RNA purification and expression analysis using microarrays and RNA deep sequencing

Steven R. Head[1], Tony Mondala[1], Terri Gelbart[2], Phillip Ordoukhanian[1], Rebecca Chappel[1], Gilberto Hernandez[1], and Daniel R. Salomon[2]

[1] Microarray and Next Generation Sequencing Core Facility, The Scripps Research Institute, La Jolla, CA

[2] Laboratory for Functional Genomics, Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA

Address Correspondence to:

shead@scripps.edu
dsalomon@scripps.edu

Summary

Transcriptome analysis or global gene expression profiling is a powerful tool for discovery as well as understanding biological mechanisms in health and disease. We present in this chapter a description of methods used to isolate mRNA from cells and tissues that has been optimized for preservation of RNA quality using clinical materials and implemented successfully in several large, multi-center studies by the authors. In addition two methods, gene expression microarrays and RNAseq, are described for mRNA profiling of cells and tissues from clinical or laboratory sources.

Key words: RNA Purification, mRNA, Global Gene Expression Profiling, Microarrays, Next Generation Sequencing, Deep Sequencing, RNAseq

---

1. Introduction

The application of transcriptome analysis, essentially global gene expression profiling, using microarrays as a method for identification of biomarkers as well as revealing insights into biological mechanisms has been well documented (Verweij CL, 2009; Cooper-Knock J, 2012). Newer technologies for transcriptome analysis include the use of high throughput or “Next Generation” deep sequencing. A major question at this point is the value and optimal applications for these newer technologies (Malone, 2011). In this chapter, we will describe protocols for gene expression profiling using microarrays and RNA deep sequencing. We cannot answer the question for the reader about what is the optimal technology for their purposes. The choice really depends on multiple factors at this point including both strategy and costs. However, neither approach is dramatically superior to the other for straightforward gene expression profiling but we will comment on the relative merits of each.

Critical to generating useful and reproducible data is the extraction of high quality RNA so this subject is covered in the first section. We describe methods for preserving and extracting high quality RNA from various commonly used sources including tissue samples such as biopsies, cell pellets collected from cultured cells or peripheral blood mononuclear cell preparations. Whole blood RNA, commonly used for clinical studies and biomarker discovery projects, is prepared from blood drawn into specialized vacutainer tubes (PaxGene Blood RNA Tubes, Qiagen) designed specifically for preserving total RNA from all the cells comprising whole blood. The PaxGene Blood RNA preparation protocol described here utilizes the automated Qiacube robotic workstation to facilitate more efficient processing of multiple samples. As a general comment, when properly implemented, automated methods have the specific value of

reducing the impacts of human errors and the inevitable variations in data, a common source of batch effects, when multiple technicians are doing individual parts of a complicated experiment.

In the second section of this Chapter, we describe the gene expression profiling of high quality RNA preparations using high-throughput microarray technology – specifically the Affymetrix Gene 1.1 ST Array Plates processed on the GeneTitan system. The Affymetrix GeneTitan is a high throughput automated array system that is capable of processing two plates in a single run where each plate is comprised of 16, 48 or 96 arrays. However, the methods described here for RNA labeling are also applicable to the use of the older, single cartridge arrays.

Finally, we describe a protocol for gene expression profiling by deep RNA sequencing. This involves preparation of an RNAseq library from high quality total RNA samples using a specific NuGEN (San Carlos, CA) kit for analysis on the Illumina HiSeq2000 deep sequencing platform. One key advantage of RNAseq relative to microarray technology is that sequencing generates data free from any bias of what may or may not be represented with specific probes on any given commercial microarray. RNAseq can also provide the opportunity to run multiple samples in parallel in a single lane through the use of barcoded libraries – libraries with short sequence tags that identify their original sample source. Thus, we routinely run 6 to 12 samples per lane on an 8 lane HiSeq2000 flow-cell. At this depth of coverage, we have determined that the results from RNAseq are comparable to the gene expression profiling results obtained with the Affymetrix 1.1 ST Arrays. Nonetheless, these two technologies are in such constant evolution that it is not possible yet to conclude which of these two approaches to expression profiling is the best.

An important point is that we recognize that many readers are not going to be performing the entire workflow from RNA purification to the actual microarrays or RNA deep sequencing. In most instances, the latter stages of the pipeline will be done by trained technicians in specialized Core facilities. However, we strongly believe that investigators need to understand many of the details of the workflow in order to take responsibility for experimental design, effectively communicate with Core facility staff and ultimately to understand the value and the limitations of the data obtained.

---

2. RNA Purification from Various Samples

2.1 Materials:

RNA purified from cultured cells, cells isolated by flow cytometry, as well as peripheral blood mononuclear cells obtained from Becton Dickinson Vacutainer Cell Preparation

Tubes (CPT) are performed using the Trizol hybrid protocol. This hybrid protocol allows the purification of total RNA, comprised of both mRNAs as well as small RNAs (e.g. microRNAs). The latter are important to researchers interested in the expression and functional roles of regulatory, non-coding small RNA. This protocol can also be used to isolate RNA from tissues preserved in the commercial preparation, RNAlater (available from both Qiagen and Ambion/Life Technologies). The extraction of total RNA including small RNA from whole blood stabilized in Qiagen RNA Blood PAXgene tubes is performed using the Qiagen miRNA Blood Extraction Kit. This is an automated protocol available on the QIAcube, a robotic workstation. We also recommend that users have access to an Agilent BioAnalyzer or equivalent instrument in order to determine RNA quality and estimate fragmentation after purification protocols.

2.1.1 RNA Extraction from Cells and Tissue

1. Trizol Reagent (Life Technologies)
2. Battery Operated Hand Homogenizer/Pellet Mixer (VWR)
3. 1.5 ml Pestle RNase and DNase free (Fisherbrand)
4. Chloroform (Sigma)
5. Ethanol 100% (Sigma)
7. RNase Zap (Life Technologies)

2.1.2 RNA Extraction from Blood PaxGene Tubes

1. PAXgene Blood RNA Tubes (Qiagen)
2. PAXgene Blood miRNA Kit (Qiagen)
2. QIAcube Rotor Adaptors (Qiagen)
3. Isopropanol (Sigma)
4. Ethanol 100% (Sigma)

2.2 Methods:

2.2.1 RNA Extraction from Cells and Tissue

1. Homogenize tissue samples (up to 100mg) in 0.2ml of Trizol reagent with battery operated hand homogenizer (VWR Pellet Mixer Cat # 47747-370 and Fisherbrand 1.5 ml Pestle RNase and DNase free Cat #V7339-901). Then bring to 1ml with Trizol. Aspirate through 21g syringe several times. Incubate for 5 min at room temperature.

2. Add 200μl of Chloroform. Shake vigorously for 15 sec and incubate at room temperature for 3 min.

3. Centrifuge at 12,000xg for 15 min at 4-8°C.

4. Transfer the top aqueous phase to fresh tube and save remaining sample for DNA and/or protein extraction.

5. For small RNA inclusion slowly add 1.4 volumes of 100% RNase-free EtOH instead of equal volume. Need at least 60%EtOH, mixing as needed.

6. Load the sample (up to 700μl) into an RNeasy column seated in a collection tube and spin for 30 sec at 8,000xg. Discard flow-through.

7. For small RNA inclusion add 700μl of RPE Buffer onto the column and spin 30 sec at 8,000xg. DO NOT USE RW1 use only RPE. Make sure ethanol has been added to the RPE buffer. Discard flow-through.

8. Transfer column into a new collection tube, add 500μl buffer RPE and spin for 30 sec at 8,000xg. Discard flow-through. Make sure ethanol has been added to the RPE buffer before use.

9. Add 500μl buffer RPE and spin 2 min at 8,000xg. Discard flow-through.

10. Spin the column for 1 min at 8,000xg to get rid of remaining buffer in the column.

11. Transfer the column to a new 1.5ml collection tube and pipet 30-50μl of RNase-free water directly onto the column membrane. Allow the sample to sit at room temperature for 1-2 min and then spin 1 min at 8,000xg to elute RNA. Quantitate on a Nanodrop spectrophotometer and check the quality on an Agilent Bioanalyzer.

12. Store RNA at -80°C until use.

Protocol Notes: Bring Trizol reagent to room temperature. RPE Buffer is supplied as a concentrate in the RNeasy kit. Add appropriate volume of 100% ethanol before using for the first time. Work in the fume hood. Do not worry about RNases in steps 1-3. Your sample is full of them anyway. They are inhibited as long as they are in the Trizol. From step 4 you should be careful to not contaminate the samples with RNases. Keep cleaning your gloves with RNase Zap through the whole process. The main source of contamination comes from your fingers by accidentally touching the inner part of the tube caps. While discarding flow-through in steps 6-9, avoid touching the mouth of the collection tubes with anything.

2.2.2 RNA Extraction from Blood PaxGene Tubes Using the QiaCube Robot (Figure 1)

The PAXgene Blood RNA tubes collect 2.5 ml of whole blood into a sterile vacutainer tube containing a proprietary additive that near instantly stabilizes the *in vivo* gene transcription profile by blocking RNA degradation while preventing the cellular gene induction that otherwise occurs *ex vivo* following blood drawing. Blood samples collected in PAXgene tubes can be stored at room temperature for 72 hours, at 2-8 C for up to 5 days, or at -20C or -80C for at least 50 months without significant RNA degradation. After blood collection the PAXgene tube should be well mixed and incubated at room temperature for two hours with occasional mixing to ensure all the cells lyse.

Samples are processed using the Qiagen PAXgene miRNA kit that preserves both mRNA and miRNA after purification. Several manual steps are involved prior to placing the samples on the robot including centrifuging, washing and resuspending the nucleic acid pellet (Steps 1-6 below). Batches of 12 samples can be run in less than 3 hours. Typical total RNA yield from healthy human whole blood is 3-7μg.

1. Thaw the PAXgene tubes if frozen. The thawed tubes should then be mixed and stored at room temperature for 2 hours with occasional mixing by gentle inversion or using a motorized rotator at the lowest speed.

2. Centrifuge the mixed PAXgene tubes for 10 minutes at 3000-5000g using a swinging bucket rotor.

3. Remove the supernatant by decanting. Add 4 ml RNase-free water to the pellet and close the tube with a clean cap.

4. Vortex until the pellet is dissolved and centrifuged for 15 minutes at 3000-5000g. Our centrifuge is limited to a maximum of 3500g. Remove the supernatant completely and let the tube drain to remove all traces of liquid.

5. Add 350μl of BM1 Buffer reagent. Vortex until the pellet is dissolved.

6. Pipet the sample into a 2ml processing tube and load onto the QIAcube shaker.

7. The QIAcube is loaded with the remaining buffers, Proteinase K, DNase I, and Buffer BR5 for elution.

8. The QIA cube rotor adaptor is loaded with a PAXgene RNA spin column, PAXgene Shredder spin column, and a microfuge sample tube in the designated positions.

9. The PAXgene miRNA protocol A is begun. When Protocol A is complete the microfuge tube containing the eluted RNA is transferred to the heating block in the QIAcube and a short 10 minute Protocol B is run to heat the sample for 5 minutes at 65C to denature the RNA for downstream reactions.

10. The RNA is then quantified on a Nanodrop spectrophotometer and the quality is assessed on an Agilent Bioanalyzer (**Figure 2**)

Please refer to the Qiagen PAXgene Blood miRNA Kit Handbook for the latest information available. This is available at the Qiagen website: http://www.qiagen.com/. If a QiaCube Robot is not available, simply follow the workflow described for the manual version of the protocol at the same site. In this case, it is recommended that all samples in one phase of a study be done by the same person and in a random order to minimize technical batch effects.

2.3 Notes:

The current trend in profiling clinical samples for global gene expression analysis is to isolate total RNA including small or microRNAs. The co-isolation of protein as well as DNA and RNA from cells and tissues can also be done by including additional steps into the Trizol hybrid method (Bai, 2012). These combination protocols were designed to take maximal advantage of precious clinical samples. They also represent the increasing and strategic use of multiple orthogonal technologies to advance a research objective (e.g. gene profiling and proteomics).

We exclusively use RNAlater for the preservation of RNA in tissue samples. One important consideration is that it takes some time for the RNAlater to permeate a tissue section or biopsy core to work optimally. Thus, it is important to consider the manufacturers guidelines for tissue size and volumes of RNAlater. Also, to insure the preservation of tissue RNA, we routinely leave a sample in RNAlater for at least 30 minutes at room temperature before freezing though it can sit overnight and be absolutely fine. In situations where we are sorting cells by flow cytometry or separating by magnetic beads and when using non-adherent cell lines, we go directly into Trizol instead. The disadvantage of Trizol is mainly the fact that it is a hazardous material and while not an issue in a single laboratory setting, it can be a serious challenge in a multicenter clinical study where shipping samples for archiving and analysis is necessary.

Whole blood samples present a unique challenge to mRNA expression due to the huge number of red blood cells containing hemoglobin-related transcripts in proportion to the relatively small number of other circulating, nucleated cells. The concern has been that these globin transcripts reduce the sensitivity and signals of the other RNA transcripts that are the target for most gene expression profiling experiments. The good news is

that we have demonstrated in our laboratory that globin reduction does not give additional benefits in terms of signal intensities or the sensitivity of differential gene expression when using the Affymetrix 1.1 ST peg arrays in combination with the current generation of labeling kits (unpublished data: http://www.genetics.ucla.edu/transplant-genomics/protocols/). Considering the added cost, time, effort and additional loss of RNA during the globin reduction steps, we don't recommend this procedure.

The next question is whether globin reduction is currently required for gene expression profiling by RNAseq experiments. One recent paper suggested there was a value in globin reduction resulting in increased sensitivity (Mastrokolias 2012). We are currently in the process of testing this in our laboratory and the results will be posted at our web site when available (see URL above).

---

3. Global Gene Expression Profiling Using Affymetrix DNA Microarrays - Gene 1.1 ST Array Plates

Gene expression profiling on the Affymetrix GeneTitan system requires that high quality total RNA samples be prepared for hybridization to the expression arrays (**Figure 3**). This involves following either one of the target preparation protocols described below (the Ambion WT Expression kit with the Affymetrix GeneChip WT Terminal Labeling Kit or the NuGEN Ovation Pico WTA V2 System with Biotin Module). We generally use the Ambion WT Expression kit when we have adequate amounts of total RNA. As little as 50ng of starting material can be used. When only limited amounts of total RNA are available, we use the NuGEN Ovation Pico WTA V2 system with Biotin Module. We have used this protocol successfully with as little as 0.5ng total RNA.

Finally, once the target RNA has been prepared and labeled, it is ready to hybridize to Affymetrix 1.1 ST array plates available in 16, 24 and 96 peg formats.

3.1 Materials:

1. Ambion WT Expression Kit (Life Technologies)
2. Affymetrix WT Labeling and Controls Kit (Affymetrix)
3. Affymetrix HT HWS Kit for GeneTitan and WT Array Plates
4. NuGEN Ovation Pico WTA System V2 (NuGEN)
5. NuGEN Encore Biotin Module (NuGEN)
6. Affymetrix GeneChip Expression Hybridization Controls (Affymetrix)

7. Affymetrix HT HWS Kit for GeneTitan and WT Array Plates

3.2 Methods:

3.2.1 Ambion WT Expression Kit

The WT Expression Kit is designed to generate amplified sense-strand cDNA ready for fragmentation and labeling using the Affymetrix GeneChip WT Terminal Labeling Kit. The labeled product is generated in about 3 days at which point it is for ready for hybridization to the Affymetrix GeneChip ST (Sense Target) Arrays. The starting requirement is 50-500ng of total RNA. For example, some sources suggest a yield of 10pg total RNA/cell, thus, approximately 5000-50,000 cells would be required using this protocol for global expression profiling. In our experience a more reasonable expectation for yield is 1pg to 5pg per cell. The key is to determine this yield experimentally for your chosen cell and protocol.

The WT Expression Kit uses a reverse transcription priming method that specifically primes non-ribosomal RNA from your sample. This is important because otherwise ribosomal RNA is so abundant in all cells that it would interfere with detection of mRNAs in the sample. The cDNA is then *in vitro* transcribed (IVT) to amplify the target and finally the RNA generated in the IVT reaction is copied back into cDNA with dUTP incorporated for the downstream fragmentation step described below.

Please refer to the manufacturer's literature: "*The Ambion® WT Expression Kit: For Affymetrix® GeneChip® Whole Transcript (WT) Expression Arrays*" available on the Life Technologies website for current protocol details and instructions (URL: www.**lifetechnologies**.com/).

3.2.2 Fragmentation and Labeling of the cDNA generated from the WT Expression Kit

In the next step of the process, the Affymetrix WT Labeling and Controls Kit (Affymetrix) is used to fragment and label the cDNA generated from the Ambion WT Expression Kit. Fragmentation is accomplished through treatment of the cDNA with uracil-DNA Glycosylase (UDG) and apurinic/apyrimidinic endonuclease (APE), which effectively cleave the cDNA wherever a uracil was incorporated during cDNA synthesis. After fragmentation, the cDNA fragments are labeled using terminal deoxynucleotidyl transferase (TdT) to incorporate a biotin label onto the 3' ends. At each step 1μl of the sample should be saved so that the fragmentation and labeling steps can be checked on the Bioanalyzer (see **Figure 4**). The fragmented sample, before and after labeling, should be run side by side for comparison. The fragmentation should result in a band with the size of 40-70 nucleotides and the subsequent fragmented and labeled cDNA

should show a slight shift to a larger size due to attachment of the labels. The amount of fragmented and labeled cDNA needed for each Affymetrix ST Peg array is 2.8μg.

Please refer to the manufacturer's literature: "*GeneChip® WT Terminal Labeling and Controls Kit*" available on the Affymetrix website for the most current protocol instructions (URL: www.**affymetrix**.com/).

3.2.3 NuGEN Ovation Pico WTA System V2

As an alternative to the protocol described above for preparation of total RNA samples for hybridization to Affymetrix gene expression arrays, the NuGEN Ovation Pico WTA V2 System can be used to prepare labeled cDNA from a range of 0.5-50ng of total RNA input. The NuGEN Encore Biotin Module is used to fragment and label the target in a manner analogous to the Affymetrix WT Labeling protocol described above in 3.2.2. The labeled product is ready for hybridization to microarrays in about 2 days.

The primer mix contains a unique mixture of random and oligo dT primers such that priming occurs across the whole transcript. The NuGEN kits use a linear amplification Ribo-SPIA technology based on SPIA (Single Primer Isothermal Amplification) for amplification. This method uses DNA/RNA chimeric primers (SPIA primers), DNA polymerase and RNase H in an isothermal assay. Since many NuGEN protocols for RNA amplification use their universal SPIA primers it is important to note that there is a potential for the generation of non-specific amplification products from carry-over (contamination) of previously amplified SPIA cDNA. Thus, care must be taken to avoid this pitfall if the laboratory chooses to adopt this approach. Precautions similar to those used to prevent PCR contamination are absolutely required.

Please refer to the NuGEN Ovation Pico WTA System V2 User Guide available on the NuGEN website for the most current protocol instructions (URL: www.nugeninc.com/nugen/).

3.2.4 Target Hybridization and Processing for Gene 1.1 ST Peg Arrays on the Affymetrix GeneTitan Instrument

Materials and Reagents

Gene Titan Hybridization, Wash, and Stain Kit (Affymetrix)

Gene Titan 1.1 ST Peg arrays in 16, 24, or 96-array formats (Affymetrix)

96-well PCR plate that can withstand heating to 95C (e.g. BioPioneer, GSO2918)

8 or 12 strip caps for PCR plates (e.g. BioPioneer, PCR-SCF-SRCAP)

Centrifuge equipped with microtiter plate rotor (e.g. Eppendorf 5810R)

Methods

Preparation of Hybridization Cocktail: Please refer to the Affymetrix "*Target Hybridization for Gene 1.1 ST Array Plates Processed on the GeneTitan Instrument*" user guide from the Affymetrix website for the most current protocol instructions.

1. Prepare master mix consisting of the components of the GeneTitan Hybridization kit following kit instructions for the number of samples being hybridized.

2. Add 39.2μl master mix to the wells of a 96-well plate followed by addition of 2.8μg of the fragmented and labeled cDNA in 32.8μl volume from each sample. Bring volume in each well to 120μl by addition of 48 microliters of the *2.5X WT Hyb Add 6* reagent that comes in the GeneTitan Hybridization kit.

3. Seal 96-well plate with strip caps and vortex briefly to mix. Briefly centrifuge plate to bring liquid to bottom of wells.

4. Heat the sealed 96-well plate to 95C for 5 minutes and then 45C for 5 minutes and centrifuge at high speed for 1 minute.

5. Transfer 90μl of each sample into each well of the hybridization plate included in the Hybridization kit.

Preparation of GeneTitan Instrument

The GeneTitan integrates hybridization, washing, and imaging in a single instrument. It supports many types of Peg arrays including 3' expression arrays, whole transcript arrays, SNP genetics arrays and genome-wide screening arrays. In addition, there are three formats of Peg plates comprised of 16, 24 or 96-arrays per plate. We will limit our current discussion to global gene expression profiling using the Whole Transcript 1.1 ST Peg arrays.

The instrument is initialized, which turns on the hybridization oven, prepares the fluidics for staining and washing and subsequently prepares the laser for scanning. Plate and sample identification files are uploaded to the instrument's computer system. Trays and lids are treated with an anti-static gun. Three stain trays are prepared with 105μl of stain in each well and a scan tray is filled with 150μl per well of an array holding buffer. The three stain plates, scan plate, hybridization plate, and the peg array plate are loaded into the GeneTitan instrument as prompted by the running computer program. The sample will now automatically be incubated at 48C for 17 hours, then stained and washed and scanned. The initial hybridization cocktail preparation and loading of the

plates into the Titan takes about an hour and then the entire time for hybridization, staining, and scanning is 20 hours for a 16 or 24 Peg array plate and 24 hours for a 96 Peg array plate. A second Peg array plate can be loaded once the first Peg array has begun the process. A delay in starting the process for the second Peg plate is necessary since each array should be scanned immediately after the stain and wash step is completed and there is only one scanning station. The delay is 5 hours if the first plate is a 96 Peg array and the delay is two and a half hours if the run is started with 16 or 24 Peg arrays.

3.3 Notes:

While a comprehensive discussion of experimental design for microarray (and RNAseq) experiments is beyond the scope of this book chapter, a few important principles should be noted. The principles of a good experimental design will apply to any microarray experiment whether using the Affymetrix arrays described here or alternative array technologies from companies such as Illumina and Agilent.

Good experimental design dictates the use of adequate sample sizes for statistical analysis and statistical power. We have found in practice with real clinical samples that a minimum of 20 samples per group in a study is necessary for statistical power. That does not mean that a trial or discovery run cannot be done with as few as 5 samples per group but care should be taken in not over-interpreting the significance of any specific result obtained unless these are validated using another technology (e.g quantitative PCR or proteomics) or multiple replications of the same experiment are done with many more samples.

Another important consideration is proper organization of samples and controls for processing to mitigate batch effects while effectively addressing the biological questions of interest. The entire process from RNA isolation to labeling to hybridization to the actual printing of the plate by the manufacturer can contribute different kinds of technical noise. This noise creates bias in the data that is not reflecting the biology that is being studied. This kind of data bias or batch effects can result from many different choices made in the process. For examples, when all the controls are done on one plate and all the experimental samples on another. Even if all samples are done on a single plate but they are prepared and labeled for hybridization on different days or by different technicians or even different kit lots, batch effects will result. It is critical for the investigator to plan out the entire organization of processing all the experimental samples before starting. However, a key point is that even given all this care, the nature of these highly complex technologies is that they generate some batch effects and

these must be effectively compensated for during the statistical analysis of the data later.

A number of new protocols are being developed to allow successful profiling of samples obtained from biopsies that have already been fixed and embedded. These allow access to large archives of samples that are clinically invaluable such as tumor biopsies and autopsy specimens and this is especially important for profiling rare samples. However, the analysis of formalin fixed paraffin embedded (FFPE) samples or other samples with degraded RNA presents unique challenges and requires alternative approaches to preparing the RNA for hybridization to microarrays and will not be discussed here.

Finally, the analysis of microarray data is complex and the reader is advised to involve an experienced statistician/bioinformatics expert for this task. Standardized analysis methods are frequently used for identification of differentially expressed genes, determination of p-values and false discovery rates as well as clustering, generation of heat maps and running class prediction algorithms.

---

4. Global Gene Expression Profiling Using RNA Deep Sequencing (RNAseq) on an Illumina HiSEQ2000

The rapid evolution of high-throughput deep sequencing technology has created new opportunities to do global gene expression profiling using RNAseq. A number of issues must be considered by investigators before deciding on whether to do any given experimental study using this new approach. The isolation and required amounts of total RNA is roughly the same for both microarrays and deep sequencing. However, as can be clearly seen when the previous section is reviewed, the workflows are very different. Microarrays take total RNA, label the mRNA transcripts, hybridize them to arrays of printed probes and then detect the amount of fluorescent signal excited by a laser scan to reveal the level of gene expression in the pool of RNA. Deep RNA sequencing starts with creating a tagged library of all the mRNA transcripts in the RNA pool.

The preparation of the sequencing library adds two key elements. The first is a set of adaptors on each end of the double stranded cDNA created that is required for binding to complementary oligonucleotides attached to the surface of the sequencing flow cell. Binding to these oligonucleotides positions each transcript in the library in the correct orientation for sequencing. The second element added is the unique barcodes. These are short DNA sequences that uniquely identify the sample and are subsequently decoded in the statistical analysis of the sequencing data. The use of barcodes enables the sequencing of multiple samples in a single lane on the flow cell and that ability to multiplex has made RNAseq a cost-effective alternative to microarrays. Nonetheless,

the workflow for RNAseq is more complicated, operating the technology platforms is considerably more demanding and the initial stages of data analysis are more computationally intensive. Another key point is that gene expression data obtained by deep RNA sequencing compared to the latest generation of microarrays reveals that both are highly correlated and valuable. Thus, there is no reason for an *a priori* decision to choose RNAseq over microarrays.

Though beyond the scope of the present chapter, RNAseq can go far beyond just gene expression profiling. Using long paired-end reads a novel transcriptome can be studied and this is now driving new research into viral, bacterial and fungal microbiomes. Paired-end reads can also be used to align intron-exon boundaries to reveal alternative splicing, a major source of transcriptional and proteomic diversity in eukaryotic biological systems. Because RNAseq provides the actual sequences of the mRNA transcripts, there is a potential to discover genetic mutations (e.g. single nucleotide polymorphisms) in transcribed regions. Finally, we believe that the quality of the data and the economics of doing miRNA profiling by miRNAseq using high levels of barcode multiplexing is now the best approach.

4.1 Materials

1. Total RNA sample
2. NuGen Encore® Complete RNA-Seq DR kit – provides all reagents, enzymes, buffers, barcoded adapters, and beads for library prep (NuGEN, Santa Carlos, CA, USA)
    a. Multiplex barcode system 1-8 (PN 0333-32)
    b. Multiplex barcode system 9-16 (PN0334-32)
3. Covaris S2 sonication system (Covaris, Woburn, MA, USA)
4. Covaris microTUBE (Covaris, Woburn, MA, USA)
5. 0.2mL PCR tubes (Corning Thermowell®, Tewksbury, MA)
6. Ethanol for bead purifications (Pharmco-AAPER, Brookfield, CT)
7. Magnetic separation stand (Invitrogen™, Life Technologies, Carlsbad, CA)
8. Qubit® Fluorometer 2.0 (Invitrogen™, Life Technologies, Carlsbad, CA)
9. Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA)
10. E-Gel® with SYBR Safe™ precast gel (Invitrogen™, Life Technologies, Carlsbad, CA, USA)
11. TrackIt™ 100bp DNA Ladder (Invitrogen™, Life Technologies, Carlsbad, CA, USA)
12. Scalpel for gel excision
13. Zymoclean™ Gel DNA Recovery Kit (Zymo Research, Irvine, CA, USA)
14. Illumina SR or PE Cluster Kit (Illumina, Inc., San Diego, CA, USA)
15. Illumina Cbot
16. Illumina HiSeq 2000

4.2 Methods

The RNAseq library is generated using the NuGen Encore Complete DR kit following their recommended protocol (URL: www.nugeninc.com/nugen/).

1. After initial QC of RNA by Bioanalyzer tracing, 100-150ng of total RNA is taken into the Encore Complete prep.
2. First strand cDNA is generated from the total RNA using the provided first strand primer, buffer, and enzyme mix in the Encore Complete kit. The first strand primer is annealed at 65C for 5 minutes. The enzyme and buffers are added to the sample and incubated at 40C for 30 minutes.
3. Immediately following, second strand cDNA is generated using the provided second strand buffer and enzyme mix. The reaction is held at 16C for 60 minutes. A stop solution from the kit is added to stop the enzymatic reaction.
4. The now double stranded cDNA is fragmented using an S2 Covaris to a median size of 200bp.
5. The fragmented cDNA is purified using 1.8 volumes of Agencourt RNAClean XP beads and a 70% ethanol bead wash solution.
6. End repair buffer and enzyme mix from the kit are added to the sample and incubated at 25C for 30 minutes and then 70C for 10 minutes. This step creates blunt ends of the cDNA allowing for efficient ligation of adapters.
7. Barcoded adapters are ligated to the cDNA using one of the sixteen barcoded adapters provided in the kit along with the ligation buffer and enzyme mix. The reaction is held at 25C for 30 minutes.
8. After ligation, two strand selection steps are completed to end up with only one cDNA strand with adapters. The first step is completed with strand selection I enzyme mix and buffer provided in the kit. The reaction is held at 72C for 10 minutes.
9. The sample is then purified using 1.8 volumes of Agencourt RNAClean XP beads and a 70% ethanol bead wash solution.
10. The second strand selection step is completed with strand selection II enzyme mix and buffer provided in the kit. The reaction is held at 37C for 30 minutes and then 95C for 30 seconds.
11. The library is amplified using the DR primer mix, DMSO, buffer and enzyme mix provided in the kit. 20 total cycles; 5 with an annealing temperature of 55C and 15 with an annealing temperature of 63C are completed.
12. A final bead purification of the amplified library is completed using 1.2 volumes of Agencourt RNAClean XP beads and a 70% ethanol bead wash solution.
13. The library is run on an Agilent 2100 Bioanalyzer DNA 1000 chip (**Figure 5)**. The final product has a size range of 200-400bp with characteristic peaks around 285bp. The spectra are used to validate the library as well as determine the presence or absence of adapter dimer. The dimer has a size of around 130bp.
14. Qubit fluorometer is used to measure the concentration of the final library.

15. For multiplex sequencing, Qubit concentrations are used to equally pool the libraries. Each pool is gel purified and size selection is used to obtain the desired insert size for sequencing. Paired-end 2x100 sequencing, for example, would require an insert size of >200bp. Material would be cut from the gel between 330-430bp to obtain insert sizes of 200-300bp including 130bp of adapter.
16. Using 2% Invitrogen precast gels, the samples are loaded into 2-3 lanes alongside a 100bp DNA Track-It ladder.
17. The gel is ran for 20 minutes and viewed under a transilluminator. Using a scalpel, the section of gel is selected between 330-430bp for a 200bp insert. (**Figure 6**).
18. The gel slice after excision is dissolved using Zymo ADB buffer and purified using a Zymo (25μg) column and eluted in water.
19. A final QC of the library is done using the Bioanalyzer and quantified using the Qubit fluorometer.
20. The libraries for each lane are placed into the cBot (**Figure 7a**) that loads the flow cell and begins the process of clustering required to initiate the sequencing protocol.
21. The clustered flowcell is then loaded onto the Illumina HiSeq 2000 for sequencing (**Figure 7b**).

4.3 Notes

One important question for any new technology is what metrics can be used to judge the quality of an RNA deep sequencing run. The Illumina HiSeq system provides real-time information on read quality as it is being generated and these will be monitored by the technical staff in the Core facility. When the data is delivered to investigators the important metrics to review include pass filter rate (ideally above 85%), total number of reads per lane (ideally 150-200 million per lane) and alignment to the genome (can vary with preparation but should be 70-80% using the protocol described here). However, it is worth noting that if a project involves less standard protocols or sample sources then the expectations for results must be adjusted. For example, we do a challenging protocol called RIPseq that first involves the immunoprecipitation of RNA-binding proteins bound to mRNA transcripts and then deep RNA sequencing of the target mRNAs and the extremely low yields of mRNA presents challenges at multiple points in the workflow.

A decision that must be made at the start of every sequencing project is the read length. For simple RNA expression profiling, the objective is to detect an mRNA and determine relative abundances. This can be done with single reads of 50 to 100bp in length. However, in other situations, we choose to do paired-end read sequencing, usually at 100bp each (e.g. 2x100). For example, we routinely do paired-end reads for alternative splicing analysis because it is very useful in identifying intron/exon splice junctions

during analysis. We also use paired-end sequencing for RNA expression profiling of non-human primates because the analysis pipeline requires alignment to much less complete maps of the nonhuman primate genomes. A key step in the process is matching sequence reads to annotated genes based on homology to their human counterparts and longer, paired-end reads improve the alignment quality.

The last major issue to note is that of “read depth”. This term is calculated by the total number of reads multiplied by the fraction (%) of reads that align to the genome. The key point is that deep sequencing generates a huge amount of data but the important data is only that which contributes to your experimental objective. For example, we can generate a billion reads with one RNAseq sample in a single run. The opportunity to do multiple samples in a single lane using barcodes allows considerable cost savings and increased efficiency. So the real question is how many reads are actually needed to determine the global gene expression profile? Our experience with activation of purified human T cells indicates that 10 million aligned reads gives results that are comparable to profiling with the latest generation of Affymetrix microarrays. Thus, we can easily run 12 different samples in a single lane for RNAseq. For the profiling of alternative splicing the read depth needs to be considerably greater, at least in the 30 million aligned read range. Therefore, for each project, investigators are encouraged to critically evaluate the required read depth to accomplish the objective.

Figure Legends

Figure 1. The QIAcube, a robotic workstation from Qiagen.

Figure 2. BioAnalyzer (Agilent) trace of high quality RNA sample. X axis shows size in bases (nucleotides; nt), Y axis shows relative fluorescent units. The peak at approximately 25 nt is a size marker. The larger peaks near 1,700 nt and 4,000 nt show the two major ribosomal RNA transcripts.

Figure 3. The GeneTitan system (Affymetrix) used to hybridize, wash, stain and scan labeled samples to Affymetrix 1.1 ST array plates in 16, 24 and 96 Peg array formats.

Figure 4. BioAnalyzer trace showing the fragmented samples, before (lane 1) and after labeling (lane 2). A sizing ladder is also shown (lane L). The fragmentation results in a band with the size of 40-70 nucleotides (i.e. bases; nt) and the subsequent fragmented and labeled cDNA shows a slight shift to a larger size due to attachment of the labels.

Figure 5. Bioanalyzer trace of a successfully completed Nugen Encore Complete DR RNAseq library.

Figure 6. Gel purification of a library for paired-end 2x100 sequencing run on a 2% agarose gel.

Figure 7. (**A**) The Illumina cBot used to cluster libraries onto flow-cells for sequencing and (**B**) the Illumina HiSeq 2000 sequencing system used to generate deep sequencing data.

References

1. Verweij CL (2009). Transcript profiling towards personalised medicine in rheumatoid arthritis. *Neth J Med.* Dec;67(11):364-71.

2. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, Shaw PJ (2012). Gene expression profiling in human neurodegenerative disease. *Nat Rev Neurol.* Aug 14. doi: 10.1038/nrneurol.2012.156. [Epub ahead of print]

3. Malone JH, Oliver B (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. May 31;9:34.

4. Bai B, Laiho M (2012). Efficient sequential recovery of nucleolar macromolecular components. *Proteomics*. Aug 14. doi: 10.1002/pmic.201200071. [Epub ahead of print]

5. Mastrokolias et al.: Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. BMC Genomics 2012 13:28.