Topic Introduction

Methods for Processing High-Throughput RNA Sequencing Data

Manuel Ares Jr.

High-throughput sequencing (HTS) methods for analyzing RNA populations (RNA-Seq) are gaining rapid application to many experimental situations. The steps in an RNA-Seq experiment require thought and planning, especially because the expense in time and materials is currently higher and the protocols are far less routine than those used for other high-throughput methods, such as microarrays. As always, good experimental design will make analysis and interpretation easier. Having a clear biological question, an idea about the best way to do the experiment, and an understanding of the number of replicates needed will make the entire process more satisfying. Whether the goal is capturing transcriptome complexity from a tissue or identifying small fragments of RNA cross-linked to a protein of interest, conversion of the RNA to cDNA followed by direct sequencing using the latest methods is a developing practice, with new technical modifications and applications appearing every day. Even more rapid are the development and improvement of methods for analysis of the very large amounts of data that arrive at the end of an RNA-Seq experiment, making considerations regarding reproducibility, validation, visualization, and interpretation increasingly important. This introduction is designed to review and emphasize a pathway of analysis from experimental design through data presentation that is likely to be successful, with the recognition that better methods are right around the corner.

BACKGROUND

Advances in genome sequencing technology have quickly been applied to the study of RNA. At the moment, there are three main HTS systems—454 or “pyrosequencing,” SOLiD sequencing, and Illumina sequencing—that operate on similar principles (Mitra et al. 2003; Ehn et al. 2004; Russom et al. 2005; Shendure et al. 2005; Guo et al. 2008). Each method creates a library of seed molecules from an RNA sample that are separated, amplified individually, and placed in an array of positions on a two-dimensional surface, with each reaction center composed of many copies of the original individual RNA (actually cDNA) molecule. Sequencing reactions are performed by flooding the array with reagents (DNA polymerase, nucleotides, etc.) in a specified reaction cycle, using light-emitting reactants that allow capture of an image of the array from the photons that arise at each cycle. The reactions are set up so that the identity of each nucleotide in the seed molecule copies can be determined by analysis of sequential images. The systems use different methods to amplify each individual seed molecule and place it on the array, and they use different light-emitting chemistries to follow the sequencing reaction. SOLiD sequencing, for example, produces a read in “color space” due to the fact that each ligation step reads a dinucleotide and overlapping dinucleotide reads are used to identify the base at each position. Here, the color space data must be converted to a sequence, a process thought to add fidelity to the sequence determination but that is not needed for the other platforms. Because an array of many millions of reaction centers is imaged multiple times, the number of reads obtained in a single run can be quite huge. The amount of data to be processed, especially considering

Adapted from *RNA: A Laboratory Manual* by Donald C. Rio, Manuel Ares Jr, Gregory J. Hannon, and Timothy W. Nilsen. CSHL Press, Cold Spring Harbor, NY, USA, 2011.

© 2014 Cold Spring Harbor Laboratory Press

Cite this introduction as *Cold Spring Harb Protoc*; doi:10.1101/pdb.top083352

that the raw data are a large stack of images, makes the analysis for all three methods computationally intensive.

Despite the various approaches taken in the different HTS methods, all start with a library of seed molecules derived from the sample by ligation of adapters, cDNA synthesis, a second ligation step in some cases, PCR amplification, and size selection. These steps are necessary to properly capture the seed molecule in a form that allows bead or bridge amplification and analysis. Because the reactivity of all RNA fragments in the sample toward the primers or ligation substrates in these reactions may not be equal, the library construction steps can introduce bias into the population of seed molecules (see below). At the same time, the addition of linkers affords an opportunity for two modifications that improve cost-effectiveness and add information: “mate pair” or “paired end” read libraries and “bar coding” of samples. In libraries constructed for mate pair reads, the bead array can be denatured and reacted with a second sequencing primer to obtain sequence for the complementary strand at the other end of the seed molecule. Because seed molecules are size-fractionated during library construction, the paired end method provides pairs of sequence reads that are an approximately known distance apart. Such distance information can be very useful for genome assembly or, in the case of RNA analysis, splicing. In barcoding, the oligonucleotides used to generate the seed molecule libraries have extra bases that allow their origin to be determined in mixed runs. For example, if an investigator wants to run a time course and obtain sequence for samples at each time point, separate libraries can be constructed with different bar codes for each time point. To save on sequencing costs, the libraries can be mixed and sequenced together, and the reads can then be sorted out by their bar codes before analysis to separate molecules by their sample identity. Such multiplexing strategies reduce costs, but of course, they divide the total number of reads, so that coverage of the transcriptome may be reduced for each time point.

Once the sequence reads are obtained, their genomic origin is identified by matching their sequence to the genome, and the number of reads found to match a particular gene can be taken as a measure of expression level. As explained below, reads can be counted to measure expression levels in each sample and, more importantly, changes in expression level, promoter usage, splicing, and polyadenylation between samples. Transcript (isoform) assembly can lead to discovery of new genes or discovery of new RNA isoforms of known genes. Unlike analog microarray signals, sequence read data are digital and subject to analysis using event-based statistical methods. As for any highly parallel analysis, a subset of the data including key findings should be validated by alternative methods. Once the analysis is validated, gene sets whose expression, splicing, polyadenylation, protein binding or other properties of interest can be tested for association with genomic data and classified with respect to function, sequence motif properties, and other features to extract conclusions and develop refined hypotheses.

EXPERIMENTAL DESIGN

Before applying high-throughput RNA sequencing methods, it is important to understand whether the method is appropriate to the cost and scale of the experiment. The outline below may serve as a guide to these considerations. If the data are to be published or archived, it is important to review the requirements for submission to the Gene Expression Omnibus (Barrett et al. 2009). There is a rapidly growing collection of open-source resources for analyzing high-throughput sequencing (HTS) data (Bateman and Quackenbush 2009). As for other bioinformatics methods, many of these can be downloaded from the web resources of Bioconductor (Gentleman et al. 2004; Reimers and Carey 2006).

Quantity of Starting Material

How much starting material will be needed to obtain sufficient reads? RNA sequencing methods can be used to detect RNAs in very small highly purified pools of RNA, such as those cross-linked to a protein of interest in a “CLIP-Seq” experiment (see CLIP (Cross-Linking and Immunoprecipitation)

Identification of RNAs Bound by a Specific Protein [Darnell 2012]). For transcriptome studies, a few hundred nanograms of oligo(dT)-selected RNA can serve as the starting material (see **Enrichment of Poly(A)⁺ mRNA Using Immobilized Oligo(dT)** [Rio et al. 2010] and **Tips for Preparing mRNA-Seq Libraries from Poly(A)⁺ mRNA for Illumina Transcriptome High-Throughput Sequencing** [Graveley 2013]). Shotgun approaches can proceed with as little as 10 ng, but beware of contaminants such as mycoplasma. Thus, only small amounts of input RNA are required to generate millions of reads. Unlike microarrays, which have specific probes that capture a target sequence with high specificity in the presence of a large excess of other sequences in the sample, RNA sequencing captures sequences mostly with respect to their presence in the sample. A total RNA sample will provide rRNA reads at higher than a 90% rate, leaving some 10% of the reads to be split among the mRNAs and other RNAs in the sample, thus reducing the sensitivity of detection and the accuracy of quantification in particular for lower-abundance transcripts. To improve collection of reads for low-abundance mRNAs, it is advisable to reduce the amount of rRNA in the sample by using oligo(dT) selection (see **Enrichment of Poly(A)⁺ mRNA Using Immobilized Oligo(dT)** [Rio et al. 2010]) or by other selection methods. Despite these and more extreme efforts, including sequential oligo(dT) selections or rRNA depletions, large amounts of rRNA will remain in the starting RNA and reduce the numbers of useful reads.

Library Construction

How will the library preparation method influence the data?

Priming or Fragmentation?

A question that influences the outcome has to do with the method of library preparation and the different biases that this choice can create. Libraries can be prepared by a variety of methods (see **Fragmentation of Whole-Transcriptome RNA Using *E. coli* RNase III** [Ares 2013], **Preparation of Small RNA Libraries for High-Throughput Sequencing** [Malone et al. 2012], and other discussions [e.g., Nagalakshmi et al. 2010], including those provided by commercial kit suppliers). Many transcriptome libraries are prepared from cDNA that is primed with random hexamers (see **Tips for Preparing mRNA-Seq Libraries from Poly(A)⁺ mRNA for Illumina Transcriptome High-Throughput Sequencing** [Graveley 2013]). This can create two kinds of biases in the read distribution. One is a consequence of the fact that cDNA is synthesized in the 5′–3′ direction, leading to higher representation of reads at the 3′ end of the transcript than at the 5′ end (Wilhelm and Landry 2009; Wilhelm et al. 2010). A second bias that has been noted appears to be caused by depletion of some primers in the hexamer mix probably due to noncorrespondence between the representation of hexamers in the synthetic primer mix as compared to the representation of six nucleotide primer-binding sites in natural RNA. A method to address this bias has been developed (Armour et al. 2009).

Rather than priming RNA directly to make cDNA, some library preparation protocols fragment the RNA (by a variety of enzymatic or chemical methods) and then add RNA linkers directly to the ends of the RNA fragments using RNA ligase. Depending on the method used, this may require removal of phosphates from the 3′ end and the addition of phosphate to the 5′ end to achieve ligation after fragmentation. These manipulations can create biases in the library due to variation in efficiency of RNA strand cleavage at different sites and efficiencies of phosphatase, kinase, and RNA ligase for different RNA ends. These biases have not been studied systematically and are likely reduced (except for fragmentation bias) by ensuring that the end-treatment reactions go to completion. Library preparation can be modified to capture transcript ends selectively, to capture reads that represent polyadenylated 3′ ends (Yoon and Brem 2010) or capped 5′ ends (Fejes-Toth et al. 2009). By using structure-sensitive nucleases for fragmentation, it may be possible to map RNA structure in a high-throughput fashion using read ends to identify sites of nuclease sensitivity (J Underwood, pers. comm.).

Amplification

Library preparation requires amplification of an originating set of RNA-derived cDNA fragments by PCR. This can create bias in the library if some fragments amplify more efficiently than others. This

problem may be more severe for longer-fragment libraries. Size selection before amplification and a check that the amplified product remains distributed around the original selected size help to ensure that the population of fragments is efficiently replicated; however, it does not prevent loss or over-amplification of subsets of product. Overamplified products can be identified at the mapping stage by the overrepresentation of fragments with identical ends, and these can be filtered out (Pepke et al. 2009). Unfortunately, lost fragments will not appear in the data, but their existence should remain somewhere in the back of the investigator's mind.

Directionality (Strand Selection)

Because RNA transcription is directional, it is important to know the genomic strand of origin of the RNA sequence that has been captured. A number of strand selection strategies are available. These take advantage either of the specific ligation of 3' blocked linkers to the 3' ends of RNA fragments or of the 5'–3' directionality of reverse transcription to preserve information on the polarity of the original RNA fragment (e.g., see Ingolia et al. 2009; Yoon and Brem 2010).

Paired Ends

One option available under most sequencing protocols is to obtain “paired-end” reads. This requires that the library be constructed with the appropriate linkers and involves capturing read data from individual beads or amplification centers by first using a sequencing primer from one end of each fragment and then clearing and regenerating single-stranded DNA and sequencing a second time with a primer from the other end of the fragment. Thus, two sequence reads, one on either strand, can be obtained from an individual seed molecule in the library. Coupled with an idea of how long the average fragment is in the library (obtained during amplification while preparing the library), paired-end data can greatly improve mapping, especially if one end of the fragment lies in a repeated region or the ends of the fragments lie in two different exons of a spliced RNA. Mapping methods are available for capturing splicing in paired-end RNA sequencing data (Au et al. 2010; Hu et al. 2010; Trapnell et al. 2010). Core facilities will usually offer paired-end sequencing; however, due to the extra reagents needed for the second round of sequencing, this option will be more expensive (although it should not be twice as expensive).

Read Length

In the early days of HTS, read lengths were short, offering perhaps 25–32 high-confidence base calls with which to determine the genomic origin of the sequence. In even moderately repeated regions of the genome, including gene families and dispersed repeats, the short length of reads meant that many could not be mapped to a single region. This is even more complicated than it might first seem, because there is some error rate in the reads that confounds mapping. In addition, there are differences in the “mappability” of genomic sites due to their proximity in sequence space to other genomic sequences, as well as the existence of single-nucleotide polymorphisms in populations relative to the reference genome used for mapping. The bottom line is that all of these problems decrease away rapidly as read length increases; in particular, the identification of splice junction reads improves greatly even in the absence of paired-end data (Au et al. 2010; Hu et al. 2010; Trapnell et al. 2010). Most core facilities will offer a variety of read-length options under different platforms, and these have corresponding cost increases due to the reagents needed for additional sequencing cycles. Thus, a decision with respect to the need for long reads and the cost must be made.

Bar Coding

The bar-coding option allows multiple libraries to be sequenced in the same sequencing lane or run as a mixture. During individual library construction, a distinct linker sequence is attached to each library. These “bar codes” constitute a very few nucleotides that can be read with high confidence early in the read such that the library origin of any individual read is immediately identifiable. This allows cost-

effective sequencing of multiple libraries at once while reducing the depth at which any of the libraries can be read (see below). When multiple samples do not need to be read at great depth, bar coding is an extremely useful method. Most cores offer this service and most of the commercial kits have bar-coded linker sets that can be used in their protocols.

Number of Reads

How many reads are needed to determine transcript structure with confidence? Like genomic tiling arrays, an RNA-Seq experiment can capture evidence for previously unannotated genes and RNA isoforms. If this is a goal, a certain number of reads, but not many sample replicates, will be needed to convince, because a clear qualitative comparison of transcript sequence content to the genome is required rather than quantitative comparison between groups of samples. The utility of RNA sequencing to uncover new transcripts is well documented (Mortazavi et al. 2008; Pepke et al. 2009; Ramskold et al. 2009; Wilhelm and Landry 2009; Hu et al. 2010; Wilhelm et al. 2010). The relationship between read density and transcript discovery has been studied, in particular with respect to alternatively spliced isoforms. Clearly, more data are always better, but there is a relationship between transcript abundance and read density that allows determination of the likelihood that a transcript of a particular abundance will be detected (Li et al. 2008; Sultan et al. 2008; Wang et al. 2008; Yeo et al. 2009; Trapnell et al. 2010). This problem is particularly acute for alternative isoforms of low-abundance mRNAs because, depending on the read length, reads that cross splice junctions are rare.

Replicates

How many replicate libraries will be needed to identify differences in gene expression or RNA processing with statistical confidence? If the goal is to establish whether a change in mRNA levels occurs between biological states (growth conditions, drug treatments, tissues, knockdown, time points), biological sampling will still be necessary. As with any method, the resolution of the experiment to see a change will depend on the sampling error or noise of both the combined technical steps and the underlying biology; with sufficient replicates, the biological variation can be accessed (Pickrell et al. 2010). Simple experiments will involve comparisons between two sets of samples that differ by one variable. In the case of multiple different treatments that must be compared with one another, a common set of control replicates makes the best reference. In either case, as with microarrays, it is important to keep technical variation to a minimum to obtain strong statistical support for biological variation in the system under study. Several laboratories have provided evidence that library preparation and sequencing sets are technically very reproducible (e.g., see Marioni et al. 2008); however, this may depend on the details of the exact method used given the many options for library preparation (see above). Thus, adequate replication of experiment and controls will be needed to show that a result is trustworthy, at a given desired resolution.

It is key to consider and address the sources of variation in the experiment by controlling the variation that can be controlled (technically) and randomizing the variation that cannot be controlled (usually this will be intrinsic to the biology). There are different kinds of “replicate” experiments, some of which capture only part of the variation in an experimental protocol. A so-called “technical replicate” involves taking the very same RNA sample, splitting it into three aliquots, and making three separate fragment libraries. Variation in such an experiment should arise from the fragmentation or priming method, ligation efficiencies, amplification, and other variables in the technical steps of library construction and amplification. Variation in these steps should be held very low if technical replicates are expected to reproducibly if not exactly represent the original samples.

Another set of replicates might involve splitting a cell culture into three aliquots, extracting RNA separately, and then analyzing each RNA sample. These replicates will capture variability in the RNA extraction steps, plus that captured in the previous example. Because RNA extraction should be well controlled, these kinds of replicates are not as useful either. If, on the other hand, separate cell cultures were grown and then each separately extracted and tested, the experiment would detect variability in cell growth, which might be harder to control. Although it is tempting to avoid the variability that

cannot be controlled in experimental design to reduce noise or because it requires more effort, the goal of the statistical analysis should be to reveal changes that occur in response to the experimental manipulation, against the background of uncontrolled variation in cell culture conditions, RNA preparation, labeling, and hybridization. Some factors are easier to control than others (e.g., genetic background, growth medium, cell handling, and sample processing) and still many others are unknown and can only be captured by replication.

Normalization

How will sequencing experiments be normalized so that they can be compared with one another? The admonition to perform replicate library preparation on different biological replicates to obtain reliable measurements of the change in gene expression may require normalization or scaling of libraries with one another. Some libraries are deeper than others. Stochastic effects, especially on rare transcripts, may lead to poor library overlap with respect to recovery of rare sequences. Methods are being evaluated for replicate library comparison and normalization (Robinson and Oshlack 2010). In addition, internal controls can be created such as “spike-in” control RNAs added to the samples in a series of known concentrations for use in evaluating library scale and the efficiency of recovery of different abundance groups. Correction for library depth may also be applied at a postmapping step to create a comparable derived metric such as “RPKM” or “reads per kilobase per million” mapped reads (Mortazavi et al. 2008; Pepke et al. 2009). A sequence derived from a pair of matched-end reads (two paired-end reads) is called a “fragment” and is counted as FPKM or fragments per kilobase per million (Trapnell et al. 2010). The measures (see below) normalize for both gene size (more reads or fragments can be mapped to larger genes) and the total number of reads or fragments (per million mapped). However, (obviously) the RPKM value for a gene from a deep library may have more statistical meaning than an equivalent value from a more shallow library (Bullard et al. 2010).

For this and other aspects of this rapidly evolving area of research, it is advisable to search the Bioconductor site at <http://bioconductor.org/> (Gentleman et al. 2004; Reimers and Carey 2006) and to check a continually updated site associated with recent publications of interest in this area (Bateman and Quackenbush 2009). Once these questions are answered, RNA can be sent to the core facility with detailed and specific instructions. Alternatively, library preparation can proceed, and the fragment libraries can be sequenced. In either case, the next phase is data analysis.

PROCESSING AND MAPPING SEQUENCING READS

Most researchers have sequencing done at a core facility where they send RNA samples for library construction, quality control, and sequencing and receive in return a large amount of data (on the order of several gigabytes delivered either by file transfer or on a portable disk drive), per experiment. Some core facilities will also perform bioinformatics analysis; however, at some point, the biologist will be handed a large amount of data that must be stored, because most core facilities lack the disk space to store data that they produce for very long. At that time, it will be important for the biologist to understand what exactly has been delivered (raw reads? mapped reads? filtered reads?) and how they have been processed and to determine what sort of additional manipulations will be needed to extract the information necessary to draw conclusions from the experiment.

Processing Raw Reads

HTS machines have internal computational methods for determining base identity. These depend in some cases on the method, e.g., in SOLiD sequencing, “color space” data are converted to a sequence. In the end, each platform produces a file that includes a base call at each read position plus a quality score, indicating the confidence that the base call at that position is correct (a so-called raw read). Depending on the platform and the sequence, raw reads have distinct kinds of error profiles that are characteristic of the sequencing method used to obtain them. Raw reads are trimmed so that low-

confidence bases, usually at the end of the read, are removed. If the sequencing run has been derived from a mixture of bar-coded libraries, the next step is to sort the reads by their bar codes into separate files that contain only reads from a given input library. Once this is accomplished, the bar codes and other recognizable linker sequences are trimmed to produce the processed reads. These are labeled by library origin from the bar code, are free of linker sequence, and retain their quality scores.

Mapping

Once the processed reads are obtained, they must be mapped to a reference genome or, if no genome is available, assembled *de novo*. A large number of methods and approaches exist for this step, and many recent reviews have appeared (Mortazavi et al. 2008; Bateman and Quackenbush 2009; Langmead et al. 2009; Pepke et al. 2009; Trapnell et al. 2009; Wilhelm and Landry 2009; Trapnell et al. 2010; Wilhelm et al. 2010). In addition, and because the developments in this field are rapidly being converted to functional and available software packages, the Bioconductor website should be searched for the latest developments (Gentleman et al. 2004; Reimers and Carey 2006). Extra efforts may be necessary to capture and map splice junction reads (Trapnell et al. 2009, 2010; Ameer et al. 2010; Au et al. 2010).

Transcript (Isoform) Assembly

For discovery of novel transcripts, mapped reads must be assembled into transcripts. This problem has also been widely addressed in recent reviews (Bateman and Quackenbush 2009; Pepke et al. 2009). Multiple packages exist for carrying out this process. A commonly used approach uses three programs in sequence. The first is Bowtie (Langmead et al. 2009), which maps reads that do not span junctions and assembles them into “transfrags” that often represent individual exons. Reads that do not map in this initial step are retained and fed to Tophat (Trapnell et al. 2009), which finds and evaluates reads that connect the transfrags produced by Bowtie. Finally, Cufflinks (Trapnell et al. 2010) is used to evaluate the joining solutions created by Tophat to create transcript models that incorporate alternative splicing patterns observed. Other packages can perform similar processing (Reimers and Carey 2006; Bateman and Quackenbush 2009; Pepke et al. 2009).

MEASURING EXPRESSION LEVELS AND CHANGES IN EXPRESSION

As for the mapping problem, methods for determining expression levels are also evolving (Reimers and Carey 2006; Bateman and Quackenbush 2009; Pepke et al. 2009). A popular metric is RPKM, which allows a provisionally comparable number to be extracted from different libraries in which the mapping depth varies, adjusted for genes of different sizes (Mortazavi et al. 2008; Pepke et al. 2009; Trapnell et al. 2010). Although this measure is convenient, it has not yet received exhaustive scrutiny, and there is a suggestion that it could be attached to statistical quality metrics to become more reliable (Bullard et al. 2010; Robinson and Oshlack 2010; Young et al. 2010), especially if it is to be used to detect and measure changes in gene expression. After expression is detected and compared, a list of gene expression or RNA-processing changes by gene or genomic location, along with the magnitude of each change and some sort of measure of the likelihood that the change is not due to chance, is obtained. These represent the statistically valid set of changes that might be due to the experimental manipulation or other feature that distinguishes the groups biologically. It now becomes important to determine how these data stand up to an orthogonal method of measurement.

VALIDATION

For any high-throughput data acquisition method, a result for any individual gene of special interest should be validated by a direct experiment. There are two philosophical reasons for performing

validation experiments. Validation is useful to biologists to find out if a favorite gene that came up in the list is truly changed in the experiment, because their goal is more likely to be the discovery of biological insights into a specific process of interest. On the other hand, bioinformaticists may need to know how many of the results are true and to understand where in the ranked list of best-scoring genes or events the predictions become incorrect to identify the scores that provide the best receiver operator characteristics curve. This will be important because bioinformaticists will require a large high-confidence set of genes for further analysis. Thus, a compromise validation list must be developed for maximum progress and happiness of both collaborators. Once settled, quantitative RT-PCR, semiquantitative RT-PCR, or other independent measurement methods should be used to determine how well the sequencing data predict the true changes as measured by the more labor-intensive approach. A number of studies have been done that indicate good agreement between RNA sequencing and microarrays (e.g., see Marioni et al. 2008), suggesting they can be used to cross-validate each other.

CLASSIFICATION OF CHANGES AND THEIR ASSOCIATED BIOLOGY

After validation, a high-confidence list of gene expression changes or RNA-processing changes is now in hand. It is time to search for biological function and genetic sequence associations (to name two common examples) that appear to be enriched in the high-confidence group relative to a randomly selected comparison groups (“background sets”), either the genome in its entirety or the genes expressed in the cell type under study, or some other group of genes selected under a different experiment. In some cases, the list of genes may provide obvious clues to the biology. This will depend on the mind-set and the experience of the investigator looking at the list. It is advisable to ask the question in a more open way using GO tools (Gene Ontology Consortium 2006) available at Bioconductor. These tools have data that link gene names to terms for cellular components, biological processes, and molecular functions that have been assigned to gene products by various methods. GO has a structured language and seeks to formalize the meaning of terms and connections. In a statistical fashion, the list of genes obtained in the experiment is used as the input, and enrichment of GO terms associated with that group of genes is evaluated compared with the distribution of GO terms among the genes in the background set. If the biological treatment in the experiment impacts expression of a functionally distinct group of genes in the cell, terms associated with this function should appear to be enriched. Recently, it has been discovered that gene lists from RNA sequencing data sets may suffer from a selection bias that overrepresents long and highly expressed genes (Young et al. 2010). It may be important to take this into account.

Sequences in promoters, near splice sites, or within untranslated mRNA regions may also be enriched in the selected gene set. Finding sequence motifs is more complex, and there are numerous approaches to this question. A good starting point is the Genomic Regions Enrichment of Annotations Tool (“GREAT”; McLean et al. 2010), which, in addition to identifying GO term enrichment, allows identification of conserved promoter elements associated with an input gene list. These and other downstream analysis methods aid in hypothesis development for further experimentation. In the end, the sequencing data are just the beginning, and it is likely that many follow-up experiments will be necessary to close the loop and obtain a deeper understanding of the biological impact of the original experiment.

VISUALIZATION: PRESENTING THE DATA

Data that arise from genomic-level experiments usually exist in large tables that are not easy for humans to grasp and digest. A large variety of visualization tools are available at Bioconductor (<http://bioconductor.org/>). Also useful is the UCSC Genome Browser (<http://genome.ucsc.edu/>);

see also Mangan et al. (2008) and Zweig et al. (2008). The Galaxy website (<http://galaxyproject.org/>) is very useful to the biologist who may have limited experience writing scripts or ensuring file format integrity (Blankenberg et al. 2010). This site is especially useful for creating track files for the UCSC Genome Browser. Finally, the display of results in color figures can create concern, especially considering the frequency of red–green color blindness in the human population. An excellent discussion of this point may be found at <http://jfly.iam.u-tokyo.ac.jp/color/>.

FINAL CONSIDERATIONS

Quality control must be maintained at every step from RNA isolation through statistical evaluation, preferably by a small number of investigators who apply consistent standards. If RNA quality is variable or poor, little in the subsequent analysis will be able to improve things. Likewise, errors in analysis can crop up, so spot validation and visual inspection of results or other more elaborate “sanity checks” may be needed to confirm that the analysis is proceeding well. Also, problems may arise as a consequence of “interdisciplinary strain,” in which collaborators struggle to communicate their needs and perspectives on the work across the gulf that separates them. In such cases, talking is best, with the recognition that goals, biases about what is important and reward systems differ in different disciplines and that it is best if all of these can be recognized and respected as valid concerns.

REFERENCES

- Ameur A, Wetterbom A, Feuk L, Gyllenstein U. 2010. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol* 11: R34.
- Ares M. 2013. Fragmentation of whole-transcriptome RNA using *E. coli* RNase III. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot074369.
- Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, et al. 2009. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* 6: 647–649.
- Au KF, Jiang H, Lin L, Xing Y, Wong WH. 2010. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 38: 4570–4578.
- Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. 2009. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res* (database issue) 37: D885–D890.
- Bateman A, Quackenbush J. 2009. Bioinformatics for next generation sequencing. *Bioinformatics* 25: 429.
- Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: A web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Unit 19.10.1–21.
- Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.
- Darnell R. 2012. CLIP (Cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot072132.
- Ehn M, Nourizad N, Bergstrom K, Ahmadian A, Nyren P, Lundeberg J, Hober S. 2004. Toward pyrosequencing on surface-attached genetic material by use of DNA-binding luciferase fusion proteins. *Anal Biochem* 329: 11–20.
- Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon G, Kapranov P, Foissac S, Willingham A, Duttagupta R, Dumais E, et al. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* 457: 1028–1032.
- Gene Ontology Consortium. 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* (database issue) 34: D322–D326.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
- Graveley BR. 2013. Tips for preparing mRNA-Seq libraries from poly(A)+ mRNA for Illumina transcriptome high-throughput sequencing. *Cold Spring Harb Protoc* doi: 10.1101/pdb.ip073908.
- Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano M, Meng Q, Cao H, Li X, et al. 2008. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci* 105: 9145–9150.
- Hu Y, Wang K, He X, Chiang DY, Prins JF, Liu J. 2010. A probabilistic framework for aligning paired-end RNA-seq data. *Bioinformatics* 26: 1950–1957.
- Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218–223.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.
- Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. 2008. Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci* 105: 20179–20184.
- Malone C, Brennecke J, Czech B, Aravin A, Hannon GJ. 2012. Preparation of small RNA libraries for high-throughput sequencing. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot071431.
- Mangan ME, Williams JM, Lathe SM, Karolchik D, Lathe WC III. 2008. UCSC genome browser: Deep support for molecular biomedical research. *Biotechnol Annu Rev* 14: 63–108.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18: 1509–1517.
- McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28: 495–501.
- Mitra RD, Shendure J, Olejnik J, Edyta-Krzyszanska O, Church GM. 2003. Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem* 320: 55–65.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621–628.
- Nagalakshmi U, Waern K, Snyder M. 2010. RNA-Seq: A method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol* Unit 4.11.1–4.11.13.

M. Ares Jr.

- Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 6: S22–S32.
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768–772.
- Ramskold D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 5: e1000598.
- Reimers M, Carey VJ. 2006. Bioconductor: An open source framework for bioinformatics and computational biology. *Methods Enzymol* 411: 119–134.
- Rio DC, Ares M, Hannon GJ, Nilsen TW. 2010. Enrichment of poly(A)⁺ mRNA using immobilized oligo(dT). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5454.
- Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11: R25.
- Russom A, Tooke N, Andersson H, Stemme G. 2005. Pyrosequencing in a microfluidic flow-through device. *Anal Chem* 77: 7505–7511.
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728–1732.
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956–960.
- Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28: 511–515.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476.
- Wilhelm BT, Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48: 249–257.
- Wilhelm BT, Marguerat S, Goodhead I, Bahler J. 2010. Defining transcribed regions using RNA-seq. *Nat Protoc* 5: 255–266.
- Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* 16: 130–137.
- Yoon OK, Brem RB. 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* 16: 1256–1267.
- Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol* 11: R14.
- Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. 2008. UCSC genome browser tutorial. *Genomics* 92: 75–84.

Cold Spring Harbor Protocols

Methods for Processing High-Throughput RNA Sequencing Data

Manuel Ares Jr.

Cold Spring Harb Protoc; doi: 10.1101/pdb.top083352

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

- [Alignment of Sequences](#) (46 articles)
- [Alignment of Sequences, general](#) (24 articles)
- [Analysis of Gene Expression](#) (197 articles)
- [Analysis of Gene Expression, general](#) (129 articles)
- [Bioinformatics/Genomics, general](#) (192 articles)
- [cDNA Libraries](#) (26 articles)
- [Libraries](#) (147 articles)
- [mRNA](#) (132 articles)
- [RNA](#) (317 articles)
- [RNA, general](#) (269 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>