Applications of New Sequencing Technologies for Transcriptome Analysis

Olena Morozova, Martin Hirst, and Marco A. Marra

BC Cancer Agency, Genome Sciences Center, Vancouver, BC V5Z 4S6, Canada;
email: omorozova@bcgsc.ca, mhirst@bcgsc.ca, mmarra@bcgsc.ca

Key Words
next-generation sequencing technology, Illumina/Solexa, 454/Roche, ABI SOLiD, Helicos HeliScope, RNA sequencing

Abstract
Transcriptome analysis has been a key area of biological inquiry for decades. Over the years, research in the field has progressed from candidate gene-based detection of RNAs using Northern blotting to high-throughput expression profiling driven by the advent of microarrays. Next-generation sequencing technologies have revolutionized transcriptomics by providing opportunities for multidimensional examinations of cellular transcriptomes in which high-throughput expression data are obtained at a single-base resolution.
**TRANSCRIPTOME ANALYSIS: A HISTORICAL PERSPECTIVE**

An intriguing enigma in molecular biology is how the identical genetic make-up of cells can give rise to different cell types, each of which plays a defined role in the functioning of a multicellular organism. This phenotypic diversity has been linked to the fact that different cell types within the organism activate (or express) different sets of genes (transcriptomes) that lead to different cell fates and functions. The correlation of cellular fate and function with gene expression patterns has thus been of prime interest to biologists for decades (Table 1).

### Candidate Gene Approaches

The earliest attempts to understand cellular transcriptomes included examinations of total cellular RNA from different organisms, tissue types, or disease states for the presence and quantity of transcripts of interest. The first candidate gene-based studies utilized Northern blot analysis (3), a low-throughput technique that required the use of radioactivity and large amounts of input RNA. This procedural complexity and requirement for relatively large amounts of RNA restricted Northern blotting to the detection of a few known transcripts at a time from samples where RNA availability was not limited. The development of reverse transcription quantitative PCR (RT-qPCR) methods (8, 51) facilitated transcript detection, increased the experimental throughput, and reduced the required quantity of input RNA.

### Microarray Technology

The development of microarrays supplanted single-gene approaches by allowing simultaneous characterization of expression levels of thousands of known or putative transcripts (59). This advance brought about a multitude of expression-profiling initiatives aiming to comprehensively characterize expression signatures of different cell types and disease states. Further developments in the microarray field enabled other transcriptomics applications, such as the detection of noncoding RNAs, single-nucleotide polymorphisms (SNPs), and alternative splicing events (43). Due to their cost-efficiency, microarrays are a commonly used tool in transcriptomics research utilized in many laboratories around the world. (For a review on microarray technology the reader is referred to Reference 54.)

Despite their power to measure the expression of thousands of genes simultaneously, microarray methods do not readily address several key aspects, notably the ability to detect novel transcripts and the ability to study the coding sequence of detected transcripts. Moreover, since microarrays are indirect methods in which transcript abundance is inferred from hybridization intensity rather than measured explicitly, the derived data are

### Table 1 Milestones in transcriptome analysis

<table>
<thead>
<tr>
<th>Year</th>
<th>Milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>Sequence of the first RNA molecule determined</td>
</tr>
<tr>
<td>1977</td>
<td>Development of the Northern blot technique and the Sanger sequencing method</td>
</tr>
<tr>
<td>1989</td>
<td>Reports of RT-PCR experiments for transcriptome analysis</td>
</tr>
<tr>
<td>1991</td>
<td>First high-throughput EST sequencing study</td>
</tr>
<tr>
<td>1992</td>
<td>Introduction of Differential Display (DD) for the discovery of differentially expressed genes</td>
</tr>
<tr>
<td>1995</td>
<td>Reports of the microarray and Serial Analysis of Gene Expression (SAGE) methods</td>
</tr>
<tr>
<td>2001</td>
<td>Draft of the Human Genome completed</td>
</tr>
<tr>
<td>2005</td>
<td>First next-generation sequencing technology (454/Roche) introduced to the market</td>
</tr>
<tr>
<td>2006</td>
<td>First transcriptome sequencing studies using a next-generation technology (454/Roche)</td>
</tr>
</tbody>
</table>
noisy, which interferes with reproducibility and cross-sample comparisons.

**Sequencing-Based Approaches to Studying Transcriptomes**

DNA sequencing approaches to transcriptome analysis have been an alternative to microarray-based methods. A key advantage of these approaches over microarray methods is the ability to directly determine the identity and, more recently, the abundance of a transcript rather than inferring it indirectly from measures of hybridization intensity used in Northern blots or microarray experiments.

Transcriptome sequencing studies have evolved from determining the sequence of individual cDNA clones (64) to more comprehensive attempts to construct cDNA sequencing libraries representing portions of the species transcriptome (60). Due to the high cost of the Sanger method (58) used in these studies and the complexity of the associated cloning step, routine full-length cDNA (FLcDNA) sequencing efforts were not feasible, resulting in low coverage, insufficient to comprehensively characterize whole transcriptomes of multicellular species. As a consequence, Sanger FLcDNA sequencing has primarily been applied to novel transcript discovery and annotation (e.g., 60).

The development of expressed sequence tag (EST) sequencing in 1991 partially addressed the cost limitation of FLcDNA sequencing by introducing a less complete, less accurate, yet cheaper approach to the detection of expressed transcripts than was possible with sequencing FLcDNAs (11). Despite the decrease in cost, however, EST sequencing with the Sanger method was still too expensive and labor intensive to be routinely used on a transcriptome-wide scale. Moreover, due to the low redundancy of sequencing reads, EST data were not suitable for estimating transcript abundance.

The report of Serial Analysis of Gene Expression (SAGE) provided a key advance in transcriptome sequencing as it facilitated the use of Sanger sequencing for gene expression profiling (72). SAGE experiments offered many advantages over microarrays, such as the ability to detect novel transcripts, the ability to obtain direct measures of transcript abundance thus allowing easier comparisons between multiple samples, and the discovery of novel alternative splice isoforms. However, SAGE studies still involved a laborious cloning procedure, were costly, and produced short sequence tags (14 or 21 bp) that are difficult to resolve for transcripts with similar coding sequence (Figure 1).

**New-Generation Sequencing Methods**

With the completion of reference genome sequencing projects for human and the major model organisms of biomedical significance, resequencing applications have come to the forefront. These have been driven by a panel of conceptually new sequencing technologies collectively referred to as “next-generation sequencers” that are more cost-effective than Sanger sequencing. The four commercially available new-generation sequencing technologies, Roche/454, Illumina, Applied Biosystems SOLiD, and most recently released Helicos HeliScope, produce an abundance of short reads at a much higher throughput than is achievable with the state-of-the-art Sanger sequencer (Table 2). Another new sequencing technology, currently being developed by...
Table 2 Commercially available sequencing technologies used for transcriptome sequencing applications

<table>
<thead>
<tr>
<th>Sequencing platform</th>
<th>ABI3730xl</th>
<th>Genome Analyzer</th>
<th>Illumina Genome Analyzer</th>
<th>ABI SOLiD</th>
<th>HeliScope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing chemistry</td>
<td>Automated Sanger sequencing</td>
<td>Pyrosequencing on solid support</td>
<td>Sequencing-by-synthesis with reversible terminators</td>
<td>Sequencing by ligation</td>
<td>Sequencing-by-synthesis with virtual terminators</td>
</tr>
<tr>
<td>Template amplification method</td>
<td>In vivo amplification via cloning</td>
<td>Emulsion PCR</td>
<td>Bridge PCR</td>
<td>Emulsion PCR</td>
<td>None (single molecule)</td>
</tr>
<tr>
<td>Read length</td>
<td>700–900 bp</td>
<td>200–300 bp</td>
<td>32–40 bp</td>
<td>35 bp</td>
<td>25–35 bp</td>
</tr>
<tr>
<td>Sequencing throughput</td>
<td>0.03–0.07 Mb/h</td>
<td>13 Mb/h</td>
<td>25 Mb/h</td>
<td>21–28 Mb/h</td>
<td>83 Mb/h</td>
</tr>
</tbody>
</table>

1Data tabulated on September 15, 2008.

Expressed sequence tag (EST): a single-pass sequencing read from the 3’ or 5’ end of a cDNA clone. In contrast, full-length cDNA (FLcDNA) sequencing involves generation and assembly of sequencing reads spanning the full length of cDNA clones.

Serial Analysis of Gene Expression (SAGE): the first sequencing-based method for high-throughput gene expression profiling. SAGE involves the generation of short sequence tags from 3’ ends of mRNA transcripts. The tags are then concatenated, sequenced, and counted providing estimates of transcript abundance.

Pacific Biosciences, has the potential to take breakthroughs in DNA sequencing even further, by enabling observation of natural DNA synthesis by a DNA polymerase as it occurs in real time. This instrument is not commercially available at the time of this writing, and hence is not discussed here. The advent of next-generation sequencing technologies has tremendously reduced the sequencing cost and experimental complexity, as well as improved transcript coverage, rendering sequencing-based transcriptome analysis more readily available and useful to individual laboratories (Figure 1). This technological advance challenged the dominant nature of microarrays, enabling many new applications to be introduced for the study of transcriptomes. These technological advances and applications are discussed below following an account of the advances in sequencing technologies. (For other reviews of next-generation sequencing technologies, see References 32, 38, 39, 46, 61.)

EVOLUTION OF DNA SEQUENCING TECHNOLOGIES

A standard DNA sequencing workflow has traditionally included three key steps, sample preparation, sequencing, and data analysis. The new sequencing technologies improve upon the Sanger protocol by advances in the first steps of the workflow, albeit often at the cost of higher error rates and shorter read lengths that can challenge data analysis.

Advances in Sample Preparation

In Sanger sequencing, a DNA sample is first sheared into fragments, then subcloned into vectors, and amplified in bacterial or yeast hosts. The amplified DNA is then isolated and sequenced with the Sanger chain termination method. Cloning-based amplification is prone to host-related biases, and is lengthy and labor intensive, restricting high-throughput Sanger sequencing to genome sequencing centers where elaborate multistep pipelines are available to automate the process. A major advantage of the second-generation sequencing platforms (e.g., 454/Roche, Illumina, and SOLiD) is elimination of the in vivo cloning step and its replacement with PCR-based amplification. Both 454/Roche (40) and Applied Biosystems SOLiD technologies circumvent the cloning requirement by taking advantage of emulsion PCR (67), which uses emulsion droplets to isolate single DNA templates in separate micro reactors where amplification is
carried out. The Illumina platform (9, 10) uses bridge amplification, a solid phase amplification approach in which DNA molecules are attached to a solid surface and amplified in situ, generating clusters of identical DNA molecules. Both of these amplification approaches result in the generation of a collection of clonal copies of the template, which are fed into subsequent steps of the sequencing pipelines. A true single-molecule method, developed by Stephen Quake’s laboratory (and recently commercialized by Helicos Biosciences), eliminates the amplification step, directly sequencing single DNA molecules bound to a surface (12). Such single-molecule sequencing approaches are referred to as third-generation technologies and have the potential to reduce sequencing costs even more steeply than second-generation instruments.

Advances in Sequencing Chemistry and Detection

The paradigm of the original Sanger method is the DNA polymerase-dependent synthesis of a complementary strand in the presence of four labeled nonreversible synthesis terminators, 2′,3′-dideoxynucleotides (ddNTPs) corresponding to the four natural 2′-deoxynucleotides (dNTPs). The four terminators are incorporated into the growing DNA strand at random in place of the corresponding dNTP, thereby producing a collection of DNA fragments of varying lengths that are then separated by polyacrylamide gel electrophoresis (58). Originally, radioactively labeled ddNTPs were used and four different reactions were required per one template molecule. Subsequently, the radioactively labeled ddNTPs were replaced with fluorescently labeled terminators that allowed the four sequencing reactions to be carried out simultaneously with different ddNTPs distinguishable by color (63). Other improvements included the replacement of slab gel electrophoresis with capillaries, the advent of capillary arrays that allowed sample multiplexing, and the deployment of production-scale sequencing workflows. As a result of these developments, the Sanger method achieved the read length, accuracy, and throughput compatible with de novo sequencing of whole genomes. To date, Sanger sequencing has been exclusively responsible for the generation of reference genome sequences of many species including that of human (35, 73).

The pyrosequencing approach was the first alternative to Sanger sequencing to achieve commercialization as part of the Roche/454 instrument (40). Pyrosequencing uses chemiluminescence-based detection of each released pyrophosphate that occurs upon the incorporation of a nucleotide by the DNA polymerase (Figure 2). The four nucleotides are added to the sequencing reaction one at a time, and the addition of the correct nucleotide is accompanied by the release of light. The amount of light produced is proportional to the number of incorporated nucleotides, allowing for the detection of homopolymers (up to the point of detection saturation). About 1.6 million pyrosequencing reactions occur in parallel, each in a separate well of a picotiter plate contributing to a much higher sequencing throughput than that achieved in a 96-well capillary array of a modern Sanger sequencer.

Similarly to 454/Roche, the Illumina Genome Analyzer also uses sequencing-by-synthesis, albeit with a different detection chemistry (10). The Illumina sequencing reaction utilizes four fluorescently labeled nucleotide analogs that serve as reversible sequencing terminators, and special DNA polymerases that are capable of incorporating these analogs into the growing oligonucleotide chain (Figure 2). At each step the correct nucleotide analog is incorporated into the growing chain and its identity is revealed by the color of its fluorescent label. Importantly, the 3′-OH group of the nucleotide is blocked to prevent further incorporation. After the imaging step, the label is washed off and the blockage is reversed, thereby allowing the synthesis to proceed. The sequencing reactions occur in a massively parallel fashion on a flow cell, a glass surface that contains tens of millions of clusters of clonally identical DNA molecules.

Gene expression profiling: the simultaneous assessment of the expression level of a large number of genes, often an entire genome, to obtain a global snapshot of the complete mRNA component of the cell at a given time; helps to distinguish between different cell types, different disease states, and different time points during development

Resequencing: sequencing of the genome or transcriptome of an individual of a species for which a reference genome sequence is available. In contrast, sequencing and assembly of the reference genome itself is termed de novo sequencing. Resequencing is commonly conducted to gauge sequence diversity within the species

Next-generation sequencers: second- and third-generation sequencing platforms. Second-generation sequencers from 454/Roche, Illumina, and Applied Biosystems sequence PCR amplified “clusters” of single-molecule templates. Third-generation sequencers from Helicos and Pacific Biosciences sequence single-molecule templates directly with no PCR amplification
Advances in sequencing chemistry implemented in next-generation sequencers. (a) The pyrosequencing approach implemented in 454/Roche sequencing technology detects incorporated nucleotides by chemiluminescence resulting from PPi release. (b) The Illumina method utilizes sequencing-by-synthesis in the presence of fluorescently labeled nucleotide analogues that serve as reversible reaction terminators. (c) The single-molecule sequencing-by-synthesis approach detects template extension using Cy3 and Cy5 labels attached to the sequencing primer and the incoming nucleotides, respectively. (d) The SOLiD method sequences templates by sequential ligation of labeled degenerate probes. Two-base encoding implemented in the SOLiD instrument allows for probing each nucleotide position twice.

The true single-molecule sequencing approach commercialized by Helicos Biosciences in the HeliScope instrument also uses a synthesis-by-synthesis procedure in which virtual terminators (nucleotide analogs that reduce the processivity of DNA polymerase) are used (42). The reduced DNA polymerase processivity allows for the accurate identification of homopolymer stretches. In the Helicos system, single-molecule DNA templates are captured on the flow cell surface; Cy3-labels attached at both ends of each DNA molecule are used to reveal the location of each template bound to immobilized primers on the surface of the flow cell. The Cy5-labeled nucleotides are added to the reaction one at a time, and the detection of incorporated nucleotides is achieved by TIRF (total internal reflection fluorescence) (Figure 2). After the addition of each nucleotide, the fluorescent labels are cleaved and the synthesis continues.

In contrast to the polymerase-based approaches discussed above, the SOLiD (Supported Oligonucleotide Ligation and Detection System) system uses a sequencing-by-ligation approach in which the sequence is inferred indirectly via successive rounds of hybridization and ligation events. This approach was first published by the Church laboratory as the “polony sequencing technique” (62). The SOLiD system uses 16 dinucleotides, each carrying a fluorescent label. Four fluorescent dyes are used in the system such that one dye labels four different dinucleotides (Figure 2). The identity
of each base is determined from the fluorescent readout of two successive ligation reactions. An advantage of the two-base encoding scheme is that each position is effectively probed twice, in principle allowing for the distinction of sequencing error from a true sequence polymorphism.

APPLICATIONS OF NEW-GENERATION TRANSCRIPTOME SEQUENCING

Protein-Coding Gene Annotation

Despite the availability of complete genome sequences from humans and other organisms, much of these genomic data are not fully or even well understood (13). A complete genome annotation would require knowledge of all transcription start and polyadenylation sites, exon-intron structures, splice variants, and regulatory sequences. Despite recent advances, complete annotation information is not available for the majority of metazoan genes (13). Sanger-based transcriptome sequencing in the form of ESTs or FLcDNAs has provided an accurate and effective means for annotating many of the more abundant protein-coding genes (1, 31, 60). However, the limitations of the Sanger sequencing method restrict the utility of these approaches to the annotation of most abundantly expressed genes. For instance, it has been estimated that most EST studies using Sanger sequencing detect only about 60% of transcripts in the cell and thus do not provide a complete representation of the transcriptome (13). This information gap can be addressed using the next-generation sequencing technologies. For instance, a single run on the 454 machine is capable of generating 400,000 ESTs (6) compared to 720 ESTs generated by Sanger sequencing in earlier studies (41).

In genome annotation studies, ESTs are aligned to reference genome sequences, thus revealing the presence of exons, introns, exon junctions, and transcription boundaries for the captured genes (Figure 3). The transcriptome sequences can be aligned to the genome of either the same species (cis alignment) or a related species (trans alignment) if a reference genome sequence is not available. To date, next-generation sequencing technologies have been used to generate EST libraries for many model organism species and human tissues (see Reference 45 for a review).

EST sequencing is particularly fruitful at providing sequence and annotation information for species where no reference genome sequence is available. In such cases, annotations can be made by comparative analysis of the derived EST sequences with reference genomes of related species (trans alignment). For instance, a recent study used 454 technology to generate 148 Mb of EST data from *Eucalyptus grandis*, a tree species with little genomic information available (52). The 454 technology has also been used to provide annotation information for the genome of wasp *Polistes metricus* (70) and maize *Zea mays* (20). Due to the longer read length compared to that produced by other new sequencing technologies (Table 2), ESTs generated by 454 can be effectively used for de

![Figure 3](https://www.annualreviews.org/doi/figure-pdf/10.1146/annurev.genom.10.081808.115102/100403)

Protein-coding gene annotation using transcriptome sequencing data. This figure illustrates how novel exons and introns can be discovered by mapping transcriptome sequencing reads to an annotated reference genome sequence.
Table 3  Applications of new sequencing technologies to the analysis of protein-coding transcriptomes of under-studied species

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Sequencing platform</th>
<th>Amount of sequence data generated, Mb</th>
<th>Reference genome size, Mb</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eucalyptus grandis</td>
<td>Flooded gum, rose gum</td>
<td>454/Roche</td>
<td>148</td>
<td>564 (for E. globulus)</td>
<td>52</td>
</tr>
<tr>
<td>Polistes metricus</td>
<td>Paper wasp</td>
<td>454/Roche</td>
<td>45</td>
<td>303 Mb (for P. dominulus)</td>
<td>70</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize</td>
<td>454/Roche</td>
<td>&gt; 26.3</td>
<td>2671</td>
<td>20</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Barrel clover</td>
<td>454/Roche</td>
<td>~26.9</td>
<td>466</td>
<td>17</td>
</tr>
<tr>
<td>Melitaea cinxia</td>
<td>Glanville fritillary butterfly</td>
<td>454/Roche</td>
<td>~66.9</td>
<td>N/A</td>
<td>74</td>
</tr>
<tr>
<td>Micropterus salmoides</td>
<td>Largemouth bass</td>
<td>454/Roche</td>
<td>&gt;58</td>
<td>978</td>
<td>22a</td>
</tr>
<tr>
<td>Manduca sexta</td>
<td>Tobacco hornworm</td>
<td>454/Roche</td>
<td>~17.7</td>
<td>N/A</td>
<td>86a</td>
</tr>
<tr>
<td>Sinorhizobium meliloti</td>
<td>Nitrogen-fixing bacterium rhizobium</td>
<td>454/Roche</td>
<td>~2.6</td>
<td>6.68</td>
<td>37a</td>
</tr>
<tr>
<td>Microtus aethiopoides</td>
<td>Parasitoid wasp</td>
<td>454/Roche</td>
<td>~26</td>
<td>N/A</td>
<td>18a</td>
</tr>
<tr>
<td>Pisum sativum</td>
<td>Pea</td>
<td>454/Roche</td>
<td>~230</td>
<td>4778</td>
<td>12a</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Grape</td>
<td>Illumina</td>
<td>~90</td>
<td>417</td>
<td>33b</td>
</tr>
<tr>
<td>Pythium ultimum</td>
<td>Plant pathogen (water mould)</td>
<td>454/Roche</td>
<td>~17.3</td>
<td>N/A</td>
<td>16a</td>
</tr>
</tbody>
</table>

1The table includes work published before or on November 15, 2008. Works involving a commonly studied organism, such as yeast, fruit fly, roundworm C. elegans, mouse, or human are excluded from the table.

2The genome sizes are from Reference 28a.

novel analyses, including assembly of the transcriptome, as was recently done for the transcriptome of the Glanville fritillary butterfly (Melitaea cinxia) (74). Although shorter reads produced by Illumina, SOLiD and HeliScope compared to the 454 technology may be more challenging for de novo sequence assembly, algorithms for assembly of such short reads have been developed (e.g., 85). The short-read data have been successfully used to detect novel exons and novel splicing events in species with an available reference genome sequence (e.g., 18, 44). For a list of under-studied organisms that enjoyed protein-coding gene annotation using new-generation transcriptome sequencing, see Table 3.

Gene Expression Profiling
Tag sequencing applications. Serial Analysis of Gene Expression (SAGE) was the first reported tag sequencing method for gene expression profiling (72). Even though it offered important advantages over competing microarray approaches (76), the SAGE method had not been used as widely as microarrays. However, the development of inexpensive next-generation sequencing technologies has revived the concept behind the original SAGE method and contributed to a growth in its popularity. The short reads produced by next-generation technologies, particularly Illumina and SOLiD, are compatible with the SAGE protocol and are arguably better suited for it than is the original Sanger sequencing. In particular, concatenation and cloning of SAGE tags are no longer required in next-generation sequencing, simplifying the SAGE procedure. To date, SAGE-like protocols have been used in conjunction with 454 (49) and Illumina (30; A.S. Morrissy et al., submitted) sequencing technologies. In addition to revisiting SAGE, other tag-based
methods for expression profiling have been developed that capitalize on the short-read structure and high throughput of the new sequencing technologies. For instance, 454 sequencing has been used in the novel method termed 5′-RATE, where tags corresponding to 5′ ends of transcripts are generated and sequenced providing information on the location of transcription start sites (27). Another novel tag-based method involving the 454 technology is 3′ UTR sequencing, wherein tags are generated from 3′ UTRs of mRNAs to allow for the distinction of closely related transcripts (21).

**Transcriptome shotgun sequencing.** Transcriptome sequences, produced by next-generation technologies, achieve sufficient sequencing depth to provide an adequate representation of the cellular transcriptome. With the elimination of the cloning step and common use of random priming, next-generation EST sequencing data became indistinguishable from those generated by transcriptome shotgun sequencing. In this approach, mRNA is reverse transcribed into cDNA, which is then fragmented and sequenced using a next-generation technology to generate reads covering the full length of a transcript (Figure 4). To date, the 454 technology has been used to generate transcriptome sequencing libraries from plants, e.g., *Arabidopsis thaliana* (78), *Medicago truncatula*, *Z. mays* (17), and other biological systems, such as *Drosophila melanogaster* (69), *Caenorhabditis elegans* (62a), and human cell lines (6, 81). The Illumina technology was used to develop a whole transcriptome shotgun sequencing (WTSS) procedure that was then applied to survey the transcriptome of HeLa cells (44). The WTSS procedure is also referred to as RNA-seq (47).

Other studies with Illumina transcriptome shotgun sequencing characterized the transcriptomes of a number of organisms and tissues, including mouse embryonic stem cells (56), human embryonic kidney and a B cell line (65), and yeast (47). To date, SOLiD technology has been used to resequence the transcriptome of human embryonic stem cells (18). Whole transcriptome sequencing data wherein sequencing reads are obtained from any location within the transcript, as opposed to a defined one as in tag-based sequencing (Figure 1), is versatile as, in addition to expression profiling, it can also be used for genome annotation, the detection of transcript aberrations, the discovery of alternative splice variants, and mutational profiling (44).

**Noncoding RNA Discovery and Detection**

Small noncoding RNAs (ncRNAs) have recently arisen as crucial regulators of development and cell fate determination. These

**Sequencing depth:** the total number of sequencing reads generated from a sequencing library. The higher the sequencing depth the higher the chance of detecting rare transcripts and sequence variants present in the cell.

---

**Reference genome sequence**

Exon1 Exon2 Intron1 Exon3

Sequencing reads

Digital expression read-out

**Figure 4**

Gene expression profiling using high-throughput transcriptome sequencing. The number of sequencing reads mapped to particular exons can be used to infer the abundance of the exons in the cell and hence the expression level of the corresponding transcripts.
18-30 nucleotide-long RNA molecules are transcribed from genomic DNA but not translated into a protein product.

Two classes of ncRNAs that have been implicated in many important processes, such as cell differentiation and oncogenesis, are microRNAs (miRNAs) and small interfering RNAs (siRNAs). These RNAs serve as posttranscriptional regulators of gene expression in a wide range of organisms (22, 77). Mature miRNAs and siRNAs bind to complementary sequences often found in UTR regions of genes and induce degradation of target mRNAs, thereby regulating their translation rates (22).

Next-generation sequencing technologies have had a profound influence on ncRNA research. Although extensively used for ncRNA expression profiling (84), microarrays are restricted to the detection of known miRNA and siRNA genes. Due to the high degree of ncRNA sequence diversity across different species, computational identification of novel ncRNA genes has had limited success (82). By contrast, massively parallel short-read sequencing technologies have been highly efficacious for the discovery of novel miRNA and siRNA genes on a genome-wide scale (Figure 5). In addition to novel miRNA discovery, sequencing-based approaches are amenable to the detection of variants of known miRNAs, RNA editing events, and miRNA-target RNA pairs (23, 55). To date, studies using 454 technology have characterized small noncoding RNAs in many species (e.g., 4, 5, 83, 86). One such study generated 166,835 sequence reads from the California poppy *Eschscholzia californica* and identified 173 distinct miRNA genes in this species (7). The higher throughput of Illumina and SOLiD technologies also enables the generation of deep miRNA libraries. Two recent studies using Illumina sequencing generated 6 and 9.5 million short sequence reads from small RNA libraries (26, 44a). Morin et al. (44a) identified 104 novel and 334 known miRNA genes expressed in human embryonic stem cells, while Glazov et al. (26) detected 449 novel and all known chicken miRNAs in the chicken embryo. Similar ncRNA profiling studies using Illumina have been conducted in other systems (34a, 50, 87).

Recent studies using 454 and Illumina sequencing have been used to identify endogenous siRNAs in mouse oocytes (66, 77) and *Drosophila* (19, 24); siRNA transcripts were previously uncharacterized in animals. Another key contribution is the discovery of a novel class of small RNAs, distinct from siRNAs and miRNAs. These RNA molecules, termed Piwi-interacting RNA (piRNA), were found to participate in RNA-protein complexes that are involved in transcriptional silencing in the germline of many species (25, 33, 36).

### Transcript Rearrangement Discovery

Genome rearrangements resulting in aberrant transcriptional events are common features of human cancers (29). Such rearrangements may include translocations, inversions, small insertions/deletions (indels), and copy number variants (CNVs) and may occur in all or a fraction of cancer cells within a tumor. While cytogenetics and microarray-based methods have been developed to identify genome rearrangements, most of them are suitable for the detection of only particular types of rearrangements and have limited resolution. Next-generation sequencing technologies offer important advantages over conventional methods such as...
microarrays or array comparative genomic hybridization (array CGH) for high-throughput detection of genome aberrations (46). In particular, transcriptome sequencing can be used to detect all types of genome rearrangements affecting coding sequences at potentially a single nucleotide resolution. Moreover, sequencing approaches are able to detect variants that are present in a subpopulation of cells (68). This point is particularly important in light of tumor heterogeneity, in which cells within a tumor can be genetically nonidentical (2).

Given the short read lengths generated by next-generation sequencers, effective detection of transcript aberrations requires the development of paired-end sequencing approaches. Several paired-end sequencing procedures have been developed for the 454 technology (34, 48). Multiplex sequencing of paired-end ditags (MS-PET) involves generation of short sequence tags from both ends of a fragment, their concatenation, and sequencing (48, 57). The method allows for the identification of fusion transcripts as well as other aberrations in human cancers (Figure 6). Similar procedures using the Illumina sequencing technology have been developed and applied to identify genome rearrangements in lung cancer at single-base resolution (14) and to map balanced chromosome rearrangements that occur in mental retardation (16).

Single-Nucleotide Variation Profiling
Recent genome sequencing efforts reveal an abundance of single-nucleotide variants present in individual human genomes (e.g., 75, 79). This variation may occur in the germline or somatic cells, such as those that comprise human tumors (28, 37, 80). An important component of genetic variation falls within coding regions of genes and may contribute to an alteration of their function. Although all types of genetic polymorphisms can be identified via resequencing of whole genomes, this method is still too costly to be conducted routinely. Instead, transcriptome sequencing studies may help reduce sequencing costs by restricting the focus of the analysis to coding parts of the genome. For instance, Morin et al. (44) generated 28.6 million 31-nucleotide reads from the HeLa transcriptome using the Illumina 1G Analyzer. As a result of this experiment, 36,445 exons were detected at tenfold coverage or more. In comparison, genome sequencing studies using the same technology require more extensive input to obtain an equivalent exonic coverage necessary for reliable detection of polymorphisms (75). In addition, comparative analysis of transcriptome and genome sequencing data can be used to reveal putative RNA-editing events that may account for site-specific differences between the genome and transcriptome of the same individual (55).

An important issue in single-nucleotide variation profiling using new sequencing technologies is the error rate associated with the sequencing chemistries and base calling. For instance, a recent study using Illumina 1G Analyzer estimated the per-base error rate to be between 0.3% and 3.8% depending on the base position in the sequencing read (19a). The reported per-base error rate for the 454/Roche technology is 4% (33a). While the developers of SOLiD cite a much higher accuracy for this platform (99.94%), this figure is not directly comparable to the figures for Illumina

Reference genome sequence

Mapping paired reads to the genome may reveal rearrangements

Generation of reads from both ends of DNA fragments

Figure 6
Transcript aberration discovery using transcriptome sequencing data. Reads are generated from both ends of DNA fragments and mapped to a reference genome sequence. If the fragment length is fixed, the distance between the reads when mapped to the reference sequence can be used to infer the presence of a rearrangement.

Paired-end sequencing:
sequencing in which reads from both ends of nucleic acid fragments are produced; used in Sanger sequencing for de novo genome assembly; particularly crucial for next-generation sequencers as the read pair information helps to reduce alignment ambiguities when mapping short sequencing reads to the genome.
and 454/Roche due to differences in data formats produced by these platforms (52a). In certain cases the sequencing chemistries utilized by a next-generation platform can introduce systematic error types. An example of this is seen with the 454/Roche platform’s difficulty with homopolymeric repeats. It is also important to note that the quality scores associated with base calls continue to evolve as developers become more familiar with the characteristics of a given platform. For example, over the last 18 months quality scores on the Illumina Genome Analyzer platform have evolved from highly inaccurate probabilities to reference alignment calibrated quality scores to alignment independent calibrated quality scores based on empirically generated data sets. The error rates and evolving quality metrics of next-generation sequencers are often circumvented by acquiring deep redundancy of read coverage to call single-nucleotide variants. For example, a recent cross-platform comparison demonstrated that a 10% false positive detection rate required 34-, 100-, and 110-fold coverage for 454/Roche, SOLiD, and Illumina Genome Analyzer, respectively (30a).

The sequencing redundancy thus contributes to increased costs associated with finding rare sequence variants and the need for subsequent validation work, often using Sanger-based resequencing. However, various computational approaches have promised to partially address the error rate by eliminating low quality data (e.g., 33a). Our own experience with Illumina Genome Analyzers over a 19 month period showed a decline in per-base error rates with the introduction of improved instrumentation, sequence chemistries, and base-calling algorithms (Supplementary Figure 1). Follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org. In addition, single-molecule DNA sequencing technologies such as those from Helicos and Pacific Biosciences may improve the error rate even further by eliminating PCR amplification from their procedures (30b).

CONCLUDING REMARKS

Complex diseases such as cancer are characterized by a variety of molecular aberrations such as gene expression changes, chromosomal rearrangements, point mutations, and epigenetic abnormalities (29). Therefore, a multidimensional understanding of the molecular features underlying a complex disease phenotype is required for the development of effective intervention strategies. Transcriptome sequencing by next-generation technologies provides resources for gene expression profiling studies as well as simultaneous identification of mutations, sequence aberrations, alternative splice variants, and RNA editing events. This review has focused on applications of next-generation sequencers to transcriptome analysis. However, the new technologies have had profound repercussions in other areas of genomics, such as genome sequencing and epigenome analysis (38, 39). Combining read outs from these different sequencing experiments presents exciting opportunities for multidimensional analyses of biological systems (15, 53).

SUMMARY POINTS

1. Research in transcriptome analysis has evolved from detection of single mRNA molecules to large-scale gene expression profiling and genome annotation efforts using microarrays and EST sequencing, respectively.

2. The development of a panel of new-generation sequencers with a much higher sequencing throughput than that of the state-of-the-art Sanger sequencer contributed to increasing the popularity of sequencing-based methods for transcriptome analysis.
3. Transcriptome sequencing data have been used for genome annotation, alternative isoform discovery, gene expression profiling, mutational profiling, noncoding RNA discovery and detection, the identification of aberrant transcriptional events, and the discovery of RNA editing sites.

4. The development of yet more efficient sequencing technologies known as third-generation sequencers promises to bring about a second DNA sequencing revolution centered around sequencing single DNA molecules.

FUTURE ISSUES

1. Most new-generation sequencers are limited by the short read length and the high error rate that hinder sequence assembly and read annotation. These technological shortcomings may be addressed in some third-generation sequencers, such as SMRT from Pacific Biosciences that promises accurate reads with length on the order of 100,000 bp.

2. Despite the many attractive prospects offered by transcriptome sequencing on next-generation platforms, still needed are advances in sequencing data analysis and the development of suitable computational tools that can be used to effectively process massive amounts of sequence data.

3. Most next-generation sequencing studies conducted to date have been of a descriptive nature involving basic data analysis; however, more in-depth data analyses are needed to fully understand the biological meaning of the data and to exploit their full potential.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Support for this work was received from the National Cancer Institute of Canada, Genome Canada, Genome British Columbia, and British Columbia Cancer Foundation. O.M. is a Junior Graduate Trainee of the Michael Smith Foundation for Health Research and is supported by a Natural Sciences and Engineering Research Council of Canada fellowship. M.A.M. is a senior scholar of the Michael Smith Foundation for Health Research. We apologize to those authors whose relevant work could not be cited here due to space limitations.

LITERATURE CITED


55. Reid JG, Nagaraja AK, Lynn FC, Drahek RB, Muzny DM, et al. 2008. Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5′-seed/cleavage/anchor regions and stabilize predicted mmu-let-7a:MRNA duplexes. Genome Res. 18:1571–81
57. Ruan Y, Ooi HS, Choo SW, Chiu KP, Zhao XD, et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using paired-end diT ags (PETs). Genome Res. 17:828–38


Contents

Chromosomes in Leukemia and Beyond: From Irrelevant to Central Players
Janet D. Rowley ................................................................. 1

Unraveling a Multifactorial Late-Onset Disease: From Genetic Susceptibility to Disease Mechanisms for Age-Related Macular Degeneration
Anand Swaroop, Emily Y. Chew, Catherine Bowes Rickman, and Gonçalo R. Abecasis .............................................. 19

 Syndromes of Telomere Shortening
Mary Armanios ................................................................. 45

Chronic Pancreatitis: Genetics and Pathogenesis
Jian-Min Chen and Claude Férec ........................................ 63

The Genetics of Crohn’s Disease
Johan Van Limbergen, David C. Wilson, and Jack Satsangi ........ 89

Genotyping Technologies for Genetic Research
Jiannis Ragoussis ............................................................... 117

Applications of New Sequencing Technologies for Transcriptome Analysis
Olena Morozova, Martin Hirst, and Marco A. Marra .................. 135

The Posttranslational Processing of Prelamin A and Disease
Brandon S.J. Davies, Loren G. Fong, Shao H. Yang, Catherine Coffinier, and Stephen G. Young ........................................... 153

Genetic Testing in Israel: An Overview
Guy Rosner, Serena Rosner, and Avi Orr-Urtreger ....................... 175

Stewardship of Human Biospecimens, DNA, Genotype, and Clinical Data in the GWAS Era
Stephen J. O’Brien .............................................................. 193

Schistosoma Genetics: New Perspectives on Schistosome Biology and Host-Parasite Interaction
Ze-Guang Han, Paul J. Brindley, Sheng-Yue Wang, and Zhu Chen ................................................................. 211

Evolution of Genomic Imprinting: Insights from Marsupials and Monotremes
Marilyn B. Renfree, Timothy A. Hore, Geoffrey Shaw, Jennifer A. Marshall Graves, and Andrew J. Pask ......................... 241
Methods for Genomic Partitioning
   Emily H. Turner, Sarah B. Ng, Deborah A. Nickerson, and Jay Shendure .......... 263

Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes
   Laurent Duret and Nicolas Galtier .................................................. 285

Inherited Variation in Gene Expression
   Daniel A. Skelly, James Ronald, and Joshua M. Akey ............................ 313

Genomic Analyses of Sex Chromosome Evolution
   Melissa A. Wilson and Kateryna D. Makova ....................................... 333

Sequencing Primate Genomes: What Have We Learned?
   Tomas Marques-Bonet, Oliver A. Ryder, and Evan E. Eichler ....................... 355

Genotype Imputation
   Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis ..................... 387

Genetics of Athletic Performance
   Elaine A. Ostrander, Heather J. Huson, and Gary K. Ostrander .................... 407

Genetic Screening for Low-Penetrance Variants in Protein-Coding Diseases
   Jill Waalen and Ernest Beutler ......................................................... 431

Copy Number Variation in Human Health, Disease, and Evolution
   Feng Zhang, Wenli Gu, Matthew E. Hurles, and James R. Lupski ................. 451

The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms
   Bryan G. Fry, Kim Roelants, Donald E. Champagne, Holger Scheib,
   Joel D. A. Tyndall, Glenn F. King, Timo J. Nekula, Janette A. Norman,
   Richard J. Lewis, Raymond S. Norton, Camila Renjifo,
   and Ricardo C. Rodriguez de la Vega .................................................. 483

Genetics, Medicine, and the Plain People
   Kevin A. Strauss and Erik G. Puffenberger ........................................ 513

Indexes

Cumulative Index of Contributing Authors, Volumes 1–10 .......................... 537
Cumulative Index of Chapter Titles, Volumes 1–10 .................................... 541

Errata

An online log of corrections to Annual Review of Genomics and Human Genetics articles may be found at http://genom.annualreviews.org/