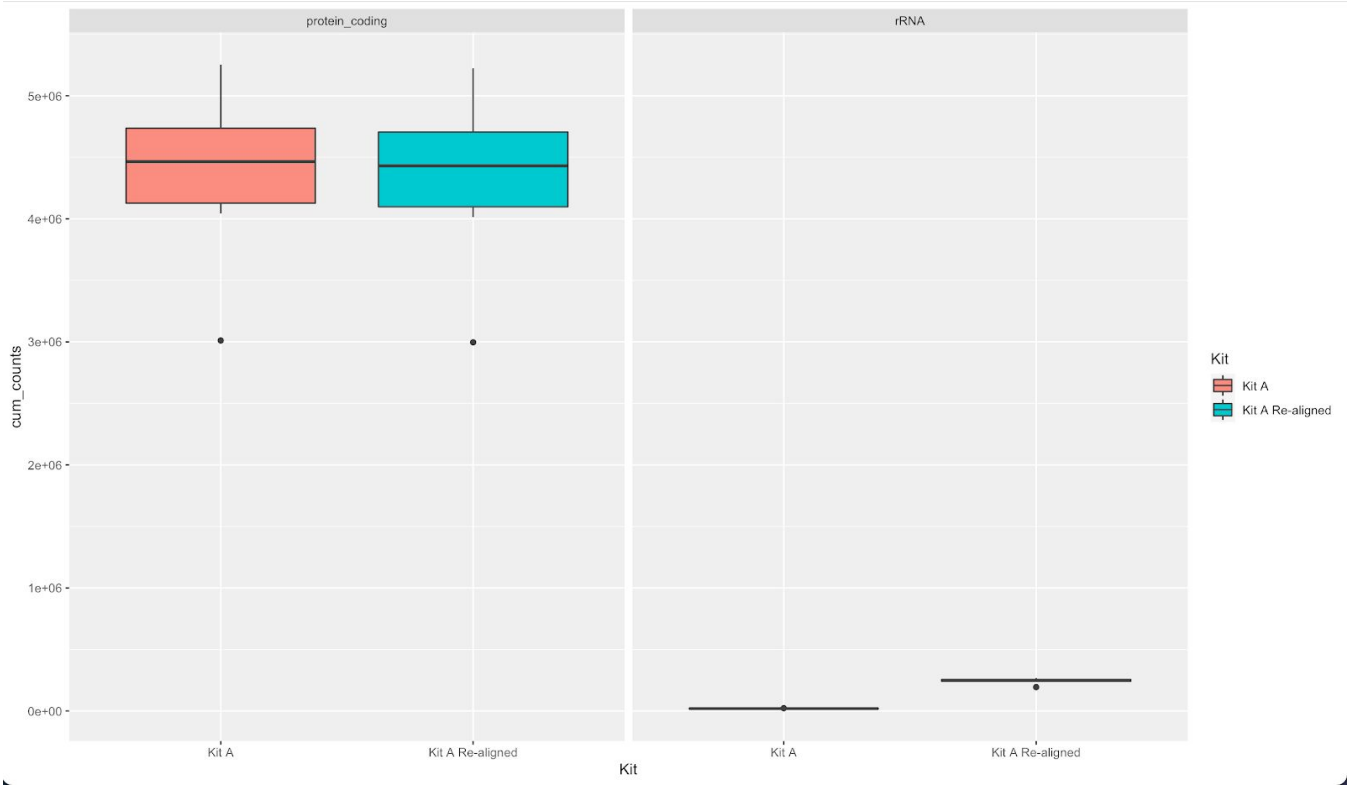

Cumulative RNA Type Coverage, Kit A with Adjusted Alignment Parameters, Filtered/Normalized
41-ARE41_S5
26-AR56a_S1
31-AR68a_S3
44-ARE44_S6
28-AR61a_S2
36-ARE36_S4
Sample
0e+00
2e+06
4e+06
6e+06
Total Counts
rna_type
antisense
bidirectional_promoter_lncrna
IG_C_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
polymorphic_pseudogene
processed_pseudogene
processed_transcript
protein_coding
pseudogene
ribozyme
rRNA
scaRNA
sense_intronic
snoRNA
snRNA
TEC
TR_V_pseudogene
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
translated_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene




The re-aligned Kit A samples demonstrate an increased percentage of aligned reads to ribosomal transcripts as evidenced by the cumulative and percentage profiles for both the raw and normalized counts, with the remaining RNA species proportions staying relatively similar to the first alignment procedure. These findings are consistent with analysis of the unmapped read sequences for Kit A, where the top overrepresented sequences that failed to align were subjected to a BLASTn search for local nucleotide similarity using the NCBI standard megablast database. The search results for each of the Kit A samples revealed hits that overwhelmingly corresponded to highly conserved ribosomal subunits and ribosomal precursor transcripts for both *M. musculus*, as well as other species. These findings therefore suggest that the majority of short, unaligned reads in the initial alignment procedure for Kit A correspond to rRNA transcripts, and the process of permitting shorter spliced read fragments to align with STAR produces a greater number of reads that align to rRNA sequences. These changes in read coverage and transcriptomic profiles are of importance to researchers because the alignment of RNA-seq reads to rRNA sequences are considered to be 'contaminating' features, and are generally not of biological importance in transcriptomic studies. Apart from very specific cases where ribosomal sequences can infer evolutionary history and phylogenetic similarity, rRNA sequences are seldomly retained for RNA-seq downstream analysis. In this instance, modifying the alignment parameters may have given only a marginal increase in the number of 'usable' RNA species for analysis, such as protein coding genes, while drastically increasing the number of aligned reads that correspond to these undesirable ribosomal features.

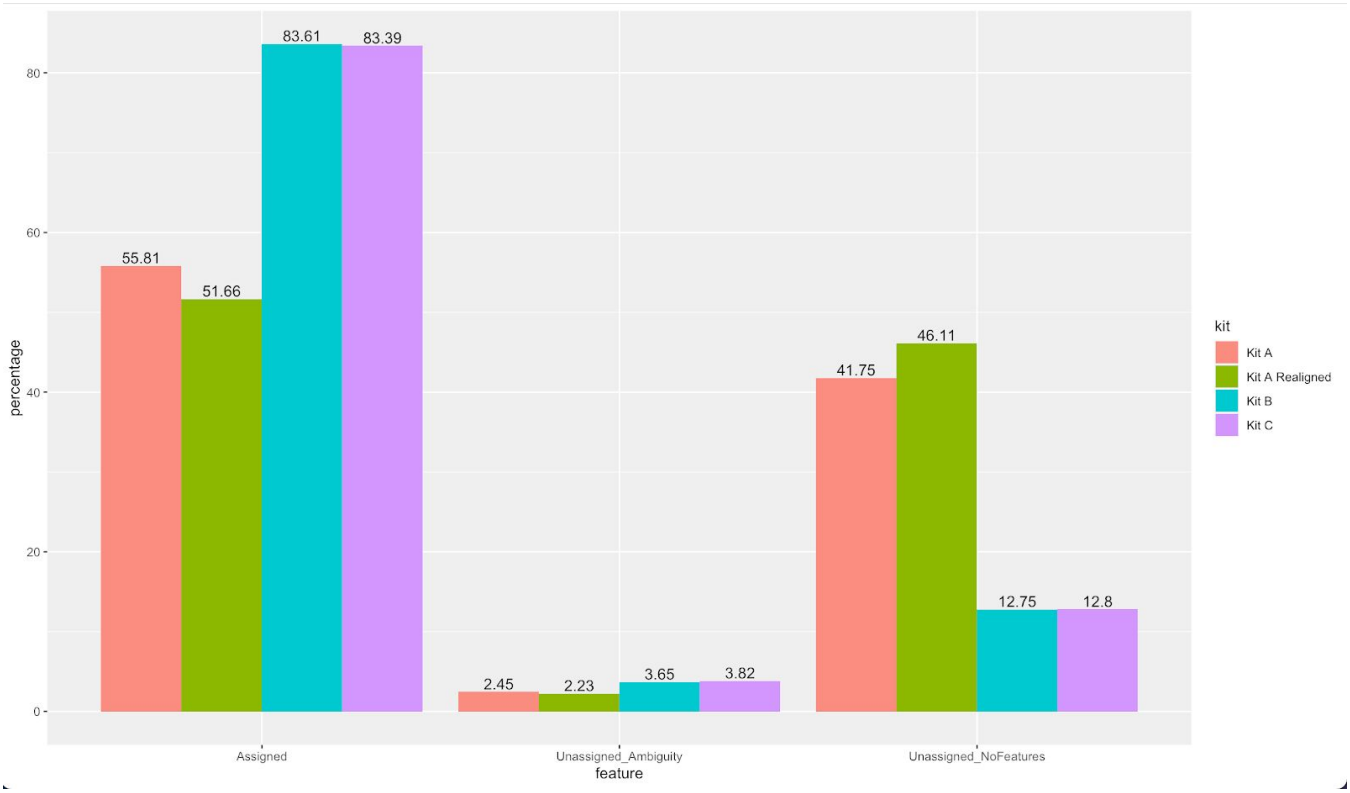
An evaluation of the average number of counts that are attributed to each RNA type serves to demonstrate the effects of the alignment parameters adjustment. When the mean counts for each RNA type category are plotted for Kit A with and without the alignment adjustment parameters, one can easily visualize any changes that the alignment modifications had to the overall capture rate of the different categories based on total, non-normalized counts, as seen below.

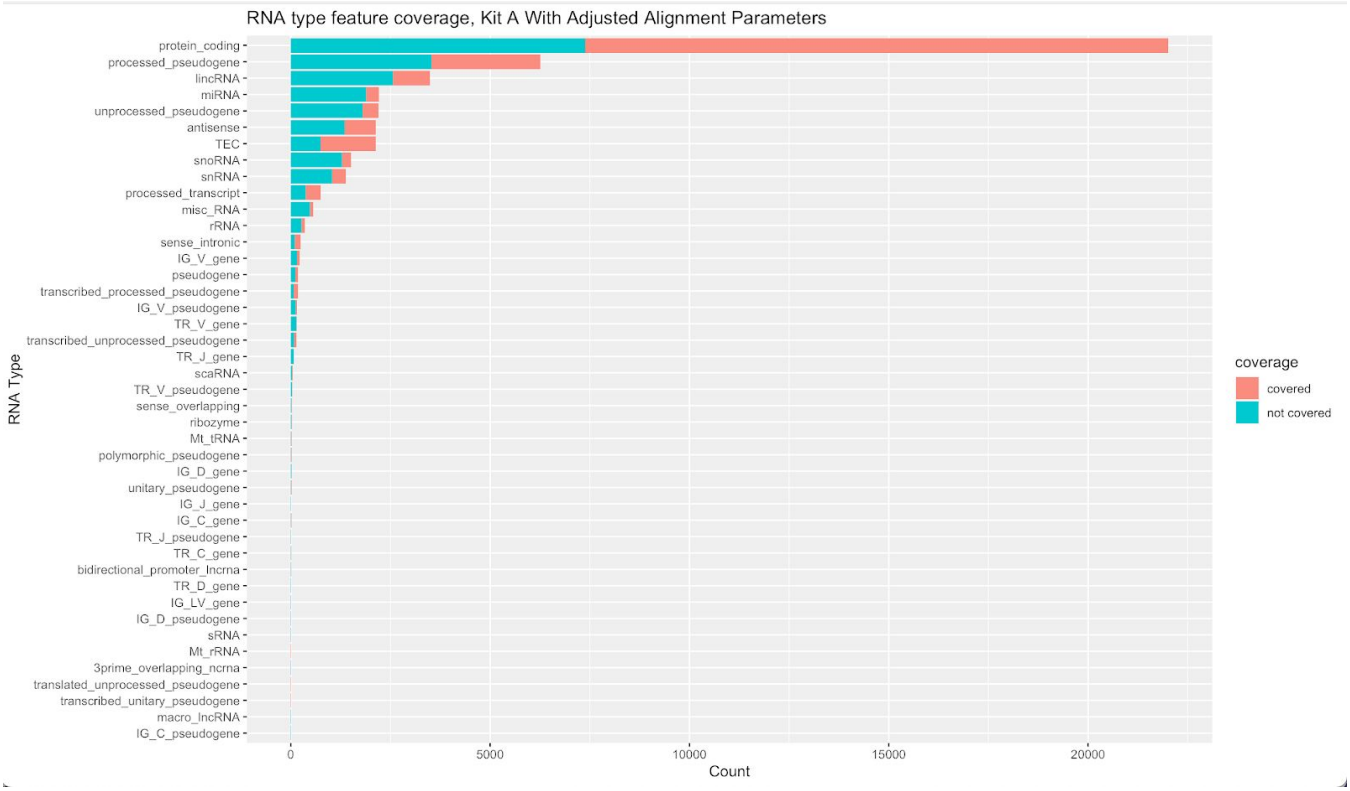

Using a subset of the above categories, it can be seen that the alignment parameters decreased the average number of protein coding read counts very slightly, while noticeably increasing the ribosomal read counts per Kit A sample. Below is a similar figure, but each box represents the total read counts for all samples, not merely the mean counts per kit.



The mean RNA species counts for the Kit A samples with and without the alignment parameter adjustment indicates that the modification of the alignment did not benefit the ability to capture an increased number of protein coding genes. Furthermore, it caused an increase in the capture of rRNA read counts.

A comparison of the assignment categories with the realigned samples to the original pipeline for each of the three kits indicates that the modification of the STAR alignment parameters causes a slight decrease in the number of assigned elements from the alignment files using featureCounts, as shown in the figure below. This suggests that permitting smaller spliced reads to align has caused a slight increase in the number of alignments to features that are not annotated in the GTF file. These values are of interest because an increase in unique alignments that results in a decrease in assigned alignments using featureCounts may act together as offsetting modifications; that is, the net result of these changes produces gene counts that are approximately the same as they were without the modification of the alignment parameters. This was reflected in the mean read count analysis shown above, where the protein coding read counts were effectively the same for the different alignment procedures, despite a noticeable increase in the unique mapping rate.






The individual annotation coverage graph shows similar coverage patterns to the samples when they were processed with the initial alignment parameters. There is not a significant or noticeable shift in the number of individual RNA species that can be captured when the alignment parameters are modified. This figure suggests that the modification of the alignment parameters produced an increased number of aligned reads to features that already had at least 1 annotation in the initial alignment step, and did not allow for previously unannotated features to be recovered during alignment. These findings suggest that contaminating rRNA reads were already detected at a low rate using the initial alignment parameters, and modification of these parameters increased the number of rRNA reads that mapped to these previously annotated features. This figure is again consistent with the hypothesis that the tissue samples used in this study had very specific transcriptome profiles that do not express all of the possible RNA transcripts that are found in the mouse genome.

Overall, the attempt to improve the alignment rates for the Kit A samples provided evidence for the altering of alignment parameters using STAR in order to improve the number of unambiguous alignments from RNA-seq libraries. However, an increase in uniquely mapping reads does not immediately result in an increase in the number of counts to RNA species of interest to a particular experiment; in reality, there is the possibility of increasing the alignment rate to ‘contaminating’ RNA species such as ribosomal transcripts, which offer little biological information to canonical transcriptomic studies. Furthermore, a slight decrease in the percentage of assigned features was observed using featureCounts, which suggests an increase in the alignment of reads to genomic regions that are not annotated for this project. As the number of protein coding read counts produced did not increase, and the overall number of ribosomal alignments increased, one can conclude that the modification of the alignment parameters was not beneficial for optimizing the RNA type coverage analysis to obtain more usable transcriptomic information from the same FASTQ files.