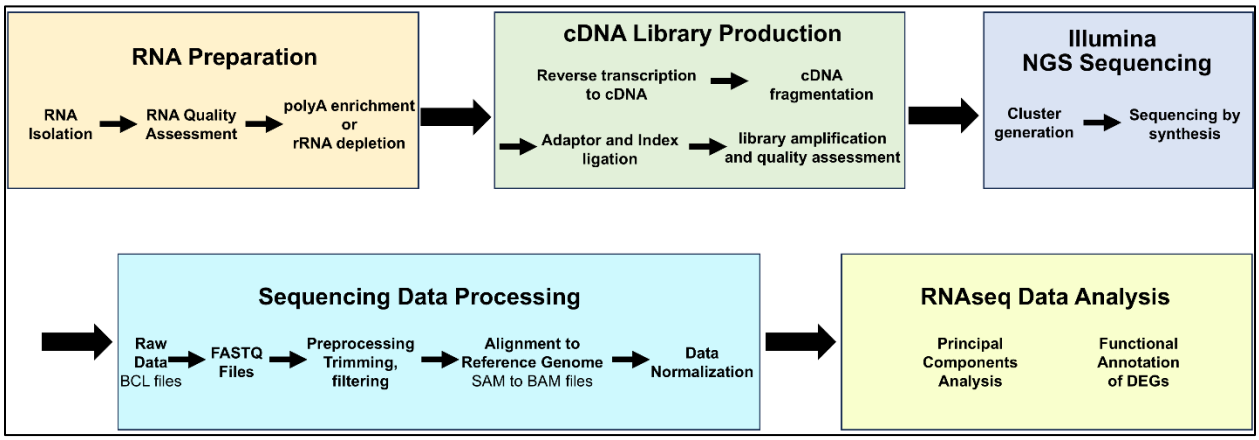
Bulk RNA-seq Instructional Manual

The purpose of this manual is to provide a practical guide for Montana State University researchers to perform bulk RNA-seq.

Author: Dana Rashid, Microbiology and Cell Biology Department, Montana State University

Funding for this manual was provided by NSF EPSCoR grant #2229308.

Fig. 1. Flow chart for bulk RNA-seq.



Introduction

What is bulk RNA-seq? Bulk RNA-seq is a method to analyze biological processes at the transcriptional level by reverse transcribing and sequencing the gamut of RNAs present in biological samples. For example, if a researcher wanted to analyze gene transcription in leukemia, they could collect blood samples from control and leukemia patients, extract RNA from those samples, generate cDNA libraries, sequence the libraries with next-generation sequencing, and then perform the bioinformatics analyses to discover the RNA expression levels in leukemia relative to non-leukemia controls (see Figure 1 for the steps involved in a bulk RNA-seq experiment). The individual data sets that are collected always compare two conditions, as in this example, between leukemia and non-leukemia. More expanded experiments can be performed, for two or more experimental conditions versus controls, but the generated data would still be comparisons between two conditions (e.g. experimental condition #1 vs control #1, and experimental condition #2 vs control #2). Bulk RNA-seq data provides a snapshot in time of differentially expressed transcripts; studying the progression of a biological process would require RNA-seq analyses of samples collected at multiple timepoints.

Bulk RNA-seq provides a list of the RNAs expressed in the samples and the relative differences in RNA levels between any two given conditions. It also provides the statistical significance of RNA expression differences between the two conditions. Bulk RNA-seq is therefore both qualitative and quantitative. Once the raw data set is generated and processed, that data can be

applied to various software programs that can, for example, group the data into known biological pathways and detail whether those pathways are up- or down-regulated in the studied process.

What can be studied with bulk RNA-seq? This method is usually used to evaluate gene expression. However, RNA-seq is not limited to coding RNAs, and is therefore not limited to gene expression analyses. Apart from mRNAs, bulk RNA-seq can also identify non-coding RNAs like miRNAs, small and long non-coding RNAs, and circRNAs. Identification of splice sites is also possible with this method. Bulk RNA-seq can be applied to whole tissues, cellular biological fluids, FACs-sorted cell populations, and even to material that has been formaldehyde-fixed. Typically, a reference genome/transcriptome is used to map the sequencing data, but bulk RNA-seq can also be used to assemble de novo transcriptomes. Bulk RNA-seq is an extraordinarily powerful tool for studying biological processes in various starting materials.

Limitations. Bulk RNA-seq is expensive, time-consuming, and laborious. It requires expertise at multiple levels, from sample processing to bioinformatics. Also, for MSU researchers, sequencing equipment for larger-scale bulk RNA-seq is not available on campus, and there is no dedicated bioinformaticist on staff (as of October 2024), requiring researchers to outsource to external sequencing facilities. Researchers can expect the processed datasets to contain hundreds to thousands of genes, which is both a challenge and a benefit of bulk RNA-seq.

Bulk RNA-seq versus PCR arrays. Bulk RNA-seq provides information on all extracted and reverse-transcribed RNAs present in a sample, whereas PCR arrays test a limited, defined group of coding or non-coding reverse-transcribed RNAs. PCR arrays are a convenient and relatively fast method to simultaneously test whether a defined group of genes or non-coding RNAs are involved in the studied process. Using the leukemia example above, RNAs extracted from control and leukemia patient blood samples would be reverse-transcribed to cDNA, PCR amplified and labeled, and the resulting amplified cDNAs would be applied to a chip in which anti-sense DNA fragments of relevant genes are arrayed in a grid. These grids or chips are commercially available representing various biological processes and species specificity, with either a pre-set or custom group of antisense DNA oligo probes. Hybridization of the labeled cDNAs to the chip probes results in a quantitative measurement (by fluorescence or chemiluminescence) of relative signal, and hence will reflect the relative abundance of the defined set of RNAs present in the sample. Since PCR arrays test for known transcripts, they are not used to discover novel genes involved in a process. Also, hybridization is not as reliable as next generation sequencing to provide relative RNA abundance; false positives and false negatives are more likely than with RNA-seq. As to cost, PCR arrays are not as expensive, but the cost differences are narrowing. Many researchers choose to perform both bulk RNA-seq and PCR arrays; the arrays can validate the RNA-seq results very well. Ideally, both methods should also be validated at the protein level, such as by Western blot, immunofluorescence staining or mass spectrometry.

Bulk RNA-seq vs Single-cell RNA-seq. Where bulk RNA-seq investigates the lump sum of transcripts in a sample, single-cell RNA-seq can differentiate the transcription profiles of individual cells and compare individual cell activities within a sample. Cells must therefore be isolated and sorted (a few hundred to tens of thousands of cells per sample), and are distinguished by barcoding. Barcodes, in the form of oligo adapters, are affixed to RNAs such that the RNAs from a given cell are labeled with the same barcode, and that barcode is unique to that cell. The major advantage of single-cell RNA-seq over bulk RNA-seq is therefore analysis of the contributions of individual cell transcriptional expression to a biological process, and the heterogeneity between cells. Cell differentiation and cell-cell communication can be

more specifically investigated. A disadvantage is that the lower capture efficiency and biases towards more abundant RNAs leads to more noise relative to signal. Also, the steps involved in removing cells from their in-situ environments, sorting, and barcoding can affect gene expression. Due to these extra steps, single-cell RNA-seq requires additional specialized equipment. Cell sorting can be accomplished by traditional methods, like FACS cell sorting, or by newer strategies such as microfluidics separation. In a related type of experiment, bulk RNA-seq can be used to investigate the transcriptome profiles of different cell types within a heterogeneous tissue if the cells are first sorted, as for single-cell RNA-seq, and then each sorted cell population sample is bulk sequenced. In this case the transcriptomes of each individual cell are not tested, but the transcriptomics is performed at the level of cell type. The choice of bulk versus single-cell RNA-seq methods depends on the biological questions being asked.

Transcriptome coverage with the different sequencing platforms. Next-generation sequencing (NGS) is usually performed using Illumina sequencing equipment, though other systems have been commercially developed. Illumina sequencers are available as different platforms, and each platform offers different capacities for the number of samples that can be processed at a time, and the number of reads that can be expected from a sequencing run. The number of reads equates to the number of sequenced cDNA fragments. The greater the number of reads, the greater the chance that the entire transcriptome is represented in the data. This parameter is referred to as 'depth'. The genome size of the organism being studied also comes into play. Bacteria, for example, have significantly smaller genomes than mammals, and therefore reasonably thorough bulk RNA-seq can be performed on a platform that generates fewer reads. To determine the most appropriate platform for a bulk RNA-seq experiment, the general rule is that bacteria require at least 8 million reads, mammals require at least 20 million reads, and plants require at least 40 million reads for gene expression comparisons. De novo transcriptome assembly, whole transcriptome analyses, and splice site analyses require more reads.

Illumina platforms range from iSeq to NextSeq sequencers; see https://www.illumina.com/systems/sequencing-platforms.html for more detailed information. Information from other platforms is available through their company websites. cDNA libraries are loaded onto flow cells specific to each platform. The flow cells must be purchased for each sequencing run; these are an expensive single purchase for RNA-seq. They also have a very limited shelf life (approximately six months); therefore, we recommend you buy the flow cells just prior to the sequencing run, when cDNA libraries that have passed quality control analyses are already in hand.

Sequencing can be performed on the DNA sense strand, called Single-End sequencing. Or, both the sense and anti-sense strands can be sequenced, referred to as Paired-End sequencing. Single-End sequencing is sufficient for evaluating gene expression changes, but Paired-End sequencing has greater depth and utility for whole transcriptome sequencing, splice site analyses, and de novo transcriptome studies.

The wide range of platforms, flow cells, and RNA-seq applications highlight the importance of working with an RNA-seq specialist at the chosen RNA-seq facility. A specialist will help guide an RNA-seq experiment from the planning to the execution stages, to maximize the chance of success. There are multiple points where an RNA-seq experiment can fail, before and during sequencing, so guidance by a specialist is strongly advised.

II. Sample preparation and cDNA library production

Number of replicates for each sample. The more replicates for each experimental condition, the greater the statistical significance of the data. Also, the lower the cut-off for differential expression, the greater the number of replicates that should be analyzed (1.5-fold up- or down-regulation of differentially expressed genes is the lowest fold cut-off to show significance between two conditions; a 2-fold minimum is preferable if possible). For a 2-fold cut-off, it is recommended that at the bare minimum, three biological replicate RNAs are analyzed for each sample type. Specifically, at least three different RNA preparations must be performed, from at least three separate starting material isolates for each sample condition. Dividing one RNA preparation to make more than one cDNA library (technical replicates) will not provide sufficient statistical significance. While three replicates is the bare minimum, preparing four to seven is strongly recommended. It is not uncommon for sequencing to reveal outlier libraries that need to be culled from the analyses (see Fig. 3).

Initial Sample Preparation. Even before RNA is isolated from a sample, the conditions of sample harvesting can make the difference between intact and degraded RNA. RNA is fragile and labile, and RNAses are everywhere. Bench surfaces, dissection tools, and pipettors should be thoroughly cleaned. We suggest avoiding RNase Away, a product touted to remove RNAses from surfaces. In practice, RNase Away can degrade RNA, and if not thoroughly removed from tools and pipettors, will degrade samples. Instead, autoclave tools and clean surfaces with 70% EtOH. If possible, set aside dedicated bench space for RNA work (a hood is not necessary).

The time between sample harvesting and RNA isolation should be minimized as much as possible. Fresh samples are better than frozen, for example, and samples snap frozen in liquid nitrogen, as small aliquots or tissue pieces are preferrable to large samples frozen at -80ºC (or worse -20ºC). The limits of sample harvesting time are specific to the sample. Colon cancer tissue samples, for example, should be processed for RNA isolation within ten minutes of collection [2](Yamagishi), whereas blood samples can be collected and stored on ice for several hours before RNA isolation and still provide good quality RNA [3].

Homogenization and/or cell lysis are critical for RNA isolation, especially for tissues. Tissues that are more easily homogenized typically are more successful than fibrous tissues for RNA quality and yield. Tissues like cartilage are among the most difficult, not only because of homogenization challenges but also because they are low in cell concentration and high in tough extracellular matrix. A common lysis solution is TRIzol (Life Technologies). For bulk RNA-seq, samples are added to TRIzol directly after harvesting and before homogenization and RNA isolation. Samples can be stored at -80ºC in TRIzol, with or without homogenization. Ideal sample types and harvesting conditions are often not possible, for example during field collections. In this case, there are room-temperature RNA preservative solutions such as RNAlater (Invitrogen) and DNA/RNA Shield (Zymo; likely the better alternative). These solutions quickly penetrate cells and tissues and can be good for sample storage, but can lower RNA yield and sometimes quality. They are not lysis solutions, and can hinder homogenization. Researchers should try several different sample preparation conditions to determine the most effective.

There are many homogenization methods (those discussed here are more relevant to bulk RNA-seq). For some samples, like blood or soft, loosely connected early embryonic tissues, TRIzol (or equivalent) alone or passing through small bore needles in TRIzol is sufficient. Other tissues require more intensive methods, such as bead beating or pulverization in liquid nitrogen. Bead beating is commonly employed and generally recommended, but requires specialized equipment. Samples in TRIzol or other solution are vigorously shaken with beater beads using

bead beater equipment, in tubes designed to withstand the mechanical stress. The beads themselves vary in size, surface texture, and composition. For RNA work, the metal or ceramic beads should be smooth; bead size and length of beating time are determined by the type of tissue. Bead-beating equipment will shake tubes at a range of intensity and time settings. Pulverization in liquid nitrogen can be performed manually or using specialized cryomill equipment. Samples are pulverized to a frozen powder and then added to a lysis or RNA-preserving solution. To aid homogenization, especially for hard-to-lyse samples, Proteinase K treatment can be performed in TRIzol, guanidine thiocyanate solutions, RNAlater, and DNA/RNA Shield. Also, Qiagen sells QiaShredder columns, which can prove useful for enhanced tissue homogenization. Researchers should search for homogenization protocols that most closely resemble their sample as a starting point for optimizing their method.

RNA isolation can be performed by alcohol precipitation, column binding and elution, or a combination of the two. Chloroform extraction commonly precedes alcohol precipitation in TRIzol-based methods. Column binding and elution, with or without alcohol precipitation, is recommended for clean, good quality RNA. RNA binding columns are typically purchased in RNA isolation kits, which are available for samples homogenized in different solutions, including TRIzol, guanidium thiocyanate, RNAlater, or DNA/RNA Shield. DNase I digestion is a common RNA isolation step. It can be performed in-solution or in-column, and can be critical for good quality RNA. It is important to remove any DNA contamination from purified RNA. Contaminating DNA can be differentially amplified during processing, and can skew the RNA-seq results. Our protocol for RNA isolation from cartilage includes harvesting in TRIzol, bead beating, chloroform extraction, isopropanol precipitation, DNase I digestion, and subsequent column clean-up and concentration with a Zymo kit (see RNA isolation from cartilage protocol, pg.X). Qiagen RNeasy kits are also a very popular choice.

It is important to empirically test the amount of starting sample material for RNA isolation. Approximately 50mg of tissue per 1 ml of TRIzol is a reasonable starting point. Too much sample in TRIzol can result in protein and/or carbohydrate contamination, rendering poor quality RNA. Too little starting material can prevent proper quality control assessments and provide insufficient RNA for cDNA library production.

Assessment of RNA yield and quality. To make good cDNA libraries for RNA-seq, it is imperative to start with good quality RNA of sufficient yield. This cannot be stressed enough: *badly degraded RNA leads to uninterpretable RNA-seq results*. It is not worth the time and expense to proceed with low quality RNA. In some cases, RNA can be somewhat repaired, and this may be the only option for certain starting samples like long-term storage formaldehyde-fixed sample, but in general, researchers should strive for high quality RNA from the outset.

RNA concentration can be determined by nanodrop analysis, but this method is not as accurate as Qubit analysis for cDNA library production. Qubit analysis entails mixing a small aliquot of the RNA sample with a proprietary solution that only binds RNA, and then reading the concentration using Qubit analyzing equipment. For our own protocol, we also run an aliquot of the RNA on an ethidium bromide-containing agarose gel. We only proceed with TapeStation RNA quality assessment if we observe clear 28s and 18s RNA bands of sufficient yield. For standard bulk RNA-seq, a minimum of approximately 100ng for each RNA sample is sufficient for quality control tests and cDNA library preparation, but at least 1µg is preferable.

RNA quality is usually assessed as an RNA Integrity Number (RIN), on a scale from 1 to 10. RINs are determined using either Bioanalyzer or TapeStation equipment. Bioanalyzer and TapeStation perform electrophoresis-based analyses, and provide reports of RNA quality and

yield (note that the yield is not as accurate as Qubit or nanodrop). An aliquot of RNA is mixed with a Bioanalyzer- or TapeStation-proprietary solution, and after a few simple subsequent steps is then loaded into the machine. The analyses are based primarily on the ratio of rRNA bands (28s and 18s in eukaryotes; 23s and 16s in prokaryotes) to each other and to the remaining RNA; rRNA should comprise approximately 80-90% of the total sample. To proceed with RNA-seq, RINs should be a minimum of 7. RNAs with values below 7 should either not be used, or, if proceeded with, will require additional processing and/or bioinformatics finagling. Truly, avoid this situation if possible. RNA quality can also be assessed by DV200 evaluation, which equates to the percentage of RNAs greater than 200 base pairs. DV200 analyses can be determined by TapeStation. From our own experience, we find that DV200 values of 75% or greater are indicators of good quality RNA. We assess RNA quality by both RIN and DV200, and do not proceed with cDNA library production unless the RNAs have RINs above 7 and DV200s above 75%. DV200 values may be a more accurate assessment of lower quality RNAs, such as those from formaldehyde-fixed material (referred as FFPE samples in the literature; formaldehyde-fixed paraffin-embedded) [4].

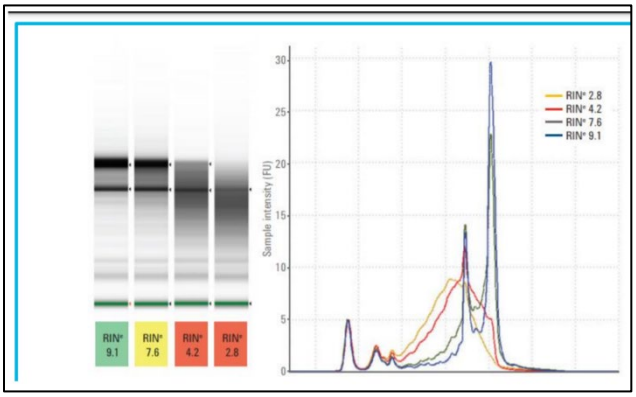


Fig. 2. TapeStation data showing intact and degraded RNA samples. Left panel, electrophoretic runs; the left 2 RNA samples are intact and acceptable for cDNA library preparation (RINs above 7); the right 2 are degraded, with RINS below 7. Right panel, a plot of the electrophoretic data. The two large peaks on the right are the 18s and 28s rRNA bands, respectively, evident in the two intact RNAs. From Agilent.com.

mRNA enrichment versus rRNA depletion. Because mRNA is only a small percent of total RNA (on the order of 1-5%), and non-coding RNAs are even less abundant, it is important to either enrich for mRNA by poly(A) tail selection, or to remove rRNA. mRNA enrichment is cleaner, but rRNA depletion generates more comprehensive RNA-seq data. For mRNA enrichment, poly dT magnetic beads bind to mRNAs, which are subsequently eluted. For rRNA depletion, the rRNAs are hybridized with DNA oligos corresponding to antisense rRNAs. RNase H is then employed to chew up the hybridized rRNAs, followed by DNase I to remove the DNA probes. For bulk RNA-seq, either mRNA enrichment or rRNA depletion can be performed, but rRNA depletion is becoming the preferred method. Prokaryote mRNAs do not have poly(A) tails, so in this case rRNA depletion is the only option. For eukaryotes, not all mRNAs and non-coding RNAs have poly(A) tails, so these would be excluded with mRNA enrichment. For single-cell RNA-seq, mRNA enrichment is still more common, but rRNA depletion is possible [5].

Ribosomal RNAs are highly conserved in eukaryotes, so using an rRNA depletion kit for human/mouse/rat should theoretically work for other species. That being said, it is possible that small variations in rRNA sequence in non-mammalian species compared to mammalian may impact the specificity of the depletion. Custom rRNA depletion kits are available and should be considered in this case. We suspect that some of our issues with cDNA library yield may be the consequence of depleting chicken rRNAs using a mammalian rRNA depletion kit.

Since this is the first step towards cDNA library generation after the RNA has been isolated, it is important to consider the total amount of RNA that will be needed from this point. Depending on the cDNA library prep kit, 25ng to 1µg is used as the RNA starting input for rRNA depletion or mRNA enrichment. In general, the lower the quality of the RNA, the higher the RNA input. For example, if the kit suggests using a range of 25ng to 250ng, and your RINs are 9-10, consider using maybe 50 to 100ng starting RNA. If your RINs are hovering around 7, then consider using maybe 200ng. There is no absolute rule here, only an educated guess. You will likely need to substantiate your RNA-seq data with RT-PCR, so you may want to save enough RNA at -80C for these analyses.

Production of cDNA libraries. NGS sequencers require that cDNA libraries are prepared in a very specific way. For this reason, it is necessary to generate the libraries using commercial kits that are compatible with the specific platforms. We are familiar with the NEBNext Ultra II DNA Library Prep Kit for Illumina, though numerous other kits are commercially available. Magnetic bead separation steps are typical in the kit protocols. The RNA-binding magnetic beads can be purchased with the kits, or separately. The protocols require temperature programs using a PCR machine.

cDNA library preparations are labor intensive and are generally best performed over two days. You should carefully follow the protocol that comes with your library prep kit. The NEBNext Ultra II DNA library protocol starts with mRNA enrichment or rRNA depletion from purified RNA (the RNA has been previously analyzed for quality and concentration and recently frozen at -80ºC). The RNAs are then fragmented, and subsequently reverse-transcribed to cDNAs. Adaptors are then ligated to the fragment ends to allow subsequent index primer ligation. From a kit containing multiple different index primers (multiplexing for Pair-End sequencing), unique 5' and 3' indexes are chosen for each sample and ligated, such that each sample has its own unique combination of 5' and 3' indexes that are recognized by the sequencer and distinguished in downstream data analysis.

cDNA library amplification is an important step that follows index ligation. *A critical lesson we have learned is that cDNA libraries sequenced in the same run and part of the same bulk RNA-seq experiment should be amplified with the same number of PCR cycles.* If different numbers of PCR cycles are used for different samples, they will not be directly comparable. The number of library amplification cycles is loosely determined by the amount of starting RNA and the RNA RIN numbers. More starting RNA and higher RIN numbers equates to fewer numbers of cycles. You should consult with your RNA-seq specialist to determine the number of cycles for your samples. Over-amplification of libraries is usually a bigger problem than under-amplification. Consistency is important: for optimal cDNA library generation, make all cDNA libraries for one experiment using the same amount of starting RNA, with similar RINs and the same number of PCR amplification cycles.

Once the cDNA libraries are generated, they must be checked for quality and concentration. For quality, the libraries should predominantly contain fragments near 300bp. The fragment size distribution can be determined using either a Bioanalyzer or TapeStation. A common contaminant in cDNA libraries is adapter dimers, observed as a peak of approximately 150bp. This can occur when there is an overabundance of adapter primers to cDNAs during library preparation, or poor primer ligation. As little as 5% adapter dimer contamination relative to total library can result in as much as 50% of the sequenced reads correlating to the adapter dimers. If adapter dimer contamination is evident, all is not lost; adapter dimers can often be removed by an additional magnetic bead purification of the cDNA libraries. Some of the library yield is lost during this step, but this loss is far preferable to dealing with adapter dimer sequences later.

For cDNA library concentration, a range of 8 ng/ul to 50 ng/ul (in a total of 20ul) is reasonable for Illumina sequencing. Qubit is the best system to assess concentration initially, followed by KAPA analysis. KAPA is a PCR-based analysis for library quantitation; kits for this analysis are available from different companies. We have opted to have KAPA analyses performed by the sequencing facility, because the kits are often more expensive than what the facility charges for analyzing a small number of libraries. Once the cDNA libraries pass the quality control analyses, they can be sequenced.

cDNA library sequencing. Our experience is with the Illumina NextSeq 2000 sequencing equipment at Clemson University; this high performance machine can generate over one billion reads per run, and can sequence up to 16 libraries. Three different flow cells are available, with varying capacities for the number of reads. Our last run, for example, required the P2 flow cell, with a capacity for up to 400 million single-/800 million paired-end reads; we had 12 cDNA libraries (three conditions, four biological replicates per condition), and averaged about 70 million reads per sample. Different sequencers and different platforms have their own requirements/limitations for the cDNA libraries they can sequence.

With the Illumina systems, all libraries are pooled together and loaded into the flow cell. Pooling the samples is possible because of the unique index primers for each library. Only a very small concentration of pooled libraries is loaded onto the flow cell; in our case, 650pM total, in 20 microliters. The library concentrations are best determined by KAPA analysis, and each library *must have the same ultimate concentration in the final pooled volume*. It is ideal if the individual concentrations of libraries before diluting and pooling are at least 1nM (libraries before pooling that are under 650pM can also be sequenced, as long as the remaining libraries have high enough concentrations to accommodate, and a higher volume of the lower concentration library is used relative to the other libraries). On the Illumina website, a calculator for library dilution can simplify the dilution determinations (https://support.illumina.com/help/pooling-calculator/pooling-calculator.htm). It is critical not to over- or under-load the flow cell, which can result in failure of the sequencing run or uninterpretable results.

Once the flow cell has been loaded into the Illumina machine, a program of multiple steps is initiated. The individual cDNAs bind to the internal glass surface of the flow cell by hybridization of complementary oligo-adapter sequences. Clustering of the cDNAs by adaptor sequences and amplification of the cDNAs occurs before sequencing. The actual sequencing is achieved by ‘sequencing by synthesis’ or SBS, where each added fluorescently tagged nucleotide is detected as polymerases generate new DNA strands from the forward strands of amplified cDNA library templates. Hundreds of millions of added nucleotides are read in parallel. Paired-end sequencing then commences to sequence the reverse strands. Ultimately, millions of reads are generated, so that all original cDNA library fragments are well-represented. The libraries were generated with unique indexes to distinguish each library; the sequenced fragments are therefore separated and grouped according to those indexes.

If becoming a bioinformaticist is not high on your agenda, especially if RNA-seq is more a tool for your studies rather than a major career pursuit, we recommend that you pay the sequencing facility to perform the data processing and analysis for you. At the Clemson University Genomics and Bioinformatics Facility (CUGBF), where we send our cDNA libraries, it is more expensive to receive the training than to have the analysis performed by them for the first couple RNA-seq experiments. Trained bioinformaticists, like those at the CUGBF facility, can generate the profiling data relatively quickly, on the order of weeks. For a first-time researcher tackling RNA-seq bioinformatics, a timeframe of six to eight months is not unreasonable. Also,

you can invest in training a student or post-doc who will ultimately leave your lab, leaving you in the situation you started with: needing RNA-seq data analysis with no one in your lab who can perform that analysis. We know this from personal experience. In addition, RNA-seq bioinformatics is constantly changing, and becoming more and more user-friendly. From our perspective, it has not been worth agonizing over Linux command line prompts for a system that will eventually become far more simplified.

Of course, if RNA-seq is going to be a major focus of your work, you should consider investing in the training; it could ultimately save money and support the career goals of students and postdocs. The training will come with a greater understanding and appreciation of the entire process. Course recommendations for bioinformatics training at MSU are MB 544 Advanced Bioinformatics, and MB 535 Genomic Analysis Lab. Training is also available through the CUGBF facility at Clemson (~30 hours, currently $100 per hour) and through the UMGC facility in Missoula. Understanding Linux programming, especially in R, is hugely helpful prior to any training. MSU has the super-computing capacity to analyze RNA-seq data with our Nex-Generation Supercomputer, Tempest. Many software programs for processing and analyzing RNA-seq data are on Tempest, and the MSU Tempest personnel are very helpful. The Tempest facility runs its own courses, and the computer processing usages required for RNA-seq are within the free access limits for MSU researchers.

III. Bioinformatics

Raw sequencing data processing. The raw sequencing data, in the form of BCL files, is first converted to FastQ files. The FastQ file data is quality-assessed and processed before the data can be analyzed (we use the FastQC program). For quality assessment, parameters such as read length, GC content, base quality, and sequencing adapter contamination can be examined (usually with a FastQ readout report). The data should pass this quality test before proceeding with further processing. Next, before aligning with the reference library, the adaptor and index sequences must be removed to prevent misalignments, in a process called ‘trimming’. Apart from the adaptor and index sequences, low quality reads and reads that are too short must also be removed. Additionally, contaminating rRNA sequences should be filtered out, especially if rRNA depletion was performed. Alternatively, rRNA sequences can be separately annotated if built into the reference library.

Reference library alignment. Once quality-controlled FastQ files are generated, the RNA-seq data can be aligned to a reference library. The reference library is the sequenced genome or transcriptome of the species you work with (mouse, human, chicken, etc.). In the event a reference library is not available, de novo transcriptomes can be assembled using closely-related species libraries. There are multiple sources of reference libraries, including Gene Expression Omnibus database (GEO; https://www.ncbi.nlm.nih.gov/geo/), European Molecular Biology Laboratory’s European Bioinformatics Institute (EMBL-EBL; https://www.ebi.ac.uk/), DNA Data Bank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/), and National Genomics Data Center (NGDC; China; https://ngdc.cncb.ac.cn/) [6]. The appropriate reference library, in the appropriate format, is uploaded into whatever software program you are using for alignment (we use STAR for mapping to a genome; SALMON is a pseudoalignment tool available on Tempest that performs a similar function). At this point, the data is in the form of SAM files (Sequence Alignment Map), which are compressed to BAM files (Binary Alignment Map). Once the transcript reads are aligned to the library, the quantitation of mapped reads (number of transcripts mapped to individual genome/transcriptome sites) is performed using another program (we use featureCounts). The mapped transcript levels must then be normalized; normalization methods vary and include counts per million (CPM), transcripts per

million (TPM), reads per kilobase of transcript per million reads mapped (RPKM), or fragments per kilobase of transcript per million reads mapped (FPKM) [6] (we use DESeq). Clearly, there are numerous programs available to perform the different processing steps. Working with a bioinformaticist is essential to define the pipeline that is best tailored to your experiment.

Bulk RNA-seq Data Analysis

Principal component analysis. The sequenced, aligned, quantitated, and normalized data is now ready to be analyzed. The first order of business is to determine how similar each replicate is to the other replicates for each experimental condition. This can be achieved by principal component analysis, which reduces the large RNA-seq dataset into a graph showing the variance between libraries (Fig. 3). In an ideal experiment, all libraries from each experimental condition will cluster together. Since experiments are seldom ideal, this analysis is especially useful to weed out outlier libraries, as indicated by the arrows in Fig. 3.

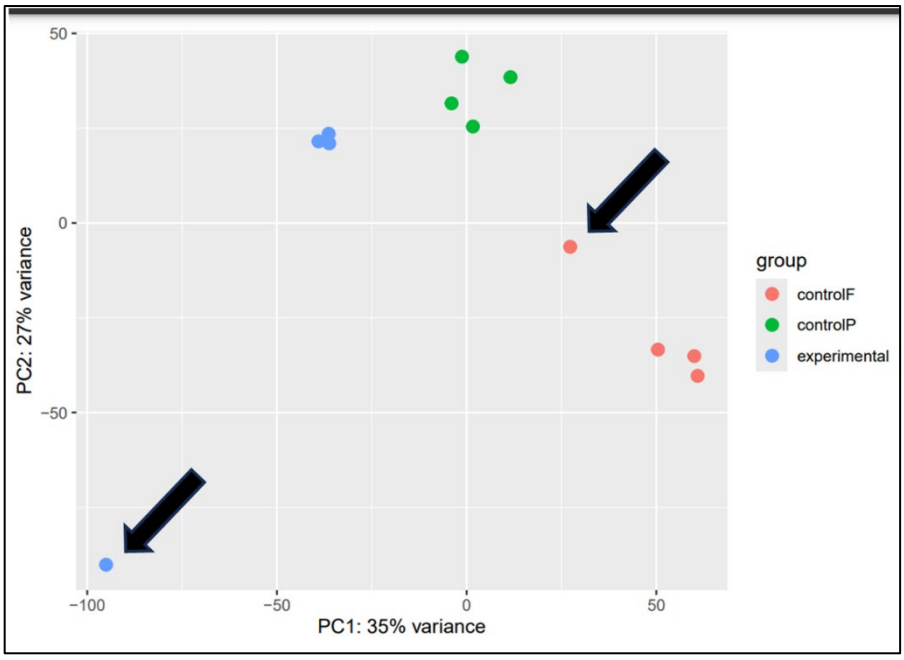
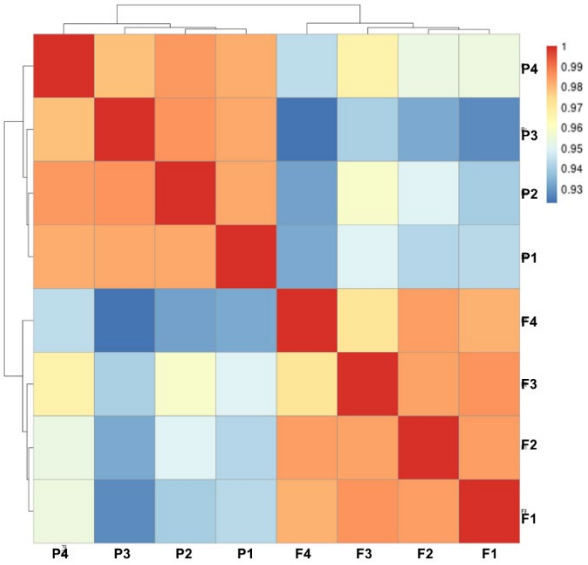


Fig. 3. Principal component analysis of a bulk RNA-seq experiment. Clusters of the experimental conditions are evident. Outlier libraries, indicated by arrows, show substantial variance from the other libraries of their condition, and would negatively impact data analysis if included.

Variance/similarity between libraries can also be reflected by heatmaps, generated in R, as shown in Fig. 4.



Differential gene expression analysis. For bulk RNA-seq, at this point, you will likely have generated an Excel sheet that lists all of the detected genes and (possibly) non-coding RNAs that mapped to your transcripts, the levels of differential expression of those genes/non-coding RNAs for each 2-part comparison (e.g. experimental #1 vs control #1) in logarithmic scale, and the statistical significance, in p-values, for the differential expression of each gene/non-coding RNA. The data will include both up- and down-regulated expression information; for example, gene X is upregulated $\log_2 1.0$ relative to the same gene X in the control, with a p-value of 0.002 (in this example, gene X is expressed two-fold higher in the experiment compared to in the control). The data can be ordered by degree of differential expression, or expression adjusted by p-value significance. You will need to set a cut-off for differential expression level and p-value, because while a great many genes/non-coding RNAs will be detected, only a subset of those will be statistically differentially expressed. The lowest cut-off for differential expression is 1.5-fold differential expression, or $\log_2$ (0.68 or -0.68); the most accepted p-value for statistical significance is less than or equal to 0.05. If possible, a more stringent differential expression cut-off of 2-fold, $\log_2 1.0$, is preferable. The lower the cut-off, the more genes/non-coding RNAs can be analyzed; experiments that yield fewer statistically significant differentially expressed genes often use the lower cut-off. For the number of differentially expressed genes you can expect, our own data has yielded over four thousand differentially expressed genes (including up-and down-regulated, 2-fold cut-off), but every experiment is unique. Many RNA-seq experiments target a relatively small window of genes/non-coding RNAs in a biological process, and the number of statistically differentially expressed transcripts can be substantially lower, in the hundreds.

Once the differential expression and p-value limits have been set, the data can be applied to various software programs that group the genes/non-coding RNAs into known pathways, group by transcription factor motifs or microRNAs, identify splice variants, and more. What we found to be extremely elucidating is to first manually investigate gene functions of the top 500 up-regulated differentially expressed genes using genecards.com. Genecards provides all known gene aliases and a concise description of the known functions for each gene. We write short descriptions of the genes' functions in the Excel file; this greatly helps us search the file for genes with related functions. We do this before we apply our data to the enrichment analysis software programs because it gives us a valuable preview of the data. Also, the software

programs don’t always indicate whether a particular gene promotes or inhibits a particular biological process or is up- or down-regulated in that process—many only note gene involvement. That being said, gene enrichment analysis software can be the most informative bioinformatics tool.

There are multiple software programs that relate differentially expressed genes to known biological processes. We are familiar with GO (Gene Ontology) [7] and KEGG (Kyoto Encyclopedia of Genes and Genomes [8]), which are fairly standard, but many others are available [6]. The data can be presented in various formats, including heatmaps and other plots (Fig.5) and tables. We have also found Pathview to be very useful [9]. Pathview provides a visual representation of gene expression within cellular compartments for specific pathways, and shows up- or down-regulation of those differentially expressed genes detected in the RNA-seq data that are involved in the pathway (Fig.6).

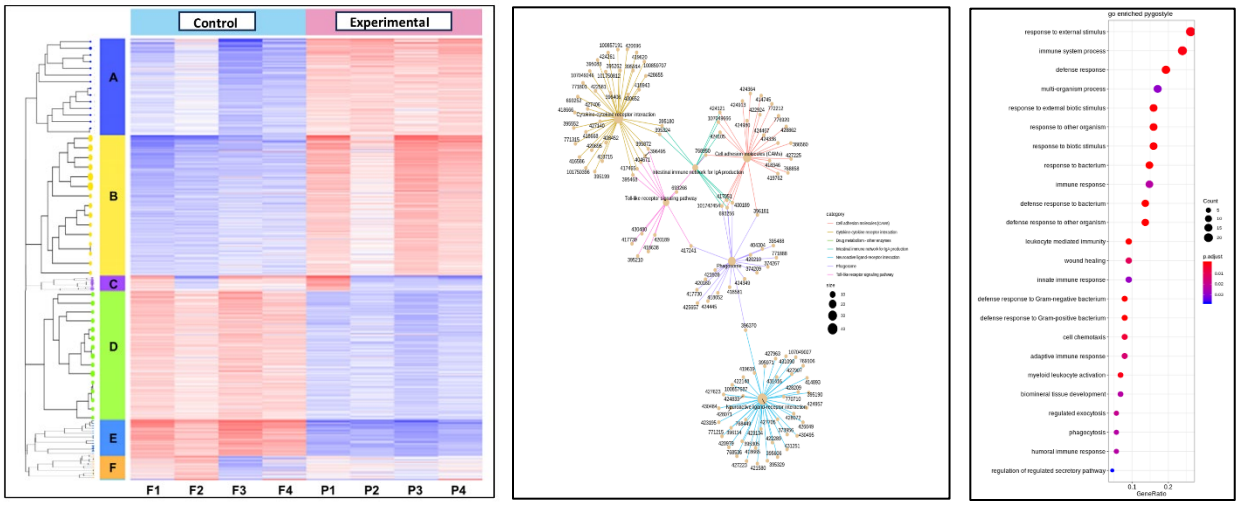


Fig.5. Sample visualizations of GO and KEGG data. Left panel, heatmap of differential expression analysis with GO annotation; the letters denote related clustering processes. **Middle panel**, KEGG data showing the relationships of genes involved in noted biological processes. **Right panel**, GO analysis showing the degree of gene representation in noted biological processes. These plots are from our own published data [1].

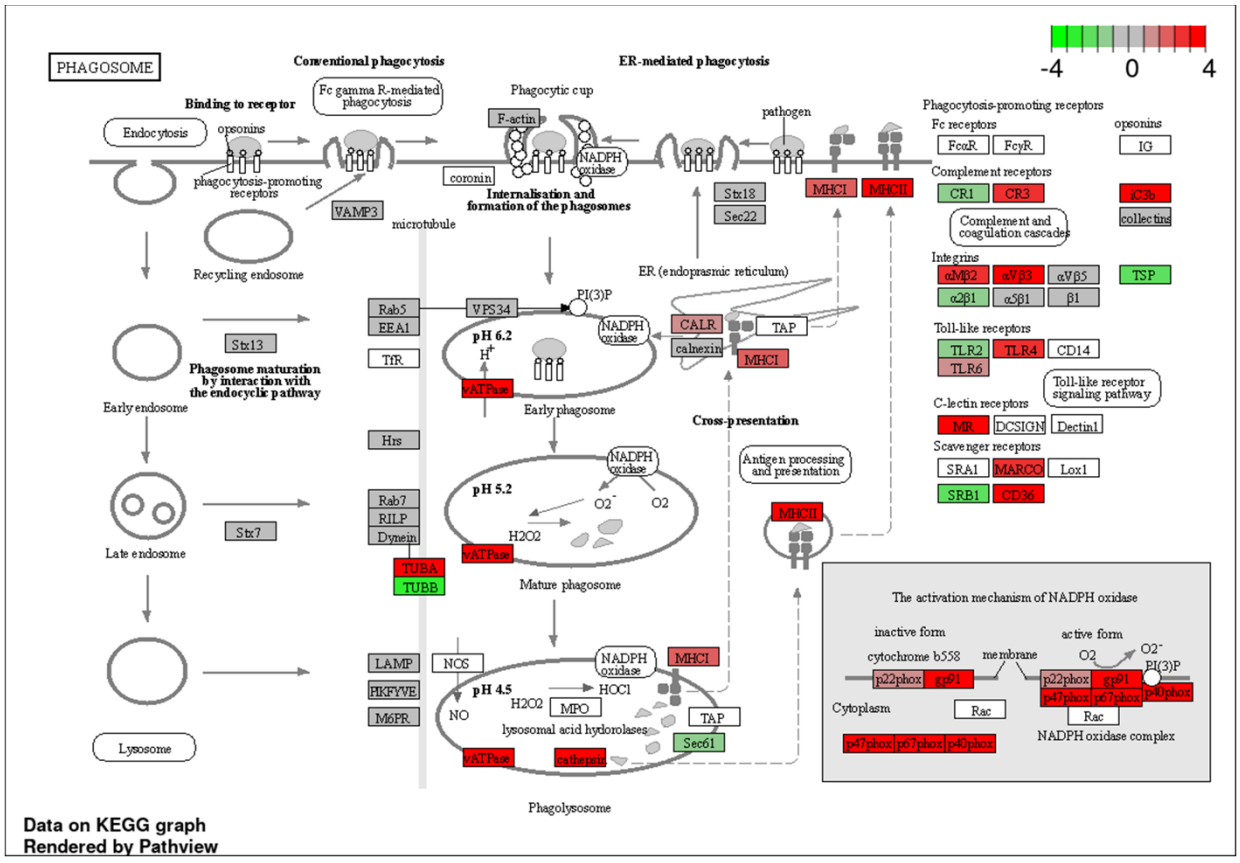


Fig.6. Pathview representation of KEGG data applied to the phagosome pathway (data generated from [1].

MSU in-house procedures and interfacing with external facilities. Because to date MSU does not have the equipment for larger-scale bulk RNA-seq NGS sequencing, MSU researchers outsource at least the sequencing of cDNA libraries. Many companies and academic RNA-seq facilities will perform all steps if necessary, including RNA isolation and bioinformatics, once provided with the biological samples. However, if problems arise before or during cDNA library preparation, the cost could rise exponentially and/or the project could ultimately fail. We have experienced this problem, primarily due to the difficulty of extracting good quality RNA from cartilage. Optimization of RNA isolation protocols is not part of an external facility's wheelhouse. RNA can be isolated in-house and sent to an external facility, though you risk RNA degradation in transit. cDNA libraries are far more stable, and have a better chance of arriving at the external facility in good shape. Also, an external facility's charges for cDNA production can be exorbitant. MSU does have the equipment necessary to generate cDNA libraries in-house, and several laboratories perform RNA-seq.

While MSU does not operate its own bulk RNA-seq sequencer, it does have the instrumentation for single-cell RNA-seq. For cell sorting and analysis, the facility has a Bio-Rad Z5 Cell Analyzer, a Stratedigm SE520EON machine, two BD FACS Caliburs, a BD FACS Aria II cell sorter within a biosafety cabinet, an ImageStream imaging cytometer, and a 10X Genomics single-cell RNA sequencer. Questions about performing single cell RNA-seq should be directed to the Cellular Analysis Core, to Diane Bimczok or Andy Sebrell (see contact information below).

Our lab has had great success working with the Clemson University Genomics Facility for bulk RNA-seq (see contact information). Our collaboration with a Clemson researcher has allowed us to perform all the procedures at Clemson ourselves, including RNA isolation (though RNA or cDNA libraries could be sent to the facility). The facility operates an Illumina NextSeq 2000, and has all the other required equipment for cDNA library preparation and quality control analyses. The people at the Clemson facility are extremely helpful and a highly dedicated group. Our resulting datasets were of very high quality and reproducibility, and we have been very happy with how the facility has helped us with the bioinformatics. The bioinformatics rates are reasonable; our current project, requiring raw data preprocessing, alignment, and differential expression analysis with GO and KEGG is projected to cost approximately $1,000 (10 hours, $100/hour; rates subject to change). For inquiries, contact Rooksie Noorai (contact information below).

The University of Montana in Missoula has established its own genomics facility (UMGC), equipped with an Illumina MiSeq sequencer. The folks in Missoula are well-equipped to make cDNA libraries. They also offer library prep training, and have a bioinformaticist on staff, Schuyler Liphardt. If a bulk RNA-seq experiment requires sequencing beyond the capacity of the MiSeq equipment, the facility will outsource the sequencing to Novogene. All of your bulk RNA-seq needs can be met through this facility. In addition to bulk RNA-seq, UMGC has 10X single cell and spatial transcriptome instrumentation, from Chromium and CytAssist. Single-cell RNA-seq and its bioinformatics analysis is therefore another RNA-seq capacity at the University of Montana.

IV. Equipment and Supplies

Required equipment

For cell sorting

An array of equipment for cell sorting and cell analysis is available in the Cellular Analysis core facility (see contact information below for more details).

For RNA extraction and cDNA library generation

A bead beater and TapeStation are available in the Animal Biosciences Building. I have not located a Qubit machine, but nanodrop machines are fairly common across campus, as are PCR machines. Well-calibrated and cleaned pipettors are a must, especially a pipettor for small volumes (less than 2 microliters). You may need to purchase your own magnetic tube holder. We have used an inexpensive knock-off from Temu that was satisfactory.

For sequencing

MSU does not have the equipment for large-scale bulk RNA-seq sequencing, but does have equipment for single cell RNA-seq in the Cellular Analysis core facility. The U of M Genomics Lab in Missoula currently operates an Illumina MiSeq sequencer; larger-scale bulk experiments will need to be sequenced outside of Montana. The most commonly used sequencing platforms are from Illumina, but several other sequencing platforms have been developed. From Illumina, larger-scale bulk RNA-seq is run on the NextSeq or NovaSeq platforms.

Computer requirements

Bioinformatics is best performed using a super computer. MSU's Tempest super computer can easily handle RNA-seq data analysis, and several relevant programs have been uploaded into the system, including Trimmomatic, Salmon, FastQC, featureCounts, DESeq, and others. Any additional needed programs can be added to the system. Outside facilities that perform bioinformatics will have access to their own super computer.

Required Supplies

For RNA isolation

You will likely need dry ice and/or liquid nitrogen. For RNA extraction we recommend Trizol, but other solutions like guanidinium-based or RNA-preservative type solutions like DNA/RNA shield (Zymo) can be used. For bead beating homogenization, metal or ceramic beads of various sizes and surface textures can be purchased either by themselves or already loaded into RNAse-free bead beater tubes. It is advised to use tubes that are designed to withstand the significant mechanical stress of bead beating; we have had success with Next Advance RINO screw capped tubes. We have not had much success with column kits that directly purify RNA from Trizol, so instead use the Zymo RNA Clean and Concentrator kit once we have isopropanol-precipitated RNA from Trizol and chloroform (see protocol X). QIAShredder may be useful (Qiagen). DNase I and/or Proteinase K may be required; it is important to purchase these enzymes as RNase-free if they do not come with an RNA purification kit. RNAeasy kits (Qiagen) are a popular column purification choice.

For cDNA library production

We use either the NEBNext® rRNA Depletion Kit v2 (Human/Mouse/Rat) or the Magnetic mRNA Isolation Kit, both from New England Biolabs. Magnetic beads can be purchased separately or with the kits; SPRI beads (NEB) or AMPure XP beads (Beckman; must be warmed to room temperature before use) are acceptable. Also from NEB, we use the NEBNext Ultra II DNA Library Prep Kit for Illumina. If you are using a different commercial sequencing platform, you should order kits specific for that platform. The NEB library prep kit requires an additional kit for multiplex oligo adapters and primers; we use the NEBNext® Multiplex Oligos for Illumina, available as either a pre-loaded plate or as individual tubes.

We recommend that you use barrier pipet tips for both RNA isolation and cDNA library preparation, and you will need *a lot* of them.

V. Contact Information

Dana Rashid, Microbiology & Cell Biology Department, MSU
I would be happy to discuss your RNA-seq experiment, and help optimize your RNA isolation protocol.
dana.rashid@montana.edu or danarashid5@gmail.com

MSU Cellular Analysis core facility
https://wetlands.msuextension.org/mbi/facilities/cellularanalysiscore.htm
Facility Director: Diane Bimczok, diane.bimczok@montana.edu
Facility Manager: Andy Sebrell, andysebrell@montana.edu

MSU Tempest Research Cluster
https://www.montana.edu/uit/rci/tempest/
Senior Manager: Coltran Hophan-Nichols, coltran@montana.edu
System Administrator: Kenny Hanson, kenny.hanson@montana.edu
System Administrator: Alexander Salois, alexander.salois@montana.edu

Clemson University Genomics and Bioinformatics Facility (CUGBF)
https://scienceweb.clemson.edu/cugbf/
Facility Director: Christopher Parkinson, viper@clemson.edu
Assistant Bioinformatics Director: Rooksana (Rooksie) Noorai, rooksan@clemson.edu
Sequencing specialist: Maslyn Greene, maslyng@g.clemson.edu

U of Montana Genomics Core (UMGC):
https://www.umt.edu/genomics-lab/default.php
Facility Director: Jeffrey Good, Jeffrey.Good@mso.umt.edu
Core Facility Manager: David Xing, David.Xing@mso.umt.edu
Bioinformatistician: Schuyler Liphart, Schuyler.Liphardt@mso.umt.edu

VI. Protocols

Protocol #1. RNA isolation from chicken intervertebral disc cartilage

Dana Rashid, MSU; Susan Chapman, Clemson U.

Note: animals must be harvested one at a time to minimize the time required for tissue extraction. Intervertebral discs are dissected, then immediately placed in 1ml Trizol in 2ml tubes, then frozen on dry ice before transferring to -80ºC. The samples can remain at -80ºC for a few months before RNA isolation.

Steps:

1. Thaw tissue/Trizol on ice. Remove tissue from Trizol to a weigh boat and quickly mince with a fresh scalpel. Transfer the minced tissue and Trizol into a new 1.5ml RINO navy screw-tapped tube pre-loaded with metal beads, then immediately bead beat: maximum setting, 120-150 seconds, 4ºC. Place the tube in dry ice immediately after bead beating. Do this for each tube individually until all tubes have been re-frozen in dry ice.
   Notes: this second freeze-thaw on dry ice is necessary for good quality RNA. Also, not all bead beaters are the same—other bead beaters may require more or less time; if more time is required, consider two bead beatings with cooling the tubes on ice between beatings. For the RINO tubes, before adding any sample, remove all the smallest stainless steel beads but keep the two larger sizes—the smaller beads have a rough texture that can break up RNA.

2. Thaw tubes on ice. Transfer homogenates to new 2ml screw capped tubes. Spin 12kxg for 15 minutes at 4ºC. While spinning, set up new 2ml screw capped tubes with 350µl chloroform each.

3. Transfer homogenate supernatants into the chloroform-containing tubes. Mix the tubes by hand by inversion until pink and frothy (don't shake).

4. Spin 12kxg for 15 minutes at 4ºC. While spinning, set up new 2ml tubes with 500-600µl isopropanol.

5. After spin, carefully transfer each homogenate aqueous phase supernatant to an isopropanol-containing tube. You should see a hazy interphase between the layers; this is the RNA (with some DNA contamination).

6. Add 1.5µl of 20mg/ml glycogen to each tube. Mix gently by inversion at room temperature for 10 min. You should see small white RNA/DNA particulates form in the tubes, which can be checked under a dissecting scope.

7. Spin 12kxg for 15 minutes at 4ºC.

8. Carefully remove the isopropanol, leaving the pellet. Add 300µl 72% EtOH (made up with RNAse-free water).

9. Spin 12kxg for 3 minutes at 4ºC.

10. Remove EtOH. The 72% EtOH step can be repeated, or the tubes are then tap spun and the remainder of EtOH is removed. Air dry the tubes for 5 minutes at room temp.

11. Resuspend the pellets in 44µl RNAse-free water (it is best to use the water from the RNA column purification kit).

12. Heat the tubes for 30 seconds at 55ºC, then put tubes immediately on ice.

Column Purification (Zymo RNA Clean and Concentrator Kit)

13. Perform DNase I in-tube treatment:
Add 5µl DNA Digestion Buffer + 1µl reconstituted DNAse (both from the kit). Let sit at room temp for 15 minutes.

14. Add 100µl RNA Binding Buffer (kit) + 150µl 100% EtOH to each tube. Pipet each mixture into a column; let sit 1 minute at room temp.

15. Spin 30 seconds 12kxg, discard flow-through.

16. Add 400µl RNA Prep Solution (kit). Spin 30 seconds, then discard flow-through.

17. Add 700µl RNA Wash Solution (kit). Spin 30 seconds, then discard flow-through.

18. Add 400µl RNA Wash Solution (kit). Spin 1 minute, then discard flow-through.

19. Elution:
Add 15-23 µl RNAse-free water to each column. Let sit 2 min at room temp. Transfer columns to new final collection tubes. Spin 1 minute to obtain the final clean RNA.
The RNAs should then be assessed for quantity and quality, and can be subsequently stored at -80ºC.

Protocol #2. RNA isolation from zebrafish embryos

Ryan North, Christa Merzdorf Laboratory

Equipment and Reagents Required

- TRIsure (BIO-38032)
- Chloroform
- Ice Cold Isopropyl Alcohol
- 75% ethanol (diluted with RNase Free H2O)
- Nuclease-free H2O (Invitrogen 10977-015)
- DNase I, RNase-free (Thermo EN0521)
    - Includes 10x reaction buffer, DNase I, 50mM EDTA
- Monarch RNA Cleanup Kit (T2030L)
    - Add 4 volumes of ethanol to Wash Buffer with new bottle
    -

TRIsure Protocol- source

1. Turn the centrifuge on so that it can reach 4C while you do the next steps
2. Place 50 embryos into a 1.5mL microcentrifuge tube, remove as much water as possible
3. Add 250uL of TRIsure reagent to the tube immediately (in fume hood)
4. Homogenize embryos with a mini pestle (or a 1000p pipette tip) until the tissue is sufficiently disrupted (30 strokes or so)
    a. Be very careful when using the pestle, avoid the TRIsure overflowing
    b. The embryos tend to get caught in the bottom of the tube, vortex and try to homogenize on the wall of the tube
5. Add 250uL TRIsure reagent to make the total TRIsure volume 0.5mL
6. Incubate samples for 5 minutes at room temp.
7. Add 125uL of chloroform per 0.5 mL of TRIsure used
8. Cap tubes, and shake vigorously for 15 seconds
9. Incubate samples for 2 minutes at room temp.
10. Centrifuge samples at 12,000xg for 15 minutes at 4C (This is about 10,000rpm on our centrifuge)
    a. You should see a pale green organic phase, an interphase, and the colorless aqueous phase
11. Transfer the aqueous phase very carefully, do not disturb the interphase, into another tube
    a. Leaving some aqueous phase in the tube is preferable, and interphase will introduce contaminants
12. Add 1mL cold isopropanol, mix and allow to sit for 10 min at room temp to precipitate RNA
13. Spin for 10min at 12,000xg at 4C
14. Remove supernantant
15. Add 0.5mL ethanol
16. Spin
17. Remove supernatant

18. Repeat steps 15- 17 4x
19. Once you remove the final supernatant, resuspend in 50uL H2O

DNase Treatment- source

1. Add the following reagents per RNase-free microcentrifuge tube:

Total RNA	1ug
10x Reaction buffer with MgCl2	1uL
DNase I, RNase free	1uL
DEPC-treated H2O	Volume to equal 10uL

2. Incubate at 37C for 30 minutes
3. Add 1uL 50 mM EDTA and incubate at 65C for 10 minutes

Monarch RNA cleanup- source

1. Add 1 volume ethanol to the RNA mixture following DNase treatment, mix well by pipetting up and down gently
2. Inset an RNA cleanup column into the collection tube
3. Transfer the RNA mixture to the column and close the cap
4. Centrifuge for 1 minute at 16,000xg
5. Discard the flow through
6. Transfer the column to a new collection tube
7. Add 500uL RNA Cleanup Wash Buffer
8. Centrifuge for 1 minute at 16,000xg
9. Discard the flow through
10. Repeat wash steps 7-9
11. Transfer the column to a new collection tube
    a. Use care to ensure that the tip of the column does not come into contact with the flow through, if in doubt re-spin for 1 minute to ensure traces of salt and ethanol are not carried over to the next step
12. Elute in nuclease-free H2O by adding 6uL H2O and spinning for 1 minute
13. Add another 6 uL to the column and spin again for 1 minute
14. Store at -80C
    a. Note: it takes about 3 freeze thaw cycles to begin to noticeable degrade the RIN of RNA, but it degrades quickly so thaw deliberately

VII. References

1. Rashid, D.J., Sheheen, J.R., Huey, T., Surya, K., Sanders, J.B., Horner, J.R., Voyich, J., and Chapman, S.C. (2023). Nonpathological inflammation drives the development of an avian flight adaptation. Proc Natl Acad Sci U S A *120*, e2219757120.
2. Yamagishi, A., Matsumoto, S., Watanabe, A., Mizuguchi, Y., Hara, K., Kan, H., Yamada, T., Koizumi, M., Shinji, S., Matsuda, A., et al. (2014). Gene profiling and bioinformatics analyses reveal time course differential gene expression in surgically resected colorectal tissues. Oncol Rep *31*, 1531-1538.
3. Huang, L.H., Lin, P.H., Tsai, K.W., Wang, L.J., Huang, Y.H., Kuo, H.C., and Li, S.C. (2017). The effects of storage temperature and duration of blood samples on DNA and RNA qualities. PLoS One *12*, e0184692.
4. Matsubara, T., Soh, J., Morita, M., Uwabo, T., Tomida, S., Fujiwara, T., Kanazawa, S., Toyooka, S., and Hirasawa, A. (2020). DV200 Index for Assessing RNA Integrity in Next-Generation Sequencing. BioMed research international *2020*, 9349132.
5. Loi, D.S.C., Yu, L., and Wu, A.R. (2021). Effective ribosomal RNA depletion for single-cell total RNA-seq by scDASH. PeerJ *9*, e10717.
6. Lee, G.-Y., Ham, S., and Lee, S.-J.V. (2024). Brief guide to RNA sequencing analysis for nonexperts in bioinformatics. Molecules and Cells *47*, 100060.
7. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res *49*, D325-d334.
8. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.
9. Luo, W., Pant, G., Bhavnasi, Y.K., Blanchard, S.G., Jr., and Brouwer, C. (2017). Pathview Web: user friendly pathway visualization and data integration. Nucleic Acids Res *45*, W501-w508.