13

RNA Sequencing with Next-Generation Sequencing

Stuart M. Brown and Jeremy Goecks

Sequencing of RNA has been an important application of DNA sequencing technology since its invention. RNA is usually sequenced by first converting it to complementary DNA (cDNA) with the reverse transcriptase enzyme (RNA-dependent DNA polymerase). Reverse transcriptase was originally isolated from Rous sarcoma retrovirus and Rauscher mouse leukemia retrovirus (R-MLV) by Baltimore (1970) and independently by Temin (1970). In 1972, Verma et al. and Bank et al. developed efficient systems to copy messenger RNA (mRNA) to cDNA by adding DNA nucleotide triphosphates and and short pieces of oligo(dT), which hybridize to the poly(A) tail of the mRNAs and act as a primer.

cDNA is frequently the subject of sequencing studies, because this an efficient method to discover the coding sequence of expressed genes or for finding gene coding regions in genomic DNA sequence. Craig Venter expanded this method by collecting large numbers of short single reads from the 3′ ends of mRNA, which were called expressed sequence tags (ESTs). Early EST sequencing of human cells was extraordinarily productive, resulting in the discovery of many thousands of new genes (Adams et al. 1991, 1992). The EST method allowed for a rough form of gene expression measurements in a variety of cell types and some differential expression studies were conducted in this manner. EST sequencing also became a valuable component of **de novo sequencing** projects, providing a layer of gene expression information and seeding annotation and gene finding efforts.

Microarray technology, developed in the 1990s, measures the hybridization of labeled cDNA to an array of DNA probes that correspond to the sequences of known genes (or ESTs). The microarray method allows for the discovery, in a genome-wide fashion, of gene expression changes (as reflected in changes of mRNA levels) resulting from any biological treatment or condition.

RNA sequencing with NGS technology (**RNA-seq**) can be used for a number of different scientific applications. The NGS reads are mapped to a **reference genome**, then the number of reads mapping within a feature of interest (such as a gene or **exon**), is a measure of expression. Direct sequencing of mRNA provides a measurement of gene expression for the entire transcriptome that is more accurate and has a greater dynamic range than microarray-based technologies (Marioni et al. 2008). Just as in microarray experiments, the most common application of RNA-seq is to identify genes that change expression between experimental conditions. RNA-seq can also be used to detect mutations in transcribed portions of the genome for the native germline cells of an individual or for **somatic** mutations in tumor cells. RNA-seq is also an excellent platform to measure alternative splicing events that produce different transcripts (and ultimately different proteins) from a single gene. Alternative transcript isoforms can be detected with great accuracy by using RNA-seq reads mapping at splice junctions, specifying both known as well as novel isoforms. With appropriate sample preparation methods, RNA-seq can also be used to interrogate a wide variety of non-protein-coding RNAs.

Protocols for sequencing of RNA have been developed by all of the major NGS vendors. **Ribosomal RNA** (**rRNA**) and transfer RNA (tRNA) are very abundant in the total RNA extracted from both prokaryotic and eukaryotic cells (~75% of RNA molecules). Sequencing of abundant non-protein-coding RNA reduces yield and sensitivity of RNA-seq methods for mRNA and increases cost. Most protocols for RNA-seq in eukaryotic cells use poly(T) oligonucleotides to isolate mRNA with poly(A) tails, or use poly(T) primers in combination with random short oligomers for reverse transcription. After poly(A) enrichment, and cDNA synthesis, most protocols shatter cDNA molecules into small fragments (from 100 to 300 bp) that are then ligated with oligomers specific for the sequencing system. Some protocols have also been developed to sequence small non-protein-coding RNA molecules such as micro-RNA (miRNA), small interfering RNA (siRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), Piwi-interacting RNA (piRNA), and others.

Another method of removing rRNA, tRNA, and other highly abundant RNAs before sequencing is called duplex-specific nuclease (DSN) normalization. This method uses a nuclease (Kamchatka crab hepatopancrease) that specifically degrades double-stranded DNA, while leaving single-stranded DNA molecules intact (Zhulidov et al. 2004). This method takes advantage of reassociation kinetics. First, total RNA is reverse transcribed to double-stranded cDNA. Then the cDNA is denatured at high temperature. Under selective annealing conditions, the most abundant cDNA molecules (cDNA clones of rRNA, tRNA, mtRNA, and the most highly transcribed messages) form double strands and are degraded, whereas less abundant molecules remain single stranded and are preserved. Illumina has presented data using DSN normalization for RNA-seq (http://www.illumina.com/documents/seminars/presentations/2010-06_sq_03_lakdawalla_transcriptome_sequencing.pdf) that

indicates very good removal of rRNA, high retention of small noncoding RNAs, and no 3′ bias compared with poly(A) purification methods. Relatively few RNA-seq experiments have been published using the DSN method, so it is not clear what bias it creates in gene expression or differential expression values.

Despite purification methods (poly(A) selection or DSN normalization), RNA-seq data may still contain substantial amounts of rRNA, tRNA, and also mitochondrial RNA. These can be filtered out in the bioinformatics pipeline by simply using a "contaminant" file of rRNA, tRNA, and mtDNA sequences for the target species to prefilter all **sequence reads** (by **alignment** with a **short read** aligner such as BWA or Bowtie) before mapping the remaining reads to the genome and/or splice junction database.

DEPTH OF COVERAGE AND NUMBER OF REPLICATES

To accurately measure changes in gene expression for a specific gene between two experimental conditions (differential expression), the number of mRNA reads sequenced from the transcript of that gene must be above the threshold of detection in each sample, or at least in samples from conditions in which that gene is expressed. In addition, given the biological and technical variability of counting mRNA molecules, the absolute counts per sample must be large enough to allow for an accurate variance measurement across several replicate samples for each experimental condition. As the total amount of sequence reads per sample increases, expression levels for each gene can be estimated more accurately and statistical power to detect differential expression (DE) increases. Somewhat counterintuitively, Tarazona et al. (2011) suggest that as the number of reads per sample increases, the number of false positives for DE calls increases for many statistical methods.

Although the goal of most RNA-seq experiments is to accurately profile the expression levels of all genes, different cell types express genes at dramatically different levels, creating unique transcriptome profiles. For any given cell type, some genes may be expressed at very high levels, perhaps as much as 5% or 10% of the total mRNA, whereas some genes are not expressed at detectable levels (i.e., less than one mRNA molecule per cell). Therefore, it is probably not possible to sequence enough reads per sample to accurately assess DE for every single gene in the genome. Blencowe et al. (2009) suggested that 700 million reads per sample were required to obtain accurate quantitation for 95% of all expressed transcripts in mammalian cells. These very low-expressed transcripts may not be good targets for DE analysis because the variance of read counts may be quite high across replicates.

Given that there is some limit on the sensitivity of RNA-seq to detect transcripts and accurately assess differential gene expression, the relevant question for most investigators becomes: What is the practical limit of sensitivity in which reliable

expression information can be obtained for the majority of transcriptionally active genes? In fact, a steep curve of diminishing returns has been observed in a number of studies that have explored the depth of **coverage** for RNA-seq. In Figure 1, data from three studies all show a consistent pattern of decreased discovery of new genes (covered by at least five reads) at increasing depth of coverage. Marioni et al. (2008) discover 232 new genes per each additional million reads at a depth of 22 million, the MAQC study finds 70 new genes per million at a depth of 45 million, and Griffith et al. (2010) find 19 new genes per million at a depth of 200 million (and by subsampling the Griffith data, the discovery rate at a depth of 20 million is 210 and at a depth of 45 million is 75). Data from Toung et al. (2011) (Fig. 2) is also consistent, with very few new transcripts discovered as sequencing depth increases from 100 million reads to nearly one billion reads.

Of course, the actual number of expressed genes and their relative abundance within the cell varies depending on organism, tissue type, cell type, and developmental status. Low abundance transcripts may be biologically important regulators, but deep coverage sequencing of in vivo samples may also capture transcripts from nontarget cell types mixed into the sample. Very deep sequencing studies also observe rare noncoding transcripts from much of the genome. In fact, the percentage of noncoding transcripts increases among newly detected genes at deeper levels of coverage, whereas discovery of new protein-coding genes reaches near saturation at much lower coverage levels (Tarazona et al. 2011).

Another important consideration for the design of an RNA-seq experiment is the number of replicates for each biological condition. Researchers generally wish to know the optimal number of replicates required to achieve a desired level of statistical power to find DE. Li et al. (2013) developed a model to calculate statistical power and estimate sample size for RNA-seq experiments based on a negative binomial model of variation in counts per gene in each sample and an exact test for DE. They show sample size requirements in a simulation experiment and by reanalysis of published data for two experiments with human tissues. In the simulation, in which variance among replicates was low ($\phi^* = 0.1$) and $\log_2$-fold change was 2.0 or more, only three to six replicates are required to find all of the DE genes with coverage greater than five reads and a false discovery rate (FDR) less than 5%. Increasing the variance to $\phi^* = 0.5$ triples the number of required replicates, and lowering the $\log_2$-fold change to 1.0 (a twofold change in expression) increases the required number of replicates to 20. In real biological data the variance often exceeds 0.6 and is overdispersed compared to the expectation of a Poisson model (Fang et al. 2012). For example, in a data set extracted from Blekhman et al. (2010), RNA-seq of liver samples are compared between three human males and three females. Average read coverage of 13,267 detected genes is 1.6 and the dispersion is $\phi^* = 0.6513$. In this data set, to discover 80% of twofold DE genes with FDR of 10% would require a sample size of 79 per condition. (See Table 1.)

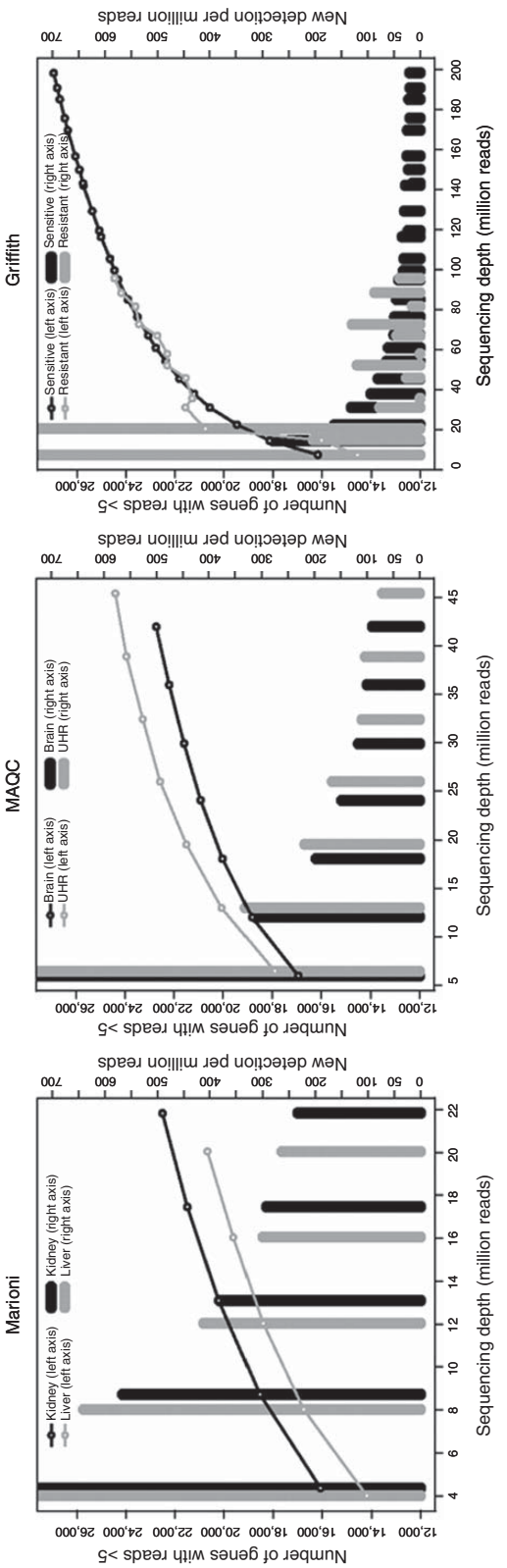

FIGURE 1. Comparison of three published studies of transcript discovery versus RNA-seq sampling depth by Marioni et al. (2008), Shi et al. (2006), and Griffith et al. (2010). (Reprinted, with permission, from Tarazona et al. 2011, © Cold Spring Harbor Laboratory Press.)

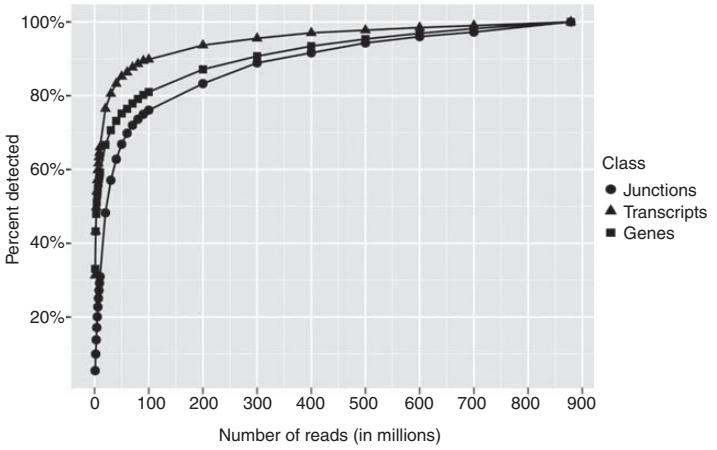

FIGURE 2. 100 million reads detect 81% of genes at FPKM $\geq$ 0.05. Each additional 100 million reads detect ~3% more genes. (Reprinted, with permission, from Toung et al. 2011, ©Cold Spring Harbor Laboratory Press.)

The general conclusion from these sample size calculations is that the majority of published RNA-seq experiments have very low power—perhaps only 50% of true DE genes can be discovered with adequate control of false discovery; and the discovery of these DE genes is biased toward genes that are highly expressed and show large fold change between conditions. Combining considerations of sequencing depth and

TABLE 1. Sample size simulation for an RNA-seq experiment with 10,000 detected genes with log-fold change [$\log_2(p^*)$] ranging from 0.5 to 2.5 and dispersion (ϕ^*) at 0.1 and 0.5

		$\mu_0^* = 1$ FDR			$\mu_0^* = 5$ FDR		
$\log_2(p^*)$	ϕ^*	1%	5%	10%	1%	5%	10%
0.5	0.1	365 (81)	305 (84)	278 (88)	104 (81)	87 (84)	79 (88)
	0.5	518 (81)	433 (84)	394 (88)	257 (81)	215 (84)	196 (89)
1.0	0.1	79 (81)	67 (84)	61 (87)	24 (82)	20 (84)	19 (91)
	0.5	119 (81)	99 (83)	91 (88)	63 (82)	53 (85)	48 (89)
1.5	0.1	31 (82)	26 (83)	24 (86)	10 (83)	9 (90)	8 (91)
	0.5	49 (81)	41 (83)	38 (88)	28 (83)	23 (84)	21 (86)
2.0	0.1	16 (85)	13 (84)	12 (86)	6 (90)	5 (92)	4 (86)
	0.5	26 (82)	22 (84)	20 (86)	16 (84)	13 (85)	12 (89)
2.5	0.1	8 (85)	7 (89)	6 (87)	3 (78)	3 (81)	3 (98)
	0.5	14 (83)	12 (87)	11 (84)	10 (82)	9 (90)	8 (91)

Reproduced, with permission, from Li et al. 2013.

The sample size is shown for minimum normalized read counts per gene (μ_0^*) of 1 and 5 and FDR rates of 1%, 5%, and 10%. Numbers in parentheses after sample size are the number of differential genes detected by the exact text using edgeR (in which the true number of DE genes is 80).

sample size leads to a general recommendation that has been repeated by a number of investigators. More biologically relevant DE genes will be discovered by sequencing more samples at lower depth of coverage rather than fewer samples at greater depth. Hart et al. (2013) surveyed 127 RNA-seq experiments and found that a sequencing depth of 10 million reads will ensure that ~90% of all annotated genes will be covered by at least 10 reads, and that no greater detection of DE at twofold expression change can be achieved with greater depth of sequencing. The larger constraint on detecting DE for a gene is the variance of the expression measurements across replicates rather than the depth of coverage (see Fig. 3). Rapaport et al. (2013) summarize the extensive RNA-seq DE benchmarking efforts of the MAQC/SEQC group with the simple statement: "Our results demonstrate that increasing the number of replicate samples significantly improves detection power over increased sequencing depth."

The abundance of transcripts from different genes observed in RNA-seq data has been shown to accurately represent the gene expression profile of various cell samples when validated by other technologies such as RNA microarray and quantitative polymerase chain reaction (qPCR) (Maroni et al. 2008). As the total yield of NGS machines has increased, the sensitivity of RNA-seq has greatly exceeded microarray-based methods of measuring transcripts from genes expressed at low levels. Because RNA-seq does not rely on existing sequence data for the creation of probes, it can measure the expression of unannotated genes and portions of known genes not previously observed in transcripts such as 5′ and 3′ extensions as well as a variety of alternatively spliced isoforms that include regions annotated as **introns**. Pickrell et al. (2010) found that ~15% of mapped human RNA-seq reads were located outside annotated exons. Figure 4 illustrates the RNA-seq mapping around the ADM

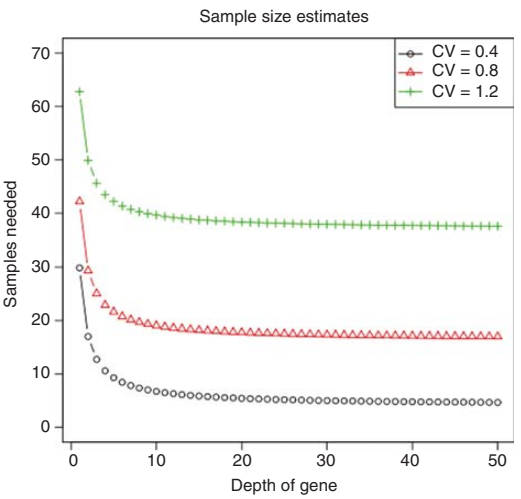

FIGURE 3. The coefficient of variation has a much larger effect than the depth of sequencing on the required sample size to detect a twofold difference in expression of a single gene with 80% power at $\alpha = 0.01$. (Reprinted, with permission, from Hart et al. 2013. The publisher for this copyrighted material is Mary Ann Liebert, Inc. Publishing.)

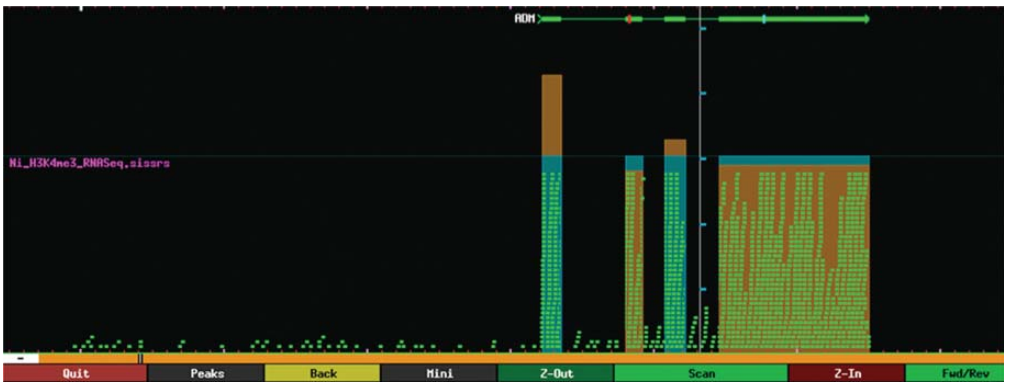

FIGURE 4. RNA-seq reads mapped to the human genome in the region of the ADM gene. (Image courtesy of P.R. Smith.)

gene, with large numbers of reads mapping to annotated exons, but some reads also map to introns and 5′ regions.

RNA-seq experiments may have goals other than quantifying gene expression and detecting expression changes across experimental conditions. Interrogation of alternative splicing requires adequate coverage of all potential splice junction sites on a transcript (an average coverage of five reads per base across every base in the entire length of the transcript), and discovery of low abundance transcript isoforms may require much deeper coverage. Discovery of **sequence variants** (single-nucleotide polymorphism [SNPs]) in RNA requires an average depth of coverage greater than 10× for every base in each expressed gene. The actual number of reads required depends of course on the size of the transcriptome for the target species.