CHAPTER 12

Using Transcriptomics to Study Behavior

J.T. Westwood
University of Toronto, Mississauga, ON, Canada

INTRODUCTION

During the lifetime of a multicellular animal, it will experience or perform a number of behaviors related to feeding, reproduction, nurturing of offspring, avoidance of danger, responses to changes in their environment, and social interactions, just to name a few. In addition, there are a number of conditions or diseases that can affect the well-being and overall behavior of individuals. There are good examples of specific genes that play a role in some of the aforementioned behaviors. For example, the *egr1* gene, which codes for a transcription factor and thus can regulate the transcription of downstream genes, has been shown to be transcriptionally induced in response to social interaction and plays a role in zebra finch song recognition during courtship, male dominance in male cichlid fish, and mothering behavior in rats (see Robinson et al.[1] for review). Some other examples include the *period* gene in *Drosophila* male courtship and the foraging gene in *Drosophila* and honeybee feeding behavior.[1]

The fact that different alleles of a gene can affect behaviors of an animal in the long run, and that the products of these genes are often transcription factors, implies that the changes in gene expression (i.e., changes in gene transcription rates) likely underlie and/or modulate the behavior. Therefore, it is logical to study gene expression changes that may underlie factors that influence behavior and brain function. Examples of factors that influence behavior and brain function include (but are not limited to) are: social interactions; different alleles of the same gene; and changes in the environment.

There are some well-studied models for specific behaviors that have been developed in insects. For some of these behaviors, gene expression studies have been conducted to gain insight into the molecular mechanisms that underlie the behavior. For example, honey bees have been a model for studying many kinds of behaviors (or factors affecting them) including, but not limited to differences in the behavior of different members of bee "society" (e.g., male drones, different workers such as nurses, guards, and foragers, and the queen), aggressive behavior in response to intruders, and pheromone and hormone manipulations that affect bee maturation and behavior. All of the aforementioned models have been examined in gene expression studies, and hundreds of reproducible

Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research
ISBN 978-0-12-804078-2, https://doi.org/10.1016/B978-0-12-804078-2.00012-X

changes in gene expression have been identified and could be linked to specific developmental and behavioral states (for a review, see Zayed and Robinson).[2]

Numerous genetic studies and mutational screens have been carried out to find genes that affect specific behaviors in *Drosophila* as well. For example, screens to find genes affecting circadian rhythm behavior have led to the discovery of the *clock* and *cycle* genes, which were later found to code for transcription factors, as well as genes like *period* and *timeless* that code for proteins that regulate the stability of *clock* and *cycle*.[3] Gene expression studies with *clock* mutants and wild-type flies uncovered 134 Clock-regulated genes that are involved in a variety of processes.[4] Another example in *Drosophila* is the *foraging* gene that encodes a cGMP-dependent protein kinase. Different natural alleles of *for* show different locomotion during larval feeding. Mutations in *dgcalpha1*, which encodes a soluble guanylyl cyclase subunit, increase both PKG activity and foraging locomotion. DNA microarray studies on this mutant in different genetic *for* backgrounds identified many genes that are differentially transcribed, and interestingly, relatively few are affected in both backgrounds. Several differentially expressed genes were found to exhibit enhanced or suppressed expression in a background-dependent manner.[5]

Gene expression analysis can also be a useful tool in the study of conditions that affect behavior such as alcoholism and drug addiction. Zebrafish that were chronically exposed to alcohol showed gene expression changes in more than 1900 genes, including a number of Cytochrome P450 family genes, ion channel genes, and solute carrier family members that provide insights on the effects of alcohol on physiology and how it might be affecting behavior.[6] For example, the solute carrier genes code for transmembrane proteins that play key roles in the transport of small molecules, including neurotransmitters across vesicular and plasma membranes, a class of molecules whose function was suspected to play roles in alcoholism but never conclusively proven before.[6]

An area of intense current scientific research is the effect of chromatin state on gene transcription with posttranslational modifications of histone proteins and DNA methylation underlying the epigenetic mechanisms that regulate gene expression. There are several examples of epigenetic alterations of behavior. In *Drosophila*, a mutation in *EHMT/G9a*, a gene that encodes a histone methyltransferase that methylates histone 3 at lysine 9 (H3K9), alters several behaviors including larval locomotor behavior, nonassociative learning, and courtship memory.[7] DNA microarray studies with the *EHMT/G9a* mutation revealed a number of genes that were differentially expressed.[7] In summary, gene expression changes may be induced by a variety of factors, and these changes may influence a broad range of behaviors that can be studied in several laboratory species.

The goal of this chapter is to provide conceptual and methodological information to researchers who are considering performing transcriptomics to investigate molecular mechanisms of brain function and/or behavior. While detailed step-by-step protocols are often useful, they can also be limited and dependent upon the exact organism and tissue being studied, as well as upon which exact experimental platform and equipment

is being used. There are numerous ways to obtain transcriptome data, and in many instances, a researcher will choose an experimental platform based on what equipment and/or services are available at local (or distant) genomic service centers. Therefore, rather than providing several different highly detailed protocols, this chapter primarily focuses on general design guidelines for gene expression studies with an emphasis on RNA isolation and cDNA production, as these are often the starting material that is given to genomic centers to generate experimental data. For the sake of simplicity, in this chapter, changes in gene expression specifically refer to changes in mRNA transcript levels.

METHODOLOGY

It is important to consider the many factors that can affect the quality and type of transcriptomic data in the design of an experiment. From choosing the appropriate source for the biologic samples, to generating high-quality RNA (and whether it is total or polyA RNA), to deciding which transcriptomic platform to employ and deciding upon an analysis pipeline. Some of the aforementioned factors have been addressed in detail elsewhere,[8,9] but many have not, so they will be reviewed here.

Experiment Design and RNA Preparation

General Experimental Design Considerations

There are many possible schemes or designs that can be employed in gene expression projects as well as factors to consider in their design. One of the most important factors is to decide how many replicates of each sample group are needed for the investigator to be able identify a sufficient number of statistically significant differentially expressed genes between groups. Results must be reproducible in any experimental system for the researcher to be confident in their validity, which in most cases means achieving statistical significance. Gene expression experiments are no exception, despite the early trend in the microarray field of using arbitrary fold-change cutoffs (e.g., twofold) to select differentially expressed genes. Although having three independent replicates is a minimum requirement for any statistical analysis, depending on the type of experiment and the nature of the biologic samples themselves, more than three replicates may be needed. If the biologic material is fairly homogeneous, for example, tissue culture cells grown in vitro at the same time or several insects of the same sex from a genetically identical population, and the experimental treatment or condition is fairly pronounced (e.g., exposure to a drug with strong physiological effects), then three replicates are likely to be sufficient. However, if the biologic samples are more heterogeneous (e.g., different individual animals that are not siblings, mixed populations of cells of different cell types, or individual cells), and the experimental treatment or condition is more subtle (e.g., age of an animal), then more than three replicates will likely be needed. Because gene expression experiments are fairly expensive, it is important to

consider the total number of replicates early, so if the total number of samples for a project is limited by budgetary constraints, then it is wiser to have fewer experimental groups with more replicates than to have more groups with only three replicates each. Often, one does not know at the start of a project how many replicates are needed until the first set of data has been analyzed. If time permits, pilot experiments with only two experimental groups (e.g., treatment vs. control) with three replicates can be informative. Similarly, if one is able to make additional biologic replicates of the sample groups (e.g., 4, 5, or 6 replicates) when the initial experiment is being done, then if it turns out that three replicates are not sufficient to give very many differentially expressed genes, the additional replicates can be processed and added to the data set. The advantages of adding biologic replicates during the initial experiment is that variables such as time of year or having a different population of cells or individuals is minimized, as is the time delay in completing the project. More information on the number of replicates needed for gene expression studies can be found in Conesa et al.[9]

Another important experimental design consideration is what biologic material one should select for RNA extraction. In the analysis of behavior, the obvious choice is the brain if the study organism is large enough to extract the tissue, as is the case for most vertebrates, from zebrafish to rat. In the case of some insects, one may use whole heads as the starting material. In mammals and other larger vertebrates, the whole brain is most likely too large of a tissue given the known idiosyncrasies of functions of particular brain areas. Therefore, analysis of the transcriptome of specific, well-defined brain regions may be required. That is, if the gene expression changes underlying the behavior are only occurring in one part of the brain, using whole-brain extract may make such changes undetectable in total brain RNA. A variety of approaches can be used to isolate RNA from specific very small areas of the brain and/or specific neurons in the brain. These include tissue dissection using laser capture microscopy; the homogenization of tissue and sorting of cells with cell-specific fluorescently (e.g. green fluorescent protein) labeled marker proteins followed by fluorescence-activated cell sorting; or the homogenization of tissue and immunoprecipitation of cells using antibodies to unique cell surface markers (immunopanning) (see Okaty et al. for review).[10] When RNA is obtained from a relatively small number of cells, the RNA (and/or resulting cDNA) often has to be amplified to have enough material to either perform microarray experiments or RNA sequencing. The fact that a smaller number of cells is used in the original sample and that the material is being amplified typically leads to more variation in the data from one sample to another. Therefore, more replicates of a sample type will likely be needed to obtain statistically significant gene expression changes.

Choice of Experimental Platform to Generate Transcriptomic Data

This is often the first choice that is made when deciding upon a study because it will usually affect all of the subsequent aspects of the study design and methodology. For

Table 12.1 Differences between the DNA microarrays and high-throughput sequencing in transcriptome experiments

	DNA microarrays	High-throughput sequencing
Species that experiment can be done	Limited to model organisms	Any (but analysis is easier with model organisms)
RNA type needed for experiment	Total or polyA	polyA (or rRNA depleted)
Time to process RNA samples prior to experiment	Moderate (for total RNA)	Relatively long
Time and complexity to prepare samples before hybridization or sequencing	Moderate	Typically long and complex (many steps)
Cost of preparing/labeling each sample	Moderate	Relatively high
Cost of array or sequencing	Moderate	Moderate to high
Time to obtain data	Moderate	Moderate to long
Time (and ease) of analyzing	Moderate	Moderate to long (and more difficult)
Ability to detect rare transcripts	Relatively low	Moderate to high

generating genome-wide transcriptomic data, there are two main technology platforms: DNA microarrays or high-throughput sequencing.

Once the cost of high-throughput sequencing became more affordable and the centers offering this type of sequencing services became more available, i.e., since about 2012, the vast majority of transcriptome profiling has been done using high-throughput DNA sequencing. While there are numerous advantages and disadvantages for each approach, one of the main disadvantages of high-throughput sequencing was the relatively high cost, but it has come down in recent years. A comparison of some of the key differences between the DNA microarray and RNA sequencing platforms is shown in Table 12.1.

Before discussing the two main experimental platforms for obtaining the transcriptome of a given sample, both approaches require obtaining high-quality RNA, a topic that will be addressed here first.

RNA Type and Isolation

A very important factor contributing to a successful transcriptome analysis experiment is to begin with high-quality RNA. For DNA microarray experiments, either total RNA or mRNA may be used, but for high-throughput sequencing, either mRNA- or rRNA-depleted RNA must be used; otherwise a majority of the sequence reads will be rRNA.

Total RNA

Total RNA may be isolated using a variety of methods, but those most commonly employed rely on homogenization of the cells, tissue, or whole organisms in the

presence of a highly denaturing guanidine isothiocyanate (GITC)-containing buffer, which immediately inactivates RNases and ensures isolation of intact RNA. Such buffers can be combined with phenol:chloroform for a single-solution approach to the isolation and purification of total RNA as originally described by Chomczynski and Sacchi.[11] Several commercially available reagents such as Invitrogen/ThermoFisher TRIzol are examples of this approach. Other commercial available RNA isolation kits combine GITC-containing buffers with purification columns or membranes, such as Zymo Direct-zol, Qiagen's RNeasy, and Ambion/ThermoFisher Trizol Plus purification kits. Our experience has led us to rely on TRIzol to extract RNA from most samples. Whereas different purification columns are required depending on the mass or type of sample tissue, the TRIzol method is robust and scalable, and samples may also be homogenized directly in the reagent, which eliminates the need to use additional RNase-protective buffers that may interfere with the extraction. In some instances, the use of total RNA may lead to poor or no reverse transcription of the template RNA and/or to poor incorporation of modified nucleotides and/or fluorescent background on the hybridized array due to contaminants present in the RNA preparation. In such cases, an additional cleanup of the total RNA using column-based purification kits can be helpful (e.g., Zymo RNA clean and concentrator or Ambion/ThermoFisher Megaclear).

There are instances where the collection of tissue may be time consuming (e.g., the dissection of specific parts of the brain) and care must be taken to reduce RNA degradation prior to the addition of GITC-containing buffer. In these cases, it is advisable to place the tissue in RNAlater (Ambion/ThermoFisher) for either a short term (4°C) or long term (−20°C) prior to RNA isolation.

mRNA- and rRNA-Subtracted RNA

As mentioned, total RNA is generally not suitable for high-throughput sequencing, so one must either purify mRNA or mRNA plus long noncoding (nc) RNA prior to the creation of DNA libraries for sequencing.

mRNA Isolation There are a few different ways to purify mRNA (i.e., polyA RNA), but the ones that employ oligo dT linked to paramagnetic beads are among the most efficient and popular. In all cases, total RNA is purified first (see earlier), and it is then mixed with the beads to bind to the polyA RNA. Two commercially available beads are Magnetic mRNA isolation beads (NEB catalog number (cat. no.). S1550S) and the Dynabeads mRNA purification kit (Invitrogen/ThermoFisher cat. no. 61006). To use these beads, one must have magnetic stands that can hold 1.5-mL microfuge tubes (e.g., Promega/ThermoFisher cat. no. PR-Z5342. NEB cat. no. S1509S, or Qiagen cat. no. 36912) or in the case of processing a large number of samples, 96-well magnetic plates (e.g., Alpaqua cat. no. 96S Super Magnet Plate).

rRNA Subtracted RNA In cases where it also is desirable to have long ncRNA data in addition to mRNA data, then one can prepare RNA samples that are depleted of ribosomal RNA. Ribosomal RNA is removed using beads that have single-stranded oligonucleotide sequences that can hybridize to rRNA, so the RNA that does not bind to the beads is mostly mRNA and large ncRNA. The preferred beads for preparing rRNA-depleted RNA are the Ribo-Zero rRNA Removal Kit (Illumina cat. no. MRZH116) or the Ribo-Zero Gold kit (Illumina cat. no. MRZG126), which also removes mitochondrial RNA. While these beads have antisense oligonucleotides specific to mammalian rRNA sequences, they also work quite well for removing most animal rRNAs.

RNA Quantification and Quality Control

It is essential that the RNA be quantified and checked for overall integrity (i.e., absence of degradation). RNA can be quantified and checked for protein contamination using a UV spectrophotometer (e.g., ThermoFisher Nanodrop). For a more accurate quantification, especially when there is only a very small amount of RNA, fluorometric dyes can be used (i.e., Quant-iT RiboGreen dye, Molecular Probes/ThermoFisher cat. nos. R11490 or R11491) in conjunction with a fluorometric plate reader or a single-sample fluorometer (e.g., ThermoFisher Quant-iT).

The integrity of total RNA samples can be assayed using devices such as the Agilent Bioanalyzer or TapeStation. These are microfluidic capillary gel electrophoresis instruments that can quickly separate RNA (or DNA or protein) samples, and the rRNA peaks can be observed. Degraded RNA samples have rRNA peaks that are not of the expected amplitude and/or show spreading. RNA integrity can also be examined using nondenaturing RNA gels, but this method requires higher amounts of RNA.

More details on RNA quantification and quality control can be found in Appendix A.

RNA Storage and Shipping

RNA samples are typically snap frozen in liquid nitrogen and stored in a −80°C freezer until they are processed further. When using the services of genomic centers, samples are usually shipped on dry ice. Sometimes, this is not practical, especially when the samples have to be shipped overseas. In those cases, products like RNAstable (Biomatrica cat. no. 93221-001) allow RNA to be lyophilized onto a stable matrix, shipped at ambient temperature, and then to be reconstituted with nuclease-free water to be processed further.

After one has obtained high-quality RNA, the processing of those RNAs will largely depend upon which experimental platform one will be using to obtain the transcriptome data. Therefore, the remaining sections are divided into the two platforms addressed in this chapter: DNA microarrays and high-throughput sequencing.

DNA Microarrays

With the advent of sequencing the entire genomes of humans and other model organisms, a variety of high-throughput approaches were developed in the late 1990s and early 21st century to study entire transcriptomes. One of these is DNA microarray technology. Early versions of microarray production involved the spotting of dsDNA fragments or ssDNA oligonucleotides onto specially coated glass microscope slides, and these were often produced by academic labs and genomic centers using robotic printers.[12,13] Several commercial microarrays were developed, including but not limited to Affymetrix (GeneChips), Illumina (GEX bead arrays), Agilent arrays, and Nimblegen/Roche arrays. The Nimblegen platform no longer exists, but the other three do. Each of the still-existing platforms attach or synthesize thousands of single-stranded DNA oligonucleotides with different sequences to a substrate. Briefly, Affymetrix arrays synthesize 20 mer oligonucleotides using photoactivatable chemistry on a transparent quartz surface, Illumina uses more traditional synthesis of 50–70 mers onto silica beads, and Agilent synthesizes 60 mers at specific locations on a coated microscope slide using inkjet technology (see Bumgarner[14] for more information). Each of these commercial array types have their advantages and disadvantages that can vary depending on the specific application for which the arrays are used (e.g., single nucleotide polymorphism analysis vs. transcriptome analysis). Both the Affymetrix and Agilent have predesigned arrays for many model organisms whose genomes have been sequenced, while Illumina Bead transcriptome arrays are mainly limited to human, mouse, and rat. One distinct advantage of the Agilent arrays is that it is possible and economical to make custom arrays.

cDNA Synthesis

Once high-quality RNA is obtained, one typically makes fluorescently labeled, double-stranded cDNA that can be hybridized to the microarray and quantified. Some manufacturers' experimental workflow (e.g., Affymetrix, Agilent) makes labeled, amplified RNA instead of cDNA. However, the production of cDNA does provides added advantages, including that the final labeled products are likely to be more stable, unused cDNA can be labeled again if the initial labeling reaction failed or hybridizations did not work, and unused cDNAs can be used for verification experiments such as qRT PCR.

A detailed protocol for cDNA preparation is provided in Appendix A.

Labeling cDNAs

cDNAs can be fluorescently labeled during cDNA synthesis using direct (e.g., incorporation of Cy dye-labeled nucleotides[15]) or indirect protocols (e.g., incorporation of amino allyl dUTP followed by conjugation to Alexa fluor reactive dyes).[16] Alternatively, cDNAs can be labeled afterward through the use of Cy dye-labeled random nonamer primers and the Klenow fragment of DNA polymerase.[17] This latter approach has the

added advantages of no inhibition or "dye bias" in cDNA synthesis, cDNAs becoming amplified during the labeling protocol by DNA polymerase, and that the initial unused cDNAs are unlabeled and therefore can be used again for other experiments if needed.

We have used this latter direct labeling of cDNAs with Cy dye-labeled nonamers successfully for Nimblegen array-based transcriptome analyses[6,18–21] as well as RNA immunoprecipitation analyses.[22–24]

A detailed protocol for labeling ds cDNA with Cy dye-labeled random nonamers is provided in Appendix B.

Array Hybridization and Scanning

Each commercial array manufacturer typically has specialized equipment and protocols for carrying out the hybridization of the labeled samples to each microarray. Most protocols require the hybridization of a single labeled cDNA (or aRNA) to the array, with or without labeled control RNA samples. Once the array has been hybridized, washed, scanned, and the signals quantified, replicate samples of sample type can be used to compare to replicate samples of a different sample type (e.g., mutant vs. control). We have found that for long oligonucleotide-based arrays (e.g., Nimblegen and Agilent), samples can be labeled with either Cy3- or Cy5-labeled nonamers and two samples with different Cy labels mixed and cohybridized to the same array at the same time. During array scanning, a quantification of each Cy dye signal can be obtained. Thus, one can use half the number of arrays to perform an experiment with no noticeable change in the outcome of the experimental data.

While there have been a few different microarray scanner manufacturers over the years, since 2005 the scanners that have been primarily made are Affymetrix/ThermoFisher GeneChip scanners, Illumina iScan bead array scanners, Agilent SureScan scanners, and Axon/Molecular Devices GenePix scanners. The latter two can be used to scan almost any microarray that is of $1 \times 3''$ standard microscope slide size, while the GeneChip and bead array scanners are specific for a specified type of array. Each array manufacturer provides detailed protocols on how their arrays should be scanned.

Scanned arrays typically generate large image (e.g., TIFF) files from which the fluorescent signals from specific locations on the array can be quantified using specialized software. Quantification software is made by each scanner manufacturer (e.g., GenePix Pro for GenepPix scanners) and/or array manufacturer (Nimblescan for Nimblegen arrays), but free quantification software (e.g., TIGR Spotfinder) is also available.

Analysis of Microarray Data

The quantified image data are exported in a tabular format from one of the aforementioned software programs and are ready to be analyzed. There are many things that need to be considered during analysis: background subtraction, filtering of low signals, flagging of aberrant features, and data normalization. These topics are discussed in detail

by Neal and Westwood.[8] A number of software packages have the ability to perform many of the aforementioned processes automatically using "default" parameters, but it is always advisable to inspect image files prior to quantification and further analysis, in case further steps need to be taken during analysis.

There are also several software packages that enable one to perform postscanning normalization as well as statistical analyses and the clustering of genes with similar expression patterns (i.e., visualization tools). These include commercial software such as GeneSpring (Agilent), Spotfire (TIBCO), Acuity (Axon/Molecular Devices), and ArrayStar (DNASTAR), as well as open-source software such as TIGR MIDAS and MeV (www.tm4.org), BASE (http://base.thep.lu.se/), and Bioconductor and R (www.bioconductor.org). Further information on the microarray analysis, and in particular statistical analysis of microarray data, can be found in Neal and Westwood.[8]

Additional information on software that can aid in the visualization of microarray data as well as help interpret the functional meaning of differentially expressed genes (e.g., gene ontology [GO], pathway and network analysis) can be found in Visualization and Pathway Tools section.

High-Throughput Sequencing

High-throughput sequencing, also known as massively parallel sequencing and next-generation sequencing, as the name implies, sequences tens of thousands to billions of DNA molecules at the same time. The first commercially available high-throughput sequencers appeared in about 2005, and they were made by 454 (later Roche), based on a pyrosequencing technology, and Solexa (now Illumina) based on sequencing by synthesis technology. There are a few other companies that also make high-throughput sequencers, but the majority of these were never produced in high numbers because they were quite expensive, and they are mainly used for their unique characteristics (e.g., long sequence read by Pacific Biosciences sequencers that facilitate assembly of entire genomes). More recently, companies such as Ion Torrent (ThermoFisher) and Nanopore make sequencers that are increasingly popular because they are relatively inexpensive to buy and operate. However, the vast majority of high-throughput DNA sequencing that is currently being done worldwide is performed on Illumina sequencers, so the subtopics in this section are focused on data generated by these sequencers.

In previous sections, we discussed RNA preparation. The guidelines provided in those sections apply for preparing samples for high-throughput sequencing as well. For Illumina transcriptome sequencing, mRNA (or rRNA-depleted RNA) is converted to cDNA libraries of a specific nucleotide size range that will be sequenced. For many researchers, the DNA library preparation (see later) and sequencing will be done by a genomics center, so all they will need to do is to prepare high-quality RNA in sufficient quantity and deliver it to the center.

DNA Library Preparation

Most DNA libraries for Illumina sequencing are prepared from kits that provide almost all of the necessary reagents to make the libraries. Kits are made by a number of manufacturers, including Illumina, New England Biolabs, and Kapa (Roche). For those researchers who will be making the libraries themselves, one can either convert their mRNA to cDNA first (e.g., using a modified version of the protocol in Appendix A), which can be used to make the library, or use a kit where the starting material is mRNA. In the past few years, most of the kits that make libraries from mRNA are "stranded" kits that preserve the identity of the sense strand. Examples of such kits are the lllumina TruSeq stranded mRNA kit, the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina kit, and KAPA (Roche) Stranded mRNA-Seq kit. KAPA also makes HyperPlus kits for making libraries from DNA (or cDNA) in a much shorter period of time than traditional kits.

If a researcher is planning on making libraries using a kit, one item that is generally not provided by a kit is the purification columns or paramagnetic beads required for binding DNA products at various steps of the library preparation. Most kit manufacturers recommend using AMPure beads (Agencort, Beckman Coulter) for DNA cleanup. Significant cost savings can be achieved using Sera-Mag Magnetic Speed-beads (SPRI beads) (GE Life Sciences cat. no. 24152105050250) and by following the protocols outlined by Rohland and Reich[25] and DeAngelis and coworkers.[26]

For researchers who are planning to make large numbers of libraries, significant cost savings can also be achieved using protocols that employ individually purchased enzymes and reagents as opposed to kits (e.g., Wang et al.[25,27]).

RNAseq Analysis

The analysis of RNAseq data is an extensive topic on its own, and only a brief overview will be provided here. A detailed review of many of the current pipelines and publicly available software for RNAseq data analysis can be found in Conesa et al.[9] Some online resources for information and/or tutorials on publicly available software, as well as sources for where the software can be run, include the Galaxy Project (https://galaxyproject.org/tutorials/rb_rnaseq/),[28] Bioconductor RNAseq workflow (https://www.bioconductor.org/help/workflows/rnaseqGene/),[29] and GenePattern (http://software.broadinstitute.org/cancer/software/genepattern/rna-seq-analysis).[30]

There are also several commercial software packages that generally rely on publicly available programs for most stages of the analysis but provide an easier to use interface and superior visualization tools. Several also have tools for things like GO term, pathway, and network analysis to get more insight when one obtains lists of differentially expressed genes. Most of these software packages are subscription based, and they can be expensive. Some of these include A.I.R. (https://transcriptomics.cloud/), ArrayStar (https://www.dnastar.com/t-sub-products-genomics-arraystar.aspx), Avadis (http://www.strand-ngs.

com/), BaseSpace apps for RNAseq analysis (Illumina, https://www.illumina.com/informatics/sequencing-data-analysis/rna.html), CLC Workbench (https://www.qiagenbioinformatics.com/solutions/gene-expression/), and Partek Genomics Suite and Pathway (http://www.partek.com/).

Quality Control Checkpoints

Poor RNA sample quality or library preparation can lead to the presence of a large number of adaptor sequences, overrepresented k-mers, duplicated reads, PCR artifacts, and other sequencing errors in the data. Samples that have a large number of these errors and artifacts should be excluded from further analysis. Programs such as FastQC (Andrews, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can perform analysis of the raw sequence reads.

In addition to quality control of raw reads, which is typically done, other quality control analysis can also be performed after read alignment and quantification of transcript numbers with the goal to again filter out samples that have large numbers of anomalies.[9] Once individual samples pass quality control, they can be compared to each other for overall sample reproducibility. Reproducibility among replicates should be generally high (e.g., Spearman $R^2 > 0.9$), but the cutoff for acceptable reproduction will largely depend upon the heterogeneity of the initial biologic samples themselves. As mentioned in the section on Experimental Design, samples with poorer reproducibility will require more replicates to obtain statistically significant differences in gene expression.

Transcript Identification

The next step in the analysis pipeline is to align the raw sequence reads. This is greatly facilitated when there is a reference genome for the organism being studied. Two commonly used programs for aligning transcript sequences are TopHat[31,32] and STAR.[33]

Transcript Quantification

Once the sequence reads have been aligned, they then need to be quantified to estimate gene and transcript expression. RPKM (reads per kilobase of exon model per million reads), FPKM (fragments per kilobase of exon model per million mapped reads), and TPM (transcripts per million) are the most commonly reported formats for reporting RNAseq gene expression. One of the most highly used programs for quantifying transcripts is Cufflinks,[34] which estimates transcript expression from mappers such as TopHat using an expectation-maximization approach. This approach takes into account biases such as the nonuniform read distribution along a gene.

Differential Gene Expression Analysis

Differential gene expression (DGE) analysis requires that gene expression values be compared between sample group types. RPKM and FPKM normalize the most

important factor for comparing samples-sequencing depth. Normalizing approaches that are based simply on total or effective counts tend to perform poorly when samples have highly and/or differentially expressed genes, but programs like TMM[35] and DESeq[36] can minimize the effects of such "abnormally" expressed genes. Different programs that perform DGE analysis use varied approaches for normalizing data, so they perform differently depending on the type of experimental data and number of sample replicates. LimmaVoom[37] performs well under many circumstances.[38–40] Two of the most commonly used programs, DESeq and edgeR, perform similarly in ranking differentially expressed genes, but DESeq has been shown to have relatively conservative false discovery rates (FDRs), whereas edgeR FDRs are relatively liberal.[38,41]

Visualization and Pathway Tools

Visualization tools can be broken into different categories with two of the main ones being viewing normalized transcript abundance along annotated genomes and viewing DGE patterns in a graphic format (e.g., heat maps). Transcript abundance can be viewed using genome browsers such as the UCSC browser,[42] integrative genomics viewer,[43] and RNAseqViewer.[44] Graphic representations of DGEs can be achieved with programs like MeV and QuickRNAseq.[45] As previously mentioned, many of the commercially available software packages have excellent tools for visualizing DGEs.

Once one has obtained lists of differentially expressed genes, the next step usually involves visualizing the results in way that is easier to understand (e.g., heat maps), as well as trying to interpret the biologic implications of the DGEs through GO function, biochemical or molecular pathway, and/or network analysis. Again, there are a variety of publicly available software applications for doing this. This is usually accomplished by comparing a specific list(s) of DGEs, based on statistical criteria, to the rest of the genome for functions that are overrepresented. Alternatively, the entire list of differentially expressed genes can be used, and a gene set enrichment analysis may be performed. Many of these tools were originally developed for microarray and other large genomic (e.g., proteomic) datasets. Tools that specifically take into account some of the biases that are inherent to RNAseq data include GOseq,[46] Gene Set Variation Analysis (GSVA),[47] and SeqGSEA.[48]

SUMMARY

In this chapter, some key aspects of gene expression experimental design and RNA preparation were addressed. Comparisons outlining some of the differences between obtaining transcriptomic data using DNA microarrays versus high-throughput sequencing were made, and workflows for both of these technology platforms were presented. Since most current gene expression experiments involve obtaining data using high-throughput sequencing, often the researcher needs only to prepare the RNA samples and provide them to a genomic service center. Therefore, the success of a gene expression study

will largely depend upon the original design of the experiment, including the source of tissue or even specific cells from which the RNA was obtained and having enough replicate samples to reach statistically significant results. Other factors that will be key to a successful study include isolation of high-quality RNA, having a robust and appropriate analysis pipeline, and the use of software tools to identify the underlying biologic and molecular processes and pathways that are being revealed by the changes in gene expression being observed.

ACKNOWLEDGMENTS

I would like to thank past and present members of the Canadian *Drosophila* Microarray Centre (CDMC) and Westwood Lab including Anna Soltyk, Scott Neal, Sarah Gonsalves, Kaigo Mo and Zak Razak for their contributions to various gene expression studies. I would also like to thank Hang Noh Lee and Brian Oliver (NIDDK, NIH, Bethesda, MD, USA) for their insights and advice on RNAseq. The gene expression studies performed in or in collaboration with the Westwood Lab was supported by an NSERC Discovery grant to J.T.W. The CDMC was originally supported by a multiuser maintenance and NET grant from CIHR and an equipment and facility grant from NSERC.

REFERENCES

1. Robinson GE, Fernald RD, Clayton DF. Genes and social behavior. *Science*. 2008;322(5903):896–900. https://doi.org/10.1126/science.1159277.
2. Zayed A, Robinson GE. Understanding the relationship between brain gene expression and social behavior: lessons from the honey bee. *Annu Rev Genet*. 2012;46:591–615. https://doi.org/10.1146/annurev-genet-110711-155517.
3. Peschel N, Helfrich-Förster C. Setting the clock–by nature: circadian rhythm in the fruit fly *Drosophila melanogaster*. *FEBS Lett*. 2011;585(10):1435–1442. https://doi.org/10.1016/j.febslet.2011.02.028.
4. McDonald MJ, Rosbash M. Microarray analysis and organization of circadian gene expression in *Drosophila*. *Cell*. 2001;107(5):567–578.
5. Riedl CAL, Neal SJ, Robichon A, Westwood JT, Sokolowski MB. *Drosophila* soluble guanylyl cyclase mutants exhibit increased foraging locomotion: behavioral and genomic investigations. *Behav Genet*. 2005;35(3):231–244. https://doi.org/10.1007/s10519-005-3216-1.
6. Pan Y, Kaiguo M, Razak Z, Westwood JT, Gerlai R. Chronic alcohol exposure induced gene expression changes in the zebrafish brain. *Behav Brain Res*. 2011;216(1):66–76. https://doi.org/10.1016/j.bbr.2010.07.017.
7. Kramer JM, Kochinke K, Oortveld MAW, et al. Epigenetic regulation of learning and memory by *Drosophila* EHMT/G9a. *PLoS Biol*. 2011;9(1):e1000569. https://doi.org/10.1371/journal.pbio.1000569.
8. Neal SJ, Westwood JT. Optimizing experiment and analysis parameters for spotted microarrays. *Methods Enzymol*. 2006;410:203–221. https://doi.org/10.1016/S0076-6879(06)10010-5.
9. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17(1):13. https://doi.org/10.1186/s13059-016-0881-8.
10. Okaty BW, Sugino K, Nelson SB. Cell type-specific transcriptomics in the brain. *J Neurosci*. 2011;31(19):6939–6943. https://doi.org/10.1523/JNEUROSCI.0626-11.2011.
11. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*. 1987;162(1):156–159. https://doi.org/10.1016/0003-2697(87)90021-2.
12. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467–470.

13. DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*. 1996;14(4):457–460. https://doi.org/10.1038/ng1296-457.
14. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol*. 2013. https://doi.org/10.1002/0471142727.mb2201s101 [Chapter 22: Units 22.1–22.1.11].
15. Neal SJ, Gibson ML, So AK-C, Westwood JT. Construction of a cDNA-based microarray for *Drosophila melanogaster*: a comparison of gene transcription profiles from SL2 and Kc167 cells. *Genome*. 2003;46(5):879–892. https://doi.org/10.1139/g03-056.
16. Semotok JL, Westwood JT, Goldman AL, Cooperstock RL, Lipshitz HD. Measuring mRNA stability during early *Drosophila* embryogenesis. *Methods Enzymol*. 2008;448:299–334. https://doi.org/10.1016/S0076-6879(08)02616-5.
17. Ouellet M, Adams PD, Keasling JD, Mukhopadhyay A. A rapid and inexpensive labeling method for microarray gene expression analysis. *BMC Biotechnol*. 2009;9(1):97. https://doi.org/10.1186/1472-6750-9-97.
18. Gonsalves SE, Moses AM, Razak Z, Robert F, Westwood JT. Whole-genome analysis reveals that active heat shock factor binding sites are mostly associated with non-heat shock genes in *Drosophila melanogaster*. *PLoS One*. 2011;6(1):e15934.
19. Le Masson F, Razak Z, Kaigo M, et al. Identification of heat shock factor 1 molecular and cellular targets during embryonic and adult female meiosis. *Mol Cell Biol*. 2011;31(16):3410–3423. https://doi.org/10.1128/MCB.05237-11.
20. Zhang J, Marshall KE, Westwood JT, Clark MS, Sinclair BJ. Divergent transcriptomic responses to repeated and single cold exposures in *Drosophila melanogaster*. *J Exp Biol*. 2011;214(Pt 23):4021–4029. https://doi.org/10.1242/jeb.059535.
21. Siddiqui NU, Li X, Luo H, et al. Genome-wide analysis of the maternal-to-zygotic transition in *Drosophila* primordial germ cells. *Genome Biol*. 2012;13(2):R11. https://doi.org/10.1186/gb-2012-13-2-r11.
22. Laver JD, Li X, Ancevicius K, et al. Genome-wide analysis of Staufen-associated mRNAs identifies secondary structures that confer target specificity. *Nucleic Acids Res*. 2013;41(20):9438–9460. https://doi.org/10.1093/nar/gkt702.
23. Laver JD, Li X, Ray D, et al. Brain tumor is a sequence-specific RNA-binding protein that directs maternal mRNA clearance during the *Drosophila* maternal-to-zygotic transition. *Genome Biol*. 2015;16(1). https://doi.org/10.1186/s13059-015-0659-4.
24. Chen L, Dumelie JG, Li X, et al. Global regulation of mRNA translation and stability in the early *Drosophila* embryo by the Smaug RNA-binding protein. *Genome Biol*. 2014;15(1). https://doi.org/10.1186/gb-2014-15-1-r4.
25. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res*. 2012;22(5):939–946. https://doi.org/10.1101/gr.128124.111.
26. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res*. 1995;23(22):4742–4743.
27. Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP. A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. Hudson ME, ed. *PLoS One*. 2011;6(10):e26426. https://doi.org/10.1371/journal.pone.0026426.
28. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016;44(W1):W3–W10. https://doi.org/10.1093/nar/gkw343.
29. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000 Res*. 2016;4:1070. https://doi.org/10.12688/f1000research.7035.2.
30. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Genet*. 2006;38(5):500–501. https://doi.org/10.1038/ng0506-500.
31. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–1111.
32. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36.

33. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
34. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–515.
35. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.
36. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106.
37. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29.
38. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013;14(9):R95. https://doi.org/10.1186/gb-2013-14-9-r95.
39. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform*. 2013;14:91. https://doi.org/10.1186/1471-2105-14-91.
40. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*. 2015;16(1):59–70. https://doi.org/10.1093/bib/bbt086.
41. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. *BMC Genom*. 2012;13:484. https://doi.org/10.1186/1471-2164-13-484.
42. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996–1006. https://doi.org/10.1101/gr.229102.
43. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–192. https://doi.org/10.1093/bib/bbs017.
44. Roge X, Zhang X. RNAseqViewer: visualization tool for RNA-seq data. *Bioinformatics*. 2014;30(6):891–892. https://doi.org/10.1093/bioinformatics/btt649.
45. Zhao S, Xi L, Quan J, et al. Quick RNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. *BMC Genom*. 2016;17:39. https://doi.org/10.1186/s12864-015-2356-9.
46. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14. https://doi.org/10.1186/gb-2010-11-2-r14.
47. Hanzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform*. 2013;14:7. https://doi.org/10.1186/1471-2105-14-7.
48. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinform*. 2013;14(suppl 5):S16. https://doi.org/10.1186/1471-2105-14-S5-S16.

APPENDIX A: PREPARATION OF cDNAs

Adapted from Roche Nimblegen Gene Expression Analysis Guide (version 5.1) and Ouellet and coworkers.[17] Note that the major modification is to make half-size reactions throughout the protocols.

Materials

- Anchored oligo dT Primer (500 ng/μL, Ambion/ThermoFisher cat. no. AB1247)
- Random Hexamers (100 μM (0.2 μg/μL), Invitrogen/ThermoFisher cat. no. SO142)
- RNase A Solution (4 mg/mL Promega cat. no. A7973)
- SuperScript II Double-Stranded cDNA Synthesis Kit (Invitrogen/ThermoFisher cat. no. 11917-010)

Step 1. Spectrophotometric Quality Control of RNA

Prior to synthesizing cDNA, verify that the RNA samples are of sufficient purity, typically with a Nanodrop spectrophotometer (ThermoFisher).

1. Quantitate each RNA sample according to the following formula:
   RNA concentration (μg/mL) = A260 × 40 × dilution factor.
   RNA samples must have a concentration ≥ 1.0 μg/μL.
2. Verify all samples meet the following requirements: A260/A280 ≥ 1.8; A260/A230 ≥ 1.8.

Step 2. Bioanalyzer/Gel QC of RNA

Verify RNA samples are of sufficient molecular weight. This can be done using an Agilent 2100 Bioanalyzer (preferred especially for smaller samples) or using a denaturing agarose gel.

1. Transfer 250 ng total RNA or 250 ng polyA + RNA to a sterile microcentrifuge tube. Store the remainder of your sample on ice or at −80°C.
2. Analyze samples using the Agilent Bioanalyzer and RNA 6000 Nano Kit.
3. Compare the Bioanalyzer traces to published electropherogram images in the Agilent Bioanalyzer manual or to samples known to be intact. Degraded samples appear as significantly lower intensity traces with the main peak area shifted to the left and typically exhibit much more noise in the trace.

   Samples exhibiting degradation should not be used for cDNA synthesis because there is a high risk for obtaining poor results.

Step 3. First Strand cDNA Synthesis

Use the Invitrogen (ThermoFisher) SuperScript II Double-Stranded cDNA Synthesis Kit to synthesize double-stranded cDNA.

1. Thaw and maintain the following components on ice. Combine components in a 0.2-mL tube on ice according to the following tables.

   Note: prepare a primer mix solution by mixing one part random hexamers to three parts anchored oligo dT primer.

	Total RNA	PolyA + RNA (mRNA) amount
RNA amount	5 μg	0.5 μg
Oligo dT/random primer mix	1 μL	1 μL
DEPC water	To volume	To volume
Total	5.5 μL	5.5 μL

2. Heat sample(s) to 70°C for 10 min in a thermocycler. Briefly spin tubes in a microcentrifuge and place them on ice for 5 min.

3. Add the following to each sample tube. (You can use a master mix when preparing multiple samples.)

Component	Volume (μL)
Reaction volume from step 2	5.5
5X first strand buffer	2
0.1 M DTT	1
5 mM dNTP Mix	1
Total	9.5

4. Mix gently (avoid vortexing). Briefly spin the tube(s) in a microcentrifuge.
5. Place sample(s) in a thermocycler set at +42°C for 2 min.
6. Add 1 μL of SuperScript II RT and mix gently (avoid vortexing).
7. Incubate the sample(s) at +42°C for 60 min.
8. STOP POINT: Briefly spin the tube(s) in a microcentrifuge. Place the sample(s) on ice until the second strand synthesis. You can store the sample(s) overnight at −20°C.

Step 4. Second Strand cDNA Synthesis

1. Add the following components to the first strand reaction(s) in the indicated order; you can use a master mix. Keep tube(s) on ice or in a PCR tube chiller rack.

Component	Volume (μL)
Reaction volume from Step 3.8	10.5
DEPC water	45
5X second strand buffer	15
10 mM dNTP mix	1.5
10 U/μL DNA ligase	0.5
10 U/μL DNA polymerase I	2
2 U/μL RNase H	0.5
Total	75

2. Mix gently (avoid vortexing). Briefly spin the tube(s) in a microcentrifuge. Incubate at +16°C for 2 h.
3. Add 1 μL of 5 U/μL T4 DNA polymerase to each reaction. Incubate at +16°C for an additional 5 min. Do not allow the reaction temperature to exceed +16°C during this step.
4. STOP POINT: Place the sample(s) on ice or in a PCR tube chiller rack, and add 5 μL of 0.5 M EDTA. You can store samples overnight at −15 to −25°C.

Step 5. RNase A Cleanup

1. Add 1 μL of 2 mg/mL RNase A solution to the tubes from Step 4.4. Use caution when working with RNase A. Change gloves after use. Use RNaseZap (Thermo-Fisher) to clean work area surfaces.

2. Mix gently (avoid vortexing). Briefly spin the tube(s) in a microcentrifuge.
3. Incubate sample(s) at +37°C for 10 min.
4. During incubation, centrifuge Phase Lock tube(s) (ThermoFisher) at 12,000x g for 2 min. Label one Phase Lock tube and two 1.5-mL centrifuge tubes for each sample with sample names.
5. Add 82 μL of phenol:chloroform:isoamyl alcohol (25:24:1) to one set of 1.5-mL centrifuge tubes.
6. Transfer the sample(s) to the tube(s) containing phenol:chloroform:isoamyl alcohol. Vortex well.
7. Transfer samples with the phenol:chloroform:isoamyl alcohol to Phase Lock tubes.
8. Centrifuge at 12,000x g for 5 min.
9. Transfer the upper, aqueous layer to a clean, labeled 1.5-mL tube.

Step 6. cDNA Precipitation

1. Add 8 μL (0.1 volume of Step 5.9) of 7.5 M ammonium acetate to the samples. Mix by repeated inversion. Briefly spin the tube(s) in a microcentrifuge.
2. Add 3.5 μL of 5 mg/mL glycogen to the samples. Mix by repeated inversion. Briefly spin the tube(s) in a microcentrifuge.
3. Add 163 μL (2 volumes of Step 5.9) of ice-cold absolute ethanol to the samples. Mix by repeated inversion.
4. Centrifuge at 12,000x g for 20 min.
5. Remove supernatant. Take care not to disturb the pellet.
6. Add 250 μL of ice-cold 80% ethanol (v/v). Mix by repeated inversion.
7. Centrifuge tubes at 12,000x g for 5 min.
8. Remove supernatant. Take care not to disturb the pellet.
9. Repeat steps 6–8.
10. Dry the pellet in a DNA vacuum concentrator.
11. Rehydrate samples with 10 μL of sterile, nuclease-free water.

Step 7. Spectrophotometric QC of cDNA

1. Quantify each cDNA sample according to the following formula:
   cDNA concentration (μg/mL) = A260 × 50 × dilution factor.
2. Verify that all samples meet the following requirements: concentration ≥100 ng/μL, A260/A280 ≥ 1.8, A260/A230 ≥ 1.8.

Step 8. Bioanalyzer/Gel QC of cDNA

1. Transfer 250 ng cDNA to a sterile microcentrifuge tube. Store the remainder of your sample on ice or at −15 to −25°C.
2. Analyze the samples using the Agilent Bioanalyzer and RNA 6000 Nano Kit.

3. Compare the Bioanalyzer traces to the traces displayed next. Verify that all samples meet the following requirement for acceptance:
   Median size ≥400 bp when compared to a DNA ladder.
   It looks similar to the examples of good cDNA sample traces displayed next.
   If using an agarose gel, compare the gel images to the Bioanalyzer's electropherogram images.

APPENDIX B: cDNA LABELING

Adapted from Nimblegen Gene Expression Analysis Guide (version 2). Note that the main modification is to perform half-size reactions.

Materials

- 5′ Cy3- and Cy5-labeled Random Nonamers (9 mer "Wobble") TriLink Biotechnologies
  (8 O.D. or 50 O.D. sizes)
- Klenow Fragment (3′→5′ exo-) 50 U/μL (NEB cat. no. M0212M)

1. Prepare random 9 mer buffer

Nuclease-free water	8.6 mL
1M tris HCl, pH 7.4	1.25 mL
1M MgCl2	125 μL
β-Mercaptoethanol	17.5 μL
Total	10 mL

2. Dilute Cy3- and Cy5-labeled 9 mers to 1 O.D./42 μL random 9 mer buffer. Aliquot to 20 μL individual reaction volumes in 0.2-mL thin-walled PCR tubes and store at −20°C; protected from light.
3. Assemble the following components in separate 0.2-mL thin-walled PCR tubes:

	Amount
cDNA	0.5 μg
Cy3 or Cy5-labeled 9 mers	20 μL
Nuclease-free water	To volume
Total	40 μL

4. Heat denature samples in a thermocycler at 98°C for 10 min. Quick-chill in an ice-water bath for 10 min.
   Important: quick-chilling after denaturation is critical for high-efficiency labeling.

5. Prepare 10 μL of the following dNTP/Klenow master mix for each sample prepared in Step 4:

   Important: keep all reagents and dNTP/Klenow master mix on ice. Do not vortex after addition of Klenow.

5X dNTP/Klenow master mix	Recipe per sample (μL)
10 mM dNTP mix	5
Nuclease-free water	4
Klenow (50 U/μL)	1
Total	10

6. Add 10 μL of the dNTP/Klenow master mix to each of the denatured samples.

	Volume (μL)
Reaction volume from step 4	40
5X dNTP/Klenow master mix	10
Total	50

7. Mix well by pipetting up and down 10 times. (Important: do not vortex after addition of Klenow.)
8. Quick-spin to force contents to bottom of the tube.
9. Incubate for 2 h at 37°C in a thermocycler protected from light.
10. Stop the reaction by addition of 5 μL 0.5M EDTA to each tube. Total volume = 55 μL.
11. Add 5.8 μL 5M NaCl to each tube. Total volume = 60.8 μL
12. Vortex briefly, spin, and transfer the entire contents to a 1.5-mL tube containing 55 μL isopropanol. Total volume = 115.8 μL
13. Vortex well. Incubate for 10 min at room temperature; protected from light.
14. Centrifuge at 12,000x g for 10 min. Remove supernatant with a pipette. Pellet should be pink for Cy3-labeled samples and light blue for Cy5-labeled samples.
15. Rinse pellet with 500 μL 80% ice-cold ethanol. Dislodge pellet from tube wall.
16. Centrifuge at 12,000x g for 2 min. Remove supernatant with a pipette.
17. Dry contents in a SpeedVac on low heat until dry (~5 min), protected from light.
18. STOP POINT: Proceed to Step 19, or store labeled samples at −20°C, protected from light, until ready to continue experiment.
19. Spin tubes briefly prior to opening. Rehydrate pellets in 12.5 μL nuclease-free water.
20. Vortex for 30 s and quick-spin to collect contents in bottom of the tube. Continue to vortex or let sit at room temperature, protected from light, until the pellet is completely rehydrated, then vortex again and quick-spin.
21. Quantify each sample using the following formula: concentration (μg/mL) = A260 × 50 × dilution factor.

22. The amount of Cy3- or Cy5-labeled cDNA sample required for each hybridization will be dependent on which array type is being used. For example, 4-plex Nimblegen or Agilent arrays would require 2 μg per sample.
23. STOP POINT: Dry contents in a SpeedVac on low heat, protected from light. Store samples at −80 or −20°C.