Published in final edited form as:

Methods Mol Biol. 2018 ; 1709: 233–252. doi:10.1007/978-1-4939-7477-1_18.

A workflow guide to RNA-seq analysis of chaperone function and beyond

Benjamin J. Lang^{1,*}, Kristina M. Holton^{2,*}, Jianlin Gong³, Stuart K. Calderwood⁴

¹Department of Radiation Oncology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, MA, 02215, USA.

²Research Computing, Harvard Medical School, Boston, MA, USA.

³Department of Medicine, Boston University School of Medicine, Boston, MA, USA.

⁴Department of Radiation Oncology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, MA, 02215, USA.

Abstract

RNA sequencing (RNA-seq) is a powerful method of transcript analysis that allows for the sequence identification and quantification of cellular transcripts. RNA-seq has many applications including differential gene expression (DE) analysis, gene fusion detection, allele-specific expression, isoform and splice variant quantification and identification of novel genes. These applications can be used for downstream systems biology analyses such as gene ontology analysis to provide insight into cellular processes altered between biological conditions. Given the wide range of signaling pathways subject to chaperone activity as well numerous chaperone functions in RNA metabolism, RNA-seq may provide a valuable tool for study of chaperone proteins in biology and disease. This chapter outlines an example RNA-seq workflow to determine differentially expressed (DE) genes between two or more sample conditions and provides some considerations for RNA-seq experimental design.

1. Introduction

RNA-seq is a versatile method of transcript analysis that utilizes sequenced RNA-derived cDNA libraries to assess gene expression. In the past decade, its use has expanded dramatically (Figure 1). This has been facilitated by decreasing costs of next-generation sequencing, the availability of commercial kits for library construction, and continued

bjlang@bidmc.harvard.edu.
*equal contributors

4.7 Additional notes

Where institutional access to a high-performance computing cluster is not available, several commercial analysis services are available. These include the amazon cloud computing service which uses a UNIX command line and many of the steps outlined in this protocol could be adapted to using the amazon cloud, see Griffith et al., for reference [54]. The Xsede project is also openly-available (<https://www.xsede.org/>) and provides computing resources to perform RNA-seq analysis. Online discussion forums such as biostars (<https://www.biostars.org/>) and seqanswers (<http://seqanswers.com/>) are valuable tools to help navigate issues commonly faced when working through an RNA-seq workflow.

Declaration of Interest

Kristina M Holton is currently a Campus Champion for the XSEDE project.

refinement of methods for downstream analysis. In addition to these factors, the increasing utilization of RNA-seq has been driven by advantages it provides over previous transcriptomic methods such as tiling microarrays and cDNA Sanger sequencing, reviewed in [1, 2]. Features of RNA-seq include its wide quantitative range, high sequencing accuracy, conduciveness to high-throughput applications, adaptability to large genomes and detection of lowly expressed genes, low background noise and that it is not completely dependent upon the existence of an annotated genome. Unlike RT-qPCR, RNA-seq allows for transcript quantification across the entire gene body simultaneously and therefore has a greater capacity to distinguish between splice variants and isoforms.

Beyond their proteostatic functions, several studies have demonstrated chaperones to have various regulatory functions in transcription and mRNA metabolism. These have included roles for HSP90 members in RNA polymerase II complex assembly, facilitation of transcriptional pausing, mRNA splicing and catalytic activity of the RNA-induced silencing complex (RISC) [3–6]. In addition, HSP70 family members bind to and mediate mRNA stability and decay [7–12], of note, these properties can be distinct between HSP paralogs and regulated by co-chaperones [13, 14]. These studies indicate a broad potential for chaperones to have significant impact upon transcriptomic profiles, however few are yet to assess chaperone functions at a transcriptomic level. Given its above-mentioned advantages, RNA-seq is likely to be a highly beneficial for future studies of RNA-chaperone functional relationships.

The RNA-seq workflow begins with the experimental conditions from which RNA samples are collected, followed by total RNA extraction, construction of cDNA libraries, sequencing of cDNA libraries and subsequent bioinformatic analyses (Figure 2). These general steps are somewhat universal for studies using RNA-seq however the parameters and method within each step are typically selected for the chosen application. This is reflected by the large variation in published RNA-seq protocols [15]. Bioinformatic analysis of RNA-seq data typically involves performing sequence quality control, alignment of sequences to a reference genome, generation of gene feature count data, followed by various assessments of gene transcript abundance or identification including DE analysis, refinement of gene models, splice variant analysis, isoform expression quantification, gene fusion detection and allele-specific expression. This chapter is intended to provide an RNA-seq workflow protocol for differential gene expression analysis of coding genes. Investigators may adapt the protocol to their chosen application and we provide some guidance for making such changes.

2. Materials

2.1 Samples

1. Samples for RNA extraction from cell culture or animal-derived tissue

2.2 Tissue harvest and RNA isolation

1. Qiagen QIAshredder cat. no. 79654
2. Qiagen RNeasy® mini kit cat. no. 217004

3. β -mercaptoethanol
4. RNase-free microcentrifuge tubes
5. RNase-away Molecular BioProducts cat. no. 7002
6. DNase I Qiagen cat. no. 79254
7. Cell culture plates
8. 70% ethanol

2.3 Assessment of the concentration, purity and integrity of the RNA sample

1. Agilent® bioanalyzer with Expert 2100 software
2. Agilent® RNA 6000 Nano microfluidic chips
3. Reagents and consumables for Agilent RNA 6000 Nano microfluidic chip

This instrumentation and service can commonly be accessed at core facilities.

2.4 cDNA library construction (Service often available at Next-generation sequencing facility)

1. Solid-Phase Reversible Immobilization (SPRI) nucleotide-binding paramagnetic beads, e.g. Agencourt® AMPure® XP (Beckman Coulter cat no. A63880) or KAPA Pure Beads (KAPA Biosystems)
2. Illumina® TruSeq index adapters
3. Thermocycler
4. Magnetic rack
5. Low-adhesion, RNase-free, DNase-free, 1.5ml microcentrifuge tubes
6. Low-adhesion, RNase-free, DNase-free, PCR tubes sized for thermocycler to be used
7. Agilent® bioanalyzer tapestation
8. Filtered pipette tips
9. KAPA® Stranded RNA-Seq with RiboErase cat. no. KK8483.

2.5 Next-generation sequencing

1. Illumina® HiSeq 2500 sequencer

2.6 Sequence processing and analysis

1. High performance computer cluster with UNIX command line
2. Programs for;
 - assessing read quality; FastqQC
 - trimming/filtering reads; Flexbar [16], Trimmomatic [17]

- read alignment; TopHat2/Bowtie 2 [18], HISAT [19], STAR [19]
 - assigning read counts to annotated genes; samtools, htseq count [20]
3. R version 3.X.X, Bioconductor software packages [21], edgeR [22], GOseq [23], ggplot2 [24].

3. Methods

3.1 Experimental design

1. Determine the number of samples needed for each condition. See section 4.1 for further guidance on choosing number of biological replicates.
2. Determine the sequencing platform to be used. For this protocol an Illumina® HiSeq 2500 sequencer platform is used.
3. Decide how the cDNA libraries will be constructed, stranded or un-stranded. This protocol uses the KAPA stranded RNA-Seq with RiboErase kit.
4. Determine the minimum read depth needed and the cost. See section 4.1.

3.2a RNA collection from cell culture

1. Prepare experimental samples ready for RNA collection.
2. Clean working area and pipettes with ethanol and RNase away.
3. Remove media from culture dishes as much as possible. Note that the maximum number of animal cells recommended for the use of the RNeasy® Plus Mini kit protocol is 1×10^7 .
4. Following the RNeasy® Plus Mini Handbook, use RLT buffer with β -mercaptoethanol to lyse the cells and collect cell lysate in a new RNase free microcentrifuge tube.
5. Homogenize the sample by passing it through a QIAshredder column.
6. Continue protocol with inclusion of the optional DNase I digestion step.
7. Continue to step 3.3 of this protocol, RNA integrity analysis or store at -80°C .

Note that RNeasy® mini kit buffers RLT and RW1 should not be mixed with bleach.

3.2b RNA extraction from mouse tissue

1. Sacrifice mice per ethical protocol.
2. Excise the tissue of interest from the mouse.
3. Mince up to 30mg of tissue to small pieces (1–3 mm) with scissors, and transfer into round bottom tube with appropriate volume of cell lysis buffer RLT with β -mercaptoethanol per RNeasy® Plus Mini Handbook.
4. Homogenize the sample by passing it through a QIAshredder column.
5. Continue protocol with inclusion of the optional DNase I digestion step.

6. Continue to step 3.3 of this protocol, RNA integrity analysis or store at -80°C .

Note that RNeasy® mini kit buffers RLT and RW1 should not be mixed with bleach.

3.3 Assess the concentration and integrity of the RNA sample

1. Quantify the concentration of RNA by absorbance spectrophotometry.
2. Assess RNA integrity using Agilent bioanalyzer. Submit 2 to 4 μl of RNA sample to core facility for analysis or if instrument is available in-house, perform analysis per manufacturer's directions.
3. Proceed with protocol if sample RNA is sufficiently intact (see section 4.3)

3.4 Stranded cDNA library construction using KAPA stranded RNA-seq Kit with RiboErase for Illumina® platforms

Follow the manufacturer's protocol to complete these steps of cDNA library construction for RNA-seq using the Illumina® platform;

1. rRNA depletion to enrich for coding genes
2. Fragmentation of RNA
3. Reverse transcription of fragmented RNA to produce 1st strand cDNA
4. Second strand synthesis, 2nd strand marked with dUTP
5. Ligation of dAMP to 3' ends of dscDNA strands
6. Ligation of indexed adaptors to each sample dscDNA
7. PCR amplification of cDNA libraries

Note that numerous SPRI purifications are involved throughout the above steps and are dependent upon the protocol used. When using Agencourt® AMPure® XP or KAPA Pure Beads®, it is essential the RNA/DNA-binding steps are performed at room temperature for the RNA/DNA to associate with the beads. RNA/cDNA/Bead solutions that are not equilibrated to room-temperature will result in sample loss.

3.5 Next-generation sequencing

1. Perform library quality control and quantification of cDNA libraries.
2. Pool/ multiplex the cDNA libraries at equal proportions to create master pool mix.
3. Run the cDNA library pool on an Illumina® HiSeq 2500 sequencer flow cell or platform of choice.

3.6 Sequence processing and analysis

1. Export FASTQ files to remote directory- these are the raw sequence files containing the sequences or 'reads' for each indexed sample.

2. Perform quality control upon ‘.fastq’ files (raw reads) to remove low quality reads and to remove or ‘trim’ the indexed adaptors from the reads.
3. Align reads to reference genome using alignment software, e.g. tophat2 [18], STAR [25], HISAT2 [19].
4. Generate a counts matrix using ‘htseq count’ to count the number of reads aligning to each annotated gene [20].
5. Compare gene-wise expression (differential gene expression ‘DE’) across samples or conditions using EdgeR [22].
6. Assess changes in cellular or biological processes between conditions through gene ontology analysis using GOseq [23].
7. Representation of data.

4. Notes

4.1 Experimental design

Immediate RNA-seq experimental design decisions include the number of biological replicates for each condition, whether the cDNA libraries will be stranded or un-stranded, whether the sequencing run will be paired end (PE) or single end (SE), read depth, read length and cost. Choosing the most cost and statistically effective combination of these parameters can be challenging as there is no current standard, however recent studies have provided direction on how each of these parameters should be prioritized according to the experimental application [26–29].

One starting point is determining how many biological replicates the experiment will entail. Biological replicates are essential for encapsulating the biological variation within replicates and between conditions in the statistical analysis [29, 30]. For DE analysis, the investigator may perform a power analysis to predict the number of replicates needed based upon desired power, biological coefficient of variation (BCV) and effect size. This can be a challenge if the BCV for the model is not known. If a costly experiment is planned, it may pay to perform a pilot study to approximate the BCV for power analysis. A typical biological coefficient of variation (BCV) for animal tissue-derived samples is 0.2–0.6 and therefore may require more replicates to achieve a given power and effect size than samples derived from cell lines (BCV closer to 0.1). A study by Schurch et al., recommended 6 replicates to be the minimum for each condition, and at least 12 replicates for each condition if all DE genes are to be determined [28]. Next generation sequencing (NGS) was shown to be highly reproducible, therefore prioritizing biological replication over technical replication will benefit DE analysis [31].

Increasing the number of sequence reads attributed to each biological sample, or ‘read depth’, increases DE power up to a point of saturation [32]. A limiting factor for this parameter is commonly the cost of additional sequencing lanes and flow cells, and it is therefore balanced across an optimal number of biological replicates, the relationship of which was examined by Liu et al., [26]. For example, the Illumina® HiSeq 2500 platform

yields approximately 180M reads per sequencing lane, if the cDNA library from one biological sample is applied to one flow cell lane, the read depth would be approximately 180M reads. For DE analysis of coding genes, allocating a pooled library of 10–12 cDNA libraries to one flow cell lane will yield approximately 18 to 15 million reads per sample, respectively. According to findings by Liu et al., aiming for 10–20 million reads per sample is the most cost-efficient distribution of reads to gain maximum statistical power for DE RNA-seq [26].

Strand-marked or ‘stranded’ cDNA libraries provide additional mapping information over un-stranded cDNA libraries. In contrast to un-stranded libraries, stranded cDNA libraries allow the RNA template strand to be recognized. This is an important distinction in regions where overlapping gene features exist on opposing DNA strands or when the anti-sense strand is also transcribed [33]. Levin et al., assessed different methods of strand marking protocols for RNA-seq and concluded dUTP marking of the second strand with paired-end sequencing to be a preferred method using the Illumina® platform [33]. Opting for paired-end (PE) sequencing is generally considered a cost-effective choice to yield greater information from the RNA-seq experiment. PE RNA-seq improves read mapping to features and is also beneficial for splice variant analysis. For DE analysis, 50nt sequence cycles/ read length is sufficient to assign reads to gene features, with very few genes mapping to multiple features [34]. Longer read lengths will significantly increase the cost and provides greater benefit for other applications.

As much as possible, the handling of RNA samples and processing of all cDNA libraries should be kept as consistent as possible throughout the protocol. Differences in sample processing can influence gene calls downstream and lead to batch effects in the data. Factors in an RNA-seq experiment that may contribute to a batch effect include: experiment date, isolation date, sample storage, sample age, sample gender, isolation area, isolation kit, researcher, and lab [27]. It is best practice to construct all cDNA libraries in concert and to sequence the samples as normalized-pooled cDNA libraries on the same flow cell lane if possible, or equal numbers of each condition spread across multiple flow cell lanes.

4.2 RNA collection

Careful RNA isolation, handling and storage is critical for generating valid RNA-seq data. If using a commercial column-based extraction method, the column should be treated with DNase to reduce the chance of genomic contamination of the library preparations. Care should also be taken to prevent any buffer residue on the outside of the column reaching the final elution and thereby avoiding unwanted impurities entering the sample. RNases are ubiquitous and steps to protect the RNA sample from RNases such as using clean gloves, and wiping the working area and pipettes with RNase solution are measure that can counter RNase contamination. Progressing through the RNA extraction protocol as promptly as possible as well as swift storage at -80°C are also measures the investigator can take to limit sample degradation. We recommend either completing the RNA extraction protocol in its entirety once the sample is lysed or if commercial RNA preservation reagents are used, then to closely follow the recommendations for their use. Repeated freeze-thaw cycles can lead to reduced RNA integrity and should be avoided.

4.3 RNA integrity analysis

Prior to construction of cDNA libraries for sequencing, it is important to assess the RNA integrity of each sample. This step not only provides confidence going forward through the workflow, but can save a great deal of time and downstream heartache where poor sample quality leads to misleading or no results. Currently the best method to determine RNA integrity within a sample is to use the Agilent Bioanalyzer system for RNA integrity number (RIN) analysis. The RIN is an arbitrary number derived from an algorithm that uses the fluorescent spectra generated by running an RNA sample through the Agilent Bioanalyzer microcapillary electrophoresis system [35, 36]. Up to 12 samples can be run on one chip at a time and the analysis simultaneously determines RNA concentration in the sample. An example of the fluorescent spectra generated is shown in Figure 3, where large peaks can be observed in samples with high integrity that correspond to 18S and 28S ribosomal RNA (Figure 3A), that are absent in degraded samples (Figure 3B). The algorithm used to generate the RIN number encompasses multiple features of RNA degradation that provides a more robust and reproducible method of determining RNA integrity than assessments based upon the 18S:28S ratio alone [35, 36].

4.4 cDNA library construction

4.4.1 —Ribosomal RNA (rRNA) makes up 80–90% of total RNA and for DE analysis of coding genes, enrichment of non-rRNA is commonly employed to ensure that rRNA does not consume most sequence reads [29]. One method to enrich for coding genes is selection of polyA-mRNAs using poly-dT beads or poly-dT primers for reverse transcription [29]. This is a highly effective method of reducing downstream reads being allocated to rRNAs, however be aware that poly-dT-based library construction can bias representation of 3' reads [1, 37]. A common alternative to polyA-based enrichment is rRNA-depletion followed by random hexamer-primed reverse transcription, which too is subject to some bias, however to a lesser degree [38]. The KAPA stranded RNA-seq with RiboErase kit used in this protocol depletes rRNA by hybridizing rRNA-complementary DNA oligonucleotides to rRNA and degradation of RNA:DNA hybrids by RNase H while sparing the non-hybridized non-rRNA RNA molecules. The starting volume of total RNA used for the rRNA-depletion stage of the KAPA stranded RNA-seq with RiboErase protocol is 10µl. Use the same mass of starting RNA across the samples, the amount that remains post-rRNA reduction will vary with variations in the proportion of rRNA across the samples, however it is not necessary to re-quantify the rRNA-depleted RNA as variations in rRNA-depleted RNA input will be compensated for downstream when the cDNA libraries are pooled based upon cDNA concentration prior to sequencing.

4.4.2 —For fragmentation of RNA, the KAPA stranded RNA-seq Kit with RiboErase kit uses heat treatment in the presence of magnesium. Varying the time and heat treatment will result in fragments of dependent length. For RNA-seq analysis of coding gene expression, we have typically used 6 min at 94°C as designated by the kit's protocol to achieve fragments of approximately 200–300bp. Optimal fragment length can be application-dependent and may also influence how many sequencing cycles one chooses. For example, some sequencing redundancy will occur if 100bp fragments are generated and more than 50

sequencing cycles PE are used at the sequencing stage as the same sequence information will be duplicated.

4.4.3 —If using a strand-specific protocol such as the KAPA stranded RNA-seq Kit with RiboErase kit, it is important to be aware that the second strand marked with dUTP will not be amplified at the latter PCR stage (3.4.7). The complement of the coding gene-region will therefore be amplified and sequenced. When using the sequence alignment software, be sure to designate the strandedness of the cDNA libraries. This also applies for counting reads to gene features at stage (Htseq count- 3.6.4) where the `--stranded= reverse` option needs to be designated.

4.4.4 —A unique indexed adapter is ligated to the cDNA libraries of each biological sample. The KAPA stranded RNA-seq Kit with RiboErase kit provides a guide to how much adapter should be included in each ligation reaction and is proportionate to the amount of input RNA. The indexed adaptors serve two purposes. First, the ends of the adaptors allow for the cDNA libraries to bind the sequencing flow cell. Second, a unique 6-nucleotide sequence identifier or ‘index’ allows for the cDNA libraries to be pooled together and run on the same flow cell lane as each sequence read is attributed to an individual biological sample based upon the unique index identifier. The adenylation of the 3’ ends of dscDNA allows for ligation of the adaptors at 3’ ends, the indexed adaptors are also ligated to the 5’ ends. The inclusion of adaptors at both ends of dscDNA allows for paired-end sequencing (PE), where the final cDNA library fragments are sequenced from both directions. PE-sequencing is desirable as it allows for sequences to be located to the genome with greater confidence and allows quantification of alternatively spliced transcripts.

4.4.5 —PCR amplification of cDNA libraries is the final step of cDNA library construction. PCR primers complementary to the Illumina® adaptors are added and the cDNA libraries are amplified by PCR for 8–16 cycles. The cycle number recommended by the KAPA protocol is dependent upon the amount of input RNA. The goal of this amplification stage is to increase the cDNA library yield while maintaining a representative distribution of the transcript-derived cDNA molecules. Molecules of cDNA representing transcripts in high abundance will be enriched more rapidly than cDNA molecules representing transcripts of low abundance and therefore choosing the optimum number of amplification cycles is a balance between increasing the cDNA library to a sufficient yield for sequencing while limiting over-amplification of highly abundant cDNAs. Over amplification will lead to highly abundant cDNAs taking up a disproportionate number of binding sites on the flow-cell and underrepresentation of lowly expressed transcripts. After the post-PCR purification stage, cDNA libraries should be stored at -20°C .

4.5 Next-generation sequencing

Once the cDNA libraries have been constructed it is important to quantify the cDNA concentration and assess the size distribution of each cDNA library. This is commonly achieved using the Agilent tapestation system to assess cDNA size distribution and by qPCR to determine cDNA library concentration [34]. The cDNA libraries are then normalized by

concentration and combined or ‘pooled’ to form a master/ parent cDNA library that will be run on an Illumina® flow cell lane.

As described in 4.1, the sequencing parameters the investigator must choose are application-dependent. These parameters include whether the sequencing run is to be paired-end (PE) or single end (SE) and the number of sequencing cycles.

4.6 Sequence processing and analysis

4.6.1 Quality control of raw reads—Quality control should always be performed as a first step to understanding the fastq sequence generated by RNA-seq. A fastq file is essentially a fasta (sequence) file, but with an added line, for the PHRED (quality) score, the scale of which depends on the sequencer generating it. Quality/degradation of input RNA, RNA library preparation, model and chemistry kit of the Next-Gen Sequencer, and length of sequence can all affect the consistency and quality of sequence throughout the read. The Java program FastQC generates an html report, including crucial graphical metrics like per base sequence quality, average quality per read, sequence duplication levels, and overrepresented sequence. With these metrics, a researcher can decide whether to trim their sequence to maximize alignment success based on a blunt end cut-off, or quality threshold. Sometimes barcodes or adapters have not been removed, and this should be addressed as well before proceeding with alignment. There are a myriad of trimming/grooming programs available, including: Cutadapt/Trim Galore [39], FASTX-Toolkit, Skewer [40], PRINSEQ [41], Trimmomatic [17], to name a few. FastQC should be re-run on trimmed/groomed sequence for validation. For more advanced QC metrics and visualization, RSeQC contains an expansive collection of quality control tools and sequencing metrics, along with format conversion.

4.6.2 Alignment to reference genome—TopHat2 with its underlying dependency on the Bowtie2 aligner were once the standard for Illumina® sequencing alignment [18]. Taking an indexed genome along with an optional genome annotation file (GAF/GTF), TopHat2 can discern reads that span splice junctions, and detect novel isoforms. TopHat2 is a command line Unix tool, and as such, is best suited to run in a High Performance Compute (HPC) environment. It can distribute its computation over multiple processors to speed up alignment, which is a workflow very amenable to HPC.

TopHat2 is now becoming deprecated, and replaced with HISAT2, by the same group at Johns Hopkins University [19]. The syntax for mismatches and multi-mappers is very similar, and HISAT2 is not plagued by many of the high memory requirements of other ultrafast aligners.

STAR aligner emerged as an ultrafast alignment alternative to TopHat2, equally capable of mapping reads across splice junctions and detecting novel isoforms [25]. STAR is a Unix tool with higher memory requirements, so is best in an HPC environment. STAR affords greater options and specificity as to mismatches, multimappers, and output format, and includes verbose alignment statistics.

4.6.3 Differential expression analysis—The powerful R package edgeR is available through Bioconductor to perform differential expression (DE) analysis to derive a list of genes that statistically differ between conditions [22]. The edgeR package is able to derive meaningful biological signal from a low number of replicates, in terms of number of significant genes and precision, compared to other algorithms [28, 42].

The edgeR package performs DE analysis on raw counts data, as opposed to RPKM or FPKM, which have been shown to not effectively eliminate gene length bias [43]. Instead, by considering each gene and acknowledging that each gene length and therefore bias is the same between conditions, direct gene-wise comparisons can be made between conditions. Gene length can later be factored in during functional enrichment analysis, where the context of the DE genes is explored.

Lowly-expressed genes (usually 5–10 counts per million (CPM)) are often considered to be noise, and are conditionally omitted out of the analysis, to prevent them from skewing the statistics. This thresholding can be set to a determined number of samples: all the genes in condition 1 must have a CPM > 5 to be considered, etc. Removing the lowly expressed genes shrinks the size of the input matrix to an easier-to-compute size.

While RPKM or FPKM normalization is not applied, the input data matrix still consists of varying library sizes (total read counts per sample) that must be normalized to allow for even-footing comparison. Highly-expressed genes may be dominating the total number of reads, and need to be compensated for. The edgeR package uses trimmed means of M normalization (TMM) [43]. TMM minimizes the log-fold changes between samples and derives a scaling factor. A scaling factor below one scales the counts up and the library size down, indicating that “small number of high count genes are monopolizing the library size.” [43]. Conversely, a scaling factor above one scales the counts down and the library size up. With the scaling factor, the counts data has a new, effective library size.

It is often useful to visualize how samples cluster using unsupervised methods. Multi-dimensional scaling (MDS) plots the leading log-fold change in the highest two dimensions. Ideally, samples from conditions will separate along the primary X-axis (Figure 4). Hierarchical clustering on the top quartile varying genes, possibly in conjunction with a heatmap, can also identify any issues with sample consistency. Algorithms like surrogate vector analysis (SVA) [44], can be used to create a modified count to use in edgeR and correct for batch effects.

Like many popular R DE algorithms, edgeR DE is based on a Negative Binomial (NB) model. A NB model is an over-dispersed Poisson model, that is, it has a variation term. The variation is how much the genes in samples within a condition differ from their mean. The biological coefficient of variation (BCV) for an ideal model cell line is around 0.1, but is often as high as 0.6 for human samples, McCarthy et al. reported BCV of 0.4 to be typical of human samples [45]. The edgeR algorithm encapsulates the BCV in two ways: common and gene-wise. The common dispersion assumes that genes all have the same mean-variance relationship. This dispersion is further refined on a tag-wise (gene) level with empirical Bayesian shrinkage; genes more consistent between replicates are ranked higher [45].

The edgeR has two major implementations: “classic” and “generalized linear model” (GLM). The classic mode is for experiments with only two conditions (one factor), and uses quantile-adjusted conditional maximum likelihood (qCML) to determine if a gene differs between conditions. The common and tagwise dispersions works on pseudo counts conditioned on the effective library size. The actual DE is an exact test using the negative binomial distribution. The resultant p-values are then corrected to q-values for multiple testing errors with a false discovery rate, using Benjamini-Hochberg (BH) correction [46].

For multiple factor experimental designs, the GLM approach is more appropriate. The common and tagwise dispersion are estimated using Cox-Reid profile-adjusted likelihood (CR) by fitting generalized linear models to the design (experimental setup) matrix. The DE can be calculated between any pairwise factors using a GLM likelihood ratio test. P-values are corrected for multiple testing with BH to get resultant q-values.

Some investigators choose to threshold q-values with a combination of log-fold change (logFC) to filter out only the statistically-significant genes with the highest differential signal. Differentially expressed genes are most commonly represented as logFC to provide a scale upon which observed fold-changes can be represented.

4.6.4 Gene ontology analysis using GOseq—Interrogating the biological context of a DE gene list is imperative to understanding what genes up-regulated and down-regulated do together. Gene Ontology (GO) is controlled vocabulary to describe a gene by three features: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). The Gene Ontology Consortium maintains a constantly-updated catalog of ontology for many model organisms, and is manually curated and evidence-based [47]. By running variations of a Fisher’s Exact Test against a DE list for each term’s membership, a p-value and corrected q-value is derived, to determine if a GO term is significant in the dataset.

For RNA-seq data, the question of gene length is not addressed in the edgeR differential, as each gene is considered significant in isolate. But there is still a systemic bias of longer and more highly-expressed transcripts being detected, and achieving greater statistical power, skewing the overall analysis [48]. To perform any sort of systems biology approach, like using GO terms, the gene length bias needs to be addressed [23].

The GSeq package in R was designed specifically to perform functional enrichment analysis of NGS data using GO terms, while addressing gene length bias and creating custom null distributions [23]. Using a Probability Weighting Function (PWF) against genes in the DE and not in the DE, the new probability that a gene is differentially expressed based on its transcript length is derived and the null distribution for each GO term is calculated. Only genes that were annotated in the analysis are included; this provides a statistical advantage over some GO approaches, which use a “one background fits all” approach. The GSeq authors find that simple hypergeometric distribution as the null distribution is not sufficient to model the data appropriately. While resampling is preferred, this is computationally intensive. Instead, GSeq uses the Wallenius non-central hypergeometric distribution, which assumes that “all genes within in a category have the same probability of being chosen, but this probability is different from the probability of choosing genes outside

this category” [23]. The resultant p-values are corrected for multiple testing using Benjamini-Hochberg FDR, and GO terms meeting the prespecified alpha are considered significant.

Because of the nature of GO terms, that is, they are organized in a Directed Acyclic Graph (DAG), where more specific, smaller membership terms belong to higher-level, larger membership terms, it is often difficult to tease out the exact context of the bigger picture. Tools like REVIGO (Reduce and Visualize Gene Ontology) take a list of significant GO terms as input and cluster related terms, creating scatterplots, interactive graphs, and tag clouds [49]. The igraph tool, available in R, Python, and C/C++, is another popular alternative to visualizing the networks from Gene Ontology [50].

4.6.5 Representation of data—Data visualization plays a crucial role in conveying the results of an RNA-seq experiment. Beginning with the cursory glance at the alignment, Integrative Genomics Viewer (IGV) can display how reads map to the reference genome (Figure 5) [52, 53]. Heatmaps with dendrograms provide a powerful visualization of a DE list, which is more manageable than the entire dataset. The heatmap.2 function in the R package ggplot2 provides powerful functionality to create information-rich heatmaps with dendrograms [24]. Some experimentalists use raw counts and adjust their color scaling. Another popular option is to calculate z-scores per gene, and graph these z-scores. Hierarchical clustering calculates the distance between samples and between genes. Samples of the same experimental factor should cluster together. Genes will cluster according to count or z-score in high-low blocks. Generic choices of clustering include Euclidean, Manhattan, and Pearson Correlation distance metrics. Branches can be cut as appropriate to denote sample or gene clusters.

Acknowledgements

We thank Thiago De Borges (BIDMC, Harvard Medical School) for critical reading of this chapter.

References

1. Wang Z, Gerstein M, and Snyder M, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 2009 10(1): p. 57–63.
2. Weber AP, Discovering New Biology through Sequencing of RNA. *Plant Physiol*, 2015 169(3): p. 1524–31. [PubMed: 26353759]
3. Boulon S, et al., HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Mol Cell*, 2010 39(6): p. 912–24. [PubMed: 20864038]
4. Iki T, et al., In vitro assembly of plant RNA-induced silencing complexes facilitated by molecular chaperone HSP90. *Mol Cell*, 2010 39(2): p. 282–91. [PubMed: 20605502]
5. Sawarkar R, Sievers C, and Paro R, Hsp90 globally targets paused RNA polymerase to regulate gene expression in response to environmental stimuli. *Cell*, 2012 149(4): p. 807–18. [PubMed: 22579285]
6. Ferraldeschi R, et al., Second-Generation HSP90 Inhibitor Onalespib Blocks mRNA Splicing of Androgen Receptor Variant 7 in Prostate Cancer Cells. *Cancer Res*, 2016 76(9): p. 2731–42. [PubMed: 27197266]
7. Duttagupta R, et al., A Yeast Homologue of Hsp70, Ssa1p, Regulates Turnover of the MFA2 Transcript through Its AU-Rich 3' Untranslated Region. *Molecular and Cellular Biology*, 2003 23(8): p. 2623–2632. [PubMed: 12665566]

8. Henics T, et al., Mammalian Hsp70 and Hsp110 Proteins Bind to RNA Motifs Involved in mRNA Stability. *The Journal of Biological Chemistry*, 1999 274(24): p. 17318–17324. [PubMed: 10358092]
9. Kishor A, et al., Hsp70 is a novel posttranscriptional regulator of gene expression that binds and stabilizes selected mRNAs containing AU-rich elements. *Mol Cell Biol*, 2013 33(1): p. 71–84. [PubMed: 23109422]
10. Laroia G, et al., Control of mRNA Decay by Heat Shock-Ubiquitin-Proteasome Pathway. *Science*, 1999 284(5413): p. 499–502. [PubMed: 10205060]
11. Marucci A, et al., The role of HSP70 on ENPP1 expression and insulin-receptor activation. *J Mol Med (Berl)*, 2009 87(2): p. 139–44. [PubMed: 19083193]
12. Wilson GM, et al., Thermodynamics and kinetics of Hsp70 association with A + U-rich mRNA-destabilizing sequences. *J Biol Chem*, 2001 276(48): p. 44450–6. [PubMed: 11581272]
13. Matsui H, Asou H, and Inaba T, *Cytokines direct the regulation of Bim mRNA stability by heat-shock cognate protein 70*. *Mol Cell*, 2007 25(1): p. 99–112. [PubMed: 17218274]
14. Zimmer C, Gabain AV, and Henics T, Analysis of sequence-specific binding of RNA to Hsp70 and its various homologs indicates the involvement of N- and C-terminal interactions. *RNA*, 2001 7: p. 1628–1637. [PubMed: 11720291]
15. Conesa A, et al., A survey of best practices for RNA-seq data analysis. *Genome Biol*, 2016 17: p. 13. [PubMed: 26813401]
16. Dodt M, et al., Flexbar- flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, 2012 1(3): p. 895–905. [PubMed: 24832523]
17. Bolger AM, Lohse M, and Usadel B, Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, 2014 30(15): p. 2114–2120. [PubMed: 24695404]
18. Trapnell C, Pachter L, and Salzberg SL, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 2009 25(9): p. 1105–11. [PubMed: 19289445]
19. Kim D, Langmead B, and Salzberg SL, HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 2015 12: p. 357–360. [PubMed: 25751142]
20. Anders S, Pyl PT, and Huber W, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 2015 31(2): p. 166–9. [PubMed: 25260700]
21. Gentleman R, et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 2004 5: p. R80. [PubMed: 15461798]
22. Robinson MD, McCarthy DJ, and Smyth GK, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010 26(1): p. 139–40. [PubMed: 19910308]
23. Young MD, Wakefield MJ, and Smyth GKO, A., *Gene ontology analysis for RNA-seq: accounting for selection bias*. *Genome Biology*, 2010 11(R14).
24. Wickham H, ggplot2: Elegant Graphics for Data Analysis. 2009: Springer-Verlag New York.
25. Dobin A, et al., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013 29(1): p. 15–21. [PubMed: 23104886]
26. Liu Y, Zhou J, and White KP, RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 2014 30(3): p. 301–4. [PubMed: 24319002]
27. Fang Z and Cui X, Design and validation issues in RNA-seq experiments. *Brief Bioinform*, 2011 12(3): p. 280–7. [PubMed: 21498551]
28. Schurch NJ, et al., *How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?* *RNA*, 2016 22(6): p. 839–51. [PubMed: 27022035]
29. Sims D, et al., Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*, 2014 15(2): p. 121–32. [PubMed: 24434847]
30. Anders S, Reyes A, and Huber W, Detecting differential usage of exons from RNA-seq data. *Genome Res*, 2012 22(10): p. 2008–17. [PubMed: 22722343]
31. Marioni JC, et al., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 2008 18(9): p. 1509–17. [PubMed: 18550803]

32. Tarazona S, et al., Differential expression in RNA-seq: a matter of depth. *Genome Res*, 2011 21(12): p. 2213–23. [PubMed: 21903743]
33. Levin JZ, et al., Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*, 2010 7(9): p. 709–15. [PubMed: 20711195]
34. Korpelainen E, et al., *RNA-seq Data Analysis: A Practical Approach Mathematical and Computational Biology Series*, ed. Britton NF, et al. 2015, Boca Raton, Florida, USA: Chapman & Hall/ CRC Press.
35. Schroeder A, et al., The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol*, 2006 7: p. 3. [PubMed: 16448564]
36. Imbeaud S, et al., Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res*, 2005 33(6): p. e56. [PubMed: 15800207]
37. Shanker S, et al., Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech*, 2015 26(1): p. 4–18. [PubMed: 25649271]
38. Hansen KD, Brenner SE, and Dudoit S, Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*, 2010 38(12): p. e131. [PubMed: 20395217]
39. Martin M, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 2011 17(1): p. 10–12.
40. Jiang H, et al., Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 2014 15: p. 182. [PubMed: 24925680]
41. Schmieder R and Edwards R, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 2011 27(6): p. 863–4. [PubMed: 21278185]
42. Seyednasrollah F, Laiho A, and Elo LL, Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*, 2015 16(1): p. 59–70. [PubMed: 24300110]
43. Robinson MD and Oshlack A, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 2010 11(3): p. R25. [PubMed: 20196867]
44. Leek JT and Storey JD, Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 2007 3(9): p. e161.
45. McCarthy DJ, Chen Y, and Smyth GK, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, 2012 40(10): p. 4288–97. [PubMed: 22287627]
46. Benjamini Y and Hochberg Y, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1995 57(1): p. 289–300.
47. Gene Ontology C, Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 2015 43(Database issue): p. D1049–56. [PubMed: 25428369]
48. Oshlack A and Wakefield MJ, Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, 2009 4: p. 14. [PubMed: 19371405]
49. Supek F, et al., REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One*, 2011 6(7): p. E21800. [PubMed: 21789182]
50. Csardi G and Nepusz T, *The igraph software package for complex network research*. *InterJournal*, 2006 *Complex Systems*: p. 1695.
51. Gong J, et al., Targeting the hsp70 gene delays mammary tumor initiation and inhibits tumor cell metastasis. *Oncogene*, 2015 34(43): p. 5460–71. [PubMed: 25659585]
52. Robinson JT, et al., Integrative Genomics Viewer. *Nature Biotechnology* 2011 29: p. 24–26.
53. Thorvaldsdottir H, Robinson JT, and Mesirov JP, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 2013 14(2): p. 178–92. [PubMed: 22517427]
54. Griffith M, et al., Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Computational Biology*, 2015 11(6): p. e1004393. [PubMed: 26248053]

Figure 1: Usage of RNA-seq has expanded in recent years.

A Pubmed search using term "RNA-seq" returned the displayed number of studies by year. Since 2008 the number of studies utilizing RNA-seq has rapidly expanded.

Figure 2:

A generalized workflow for an RNA-seq protocol examining differential gene expression (DE), and the topics of this chapter.

Figure 3:

Agilent Bioanalyzer analysis of RNA integrity. (A) An example fluorescence spectra generated by the Agilent Bioanalyzer from an intact RNA sample with RIN value of 10 with distinct peaks corresponding to 18S and 28S rRNA (B) A degraded RNA sample of low RIN value showing absence of fluorescent signal corresponding to 18S and 28S rRNA and high level of small RNA molecules.

Figure 4: Multi-dimensional scaling (MDS) plots are an informative visual representation of relatedness between samples. Ideally samples will be separated across dimension 1 based upon the biological condition such as in (A). Where samples are not separated across dimension 1 based upon biological condition such as in (B) may be indicative of a batch effect heavily influencing the gene count data.

Figure 5: Visualization of mapped reads using IGV.

An example representation of read coverage across the mouse *Hspa1a* and *Hspa1b* genes in MMT Hsp70^{-/-} and MMT WT mice [51]. Here, *Hspa1a* and *Hspa1b* coding regions have been largely deleted in the MMT Hsp70^{-/-} sample, however are still assigned read counts by htseq.