APPLICATIONS OF NEXT-GENERATION SEQUENCING

RNA sequencing: advances, challenges and opportunities

Fatih Ozsolak and Patrice M. Milos

Abstract | In the few years since its initial application, massively parallel cDNA sequencing, or RNA-seq, has allowed many advances in the characterization and quantification of transcriptomes. Recently, several developments in RNA-seq methods have provided an even more complete characterization of RNA transcripts. These developments include improvements in transcription start site mapping, strand-specific measurements, gene fusion detection, small RNA characterization and detection of alternative splicing events. Ongoing developments promise further advances in the application of RNA-seq, particularly direct RNA sequencing and approaches that allow RNA quantification from very small amounts of cellular materials.

Next generation DNA sequencing
(Often abbreviated to NGS.) Non-Sanger-based high-throughput DNA sequencing technologies. Compared to Sanger sequencing, NGS platforms sequence as many as billions of DNA strands in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes.

Helicos BioSciences Corporation, One Kendall Square, Cambridge, Massachusetts 02139, USA.
e-mails: fozsolak@helicosbio.com; pmilos@helicosbio.com
doi:10.1038/nrg2934
Published online
30 December 2010

Over the past 10 years we have come to appreciate the dynamic state of genomes, including both DNA modifications and RNA quantitative and qualitative changes, which have been characterized in species ranging from simple model organisms to humans. This advance has occurred through the use of various genomic measurements, including comprehensive transcriptomics studies[1]. We now have a new appreciation for the complexity of the transcriptome, encompassing a multitude of previously unknown coding and non-coding RNA species, particularly small RNAs (sRNAs), including microRNAs, promoter-associated RNAs and newly discovered antisense 3′ termini-associated RNA, to name a few[2,3].

Initial transcriptomics studies largely relied on hybridization-based microarray technologies and offered a limited ability to fully catalogue and quantify the diverse RNA molecules that are expressed from genomes over wide ranges of levels. The introduction of high-throughput next-generation DNA sequencing (NGS) technologies[4] revolutionized transcriptomics by allowing RNA analysis through cDNA sequencing at massive scale (RNA-seq). This development eliminated several challenges posed by microarray technologies, including the limited dynamic range of detection[5]. NGS platforms used for RNA-seq are commercially available from four companies — Illumina, Roche 454, Helicos BioSciences and Life Technologies — and new technologies are in development by others[4]. Given the importance of sequencing capabilities, such as throughput, read length, error rate and ability to perform paired reads, for RNA-seq as well as genomic studies, NGS companies are constantly improving their platforms to provide the best sequencing performance at the lowest cost[4].

New methodologies for RNA-seq studies have been providing a progressively fuller knowledge of both the quantitative and qualitative aspects of transcript biology in both prokaryotes[6] and eukaryotes[5]. Here we discuss these advances, which have included the development of approaches to allow a more comprehensive understanding of transcription initiation sites, the cataloguing of sense and antisense transcripts, improved detection of alternative splicing events and the detection of gene fusion transcripts, which has become increasingly important in cancer research — all at a data scale that was unimagined just several years ago. Recently developed approaches also allow the selection of specific RNA molecules before RNA-seq, allowing transcriptomics studies with more focused aims. In this Review, we provide an overview of these methods, touching only briefly on the types of biological insight that they allow, and focusing on the technologies themselves. We provide a comparison of the different approaches that are available for each application and discuss the current limitations and the potential for future improvements. We conclude by discussing two new developments in RNA-seq technologies: direct RNA sequencing (DRS)[7] and methods for the reliable profiling of minute RNA quantities, which is important for translational research and clinical applications of RNA-seq.

Mapping transcription start sites

The mapping of transcription start sites (TSSs) at nucleotide resolution is necessary to fully define RNA

products and to identify adjacent promoter regions that regulate the expression of each transcript. One of the first high-throughput TSS mapping methods was the cap analysis of gene expression (CAGE) approach, which was initially developed for Sanger sequencing[8,9]. This involved sequencing of cloned cDNA products derived from RNAs with intact 5′ ends (for example, containing a 5′ cap structure). Although useful, the technology required high quantities of input RNA and generated only short reads (~20 nucleotides) per TSS.

These limitations prompted the adaptation of the CAGE approach for NGS platforms, which has resulted in the discovery of the unexpected complexity of TSS distribution across genomes and in the regions surrounding individual promoters. Methods that combine RNA-seq with CAGE include deepCAGE[10], PEAT[11], nanoCAGE and CAGEscan[12], which collectively resolve several technical challenges of the initial Sanger sequencing-based CAGE strategies (TABLE 1). First, nanoCAGE[12] now allows TSS mapping from total RNA quantities as small as 10 nanograms through the use of various amplification strategies. Second, the compatibility of PEAT and CAGEscan with paired-end sequencing (a capability that is enabled by platforms such as Illumina, but is lacking in others such as Helicos) allows examination of the connectivity of TSSs with downstream regions and facilitates the assignment of identified TSSs to specific transcripts. In addition, paired-end sequencing partly alleviates the difficulty of aligning single short reads to repeat regions and thus allows a subset of repeat elements to be at least partially characterized by RNA-seq.

However, there are several caveats of these NGS-based approaches. One is that no attempt has been made to examine whether the amplification and other manipulation steps that are carried out distort the resulting view of how frequently each TSS is used. Spike-in experiments would be useful to address this issue. In addition, multiple difficulties were encountered during the development of protocols involving cDNA synthesis and amplification[12]. For example, researchers observed artefacts such as primer dimers that dominated sequencing data sets and reduced effective coverage, prompting the use of semisuppressive PCR to reduce primer dimer frequency[12]. Thus, although these methods may be useful for qualitative applications, establishing and improving their quantitative capabilities will probably require additional development.

General limitations of RNA-based TSS mapping approaches include their dependence on cDNA synthesis or hybridization steps, the efficiency of which is dependent on RNA sequence and structure. In addition, RNA-based TSS mapping is challenging for short-lived transcripts such as primary microRNAs, which are transcribed generally at high levels but are scarce owing to their rapid degradation. These limitations may be partly alleviated when combined with other methods such as chromatin-based TSS prediction, which relies on detecting histone modifications that are indicative of active transcription[13,14]. Such integration may also be useful in light of the recent suggestion that post-transcriptional processing results in 5′ cap-like structures in RNA fragments[15]. Thus, relying solely on CAGE data for TSS mapping may result in difficulties in separating transcription initiation events from RNA processing events.

Strand-specific RNA-seq

Transcriptomic studies in a range of species have revealed a pervasive presence of antisense transcription events[16]. Although these events were once considered to reflect biological or technical noise, it is now clear that antisense transcripts are functional and have various roles in both normal physiological states and disease states[16]. There is therefore an increasing interest in profiling transcriptomes at greater depths to fully characterize sense and antisense transcription products. Standard RNA-seq approaches generally require double-stranded cDNA synthesis, which erases RNA strand information. In addition, during first-strand cDNA synthesis, spurious second-strand cDNA artefacts can be introduced, owing to the DNA-dependent DNA polymerase (DDDP) activities of reverse transcriptases[17–19], which can confound sense versus antisense transcript determination[20]. Actinomycin D has been suggested as a potential agent to reduce DDDP activities of reverse transcriptases[18], but the extent to which it is effective, and whether or not it introduces additional artefacts, has not been fully examined. To overcome these difficulties, several strategies for strand-specific analyses of transcriptomes have been developed.

The strategies that have been developed to generate strand-specific information generally rely on one of three approaches. The first involves the ligation of adaptors in a predetermined orientation to the ends of RNAs or to first-strand cDNA molecules[21–23]. The

Table 1 | **Next generation sequencing-based approaches for transcription start site mapping**

TSS method	RNA sequence data	Starting RNA	Refs
CAGE	5′ end of transcripts	50 µg total RNA	9
DeepCAGE	5′ end of transcripts	10 ng total RNA	10
nanoCAGE	5′ end of transcripts	10 ng total RNA	12
CAGEscan	5′ end of transcripts and either 3′ end or internal RNA sequence	10 ng total RNA	12
PEAT	5′ end of transcripts paired with random reads along the RNA	150 µg total RNA	11

CAGE, cap analysis of gene expression; CAGEscan, paired read to combine 5′ CAGE with downstream sequence; DeepCAGE, high-throughput CAGE sequencing; nanoCAGE, low-quantity CAGE; ng, nanograms; PEAT, paired end analysis of transcription start sites; TSS, transcription start site.

Semisuppressive PCR
A PCR strategy that aims to reduce primer dimer accumulation by preferentially amplifying longer DNA fragments.

known orientations of these adaptors are used as reference points to obtain RNA strand information. A second approach is the direct sequencing of the first-strand cDNA products that are generated, either in solution[24,25] or on surfaces[26]. Last, a third approach is the selective chemical marking of the second-strand cDNA synthesis products or RNA[27,28]. These strategies have already begun to contribute to our understanding of transcriptomes, including mapping of translation states of RNAs (for example, polysome profiling)[29] and identification of novel promoter-associated RNAs[22].

A recent study that used the *Saccharomyces cerevisiae* genome as a reference compared the performance of several of these strategies, and the authors observed differences in these methods with respect to their level of strand specificity, evenness of coverage, agreement with known annotations, library complexity (for example, number of unique read start positions, which indicates the protocols' abilities to avoid amplification artefacts such as duplicate reads) and ability to generate quantitative expression profiles[30]. However, in-depth comparative studies that characterize the biases and artefacts that are introduced by each of these approaches are still lacking, and scientists working with these data sets should be aware of several issues.

First, given the tendency of reverse transcriptase to generate spurious second-strand cDNA products during first-strand cDNA synthesis[17–19], it is not clear whether the approaches that rely on sequencing first-strand cDNA products (either directly or by intra- or intermolecular ligation) are absolutely strand specific. The strand specificity of such approaches has been reported by quantifying the ratio of reads that map in the antisense orientation to the known, well-annotated genes, relative to the reads that map in the sense orientation. This investigation revealed that a small fraction of reads obtained with these approaches still align in the antisense orientation; thus, these approaches may not be entirely strand-specific[30]. Furthermore, cDNA products that contain both first- and second-strand cDNA products may not align properly to reference sequences. Given the incomplete annotations of sense and antisense transcripts in genomes, even in those of well-studied species such as *S. cerevisiae*, the true extent of strand specificity of these approaches should be carefully assessed. Ideally, such assessment should be performed with chemically synthesized RNA spike pools of defined sequence.

Second, ligation tends to have sequence preferences[31,32]. Thus, the approaches that rely on ligation may suffer from various representational biases. Examples of such bias are found in transcriptome profiling[23] and ribosome profiling experiments[29], in which extremely uneven coverage was seen for libraries prepared using ligation, compared with libraries prepared using enzymatic 3′ polyadenylation[29]. Third, the in-solution or on-surface amplification step included in some of these approaches may introduce additional artefacts — for example, in the form of GC biases and duplicate reads[33–35]. Examination of such effects revealed a duplicate read fraction in the range of 6.1% to 94.1% for standard and strand-specific Illumina RNA-seq strategies, and the existence of GC bias towards RNA templates with neutral GC content[23]. It is hoped that many of these limitations will be overcome by the sequencing technologies that are in development or with modifications and improvements to existing sequencing technologies[4].

Characterization of alternative splicing patterns

Given the importance of alternative splicing patterns in development and the fact that 15–60% of known disease-causing mutations affect splicing[36,37], it will be crucial to catalogue the complete repertoire of splicing events and to understand how altered splicing patterns contribute to development, cell differentiation and human disease. Initial splice-site mapping studies using RNA sequencing-based approaches were limited by read length, which prevented the reliable alignment to the genome of the two independent exonic portions of each read, representing the exon splicing event. Thus, initial RNA-seq-based studies of alternative splicing used computational strategies to compensate for this limitation. The reference sequence used for alignment was supplemented with 'artificial' sequences that surround all possible splice junctions between the annotated exons of genes, allowing the reads to be aligned[38–41]. These approaches changed our view of human splicing, as more than 95% of human multi-exon genes were found to be alternatively spliced, with ~110,000 novel splice sites per tissue[42]. By counting the number of reads mapping to each exon and spanning each splice junction, these approaches also allowed the splice efficiency of each junction to be determined and the levels of distinct isoforms to be quantified[43,44].

Improvements to current sequencing technologies now enable longer read lengths, allowing better mapping of the reads to the alternatively spliced exons. This improvement comes from being able to partition the reads into multiple pieces and to align each piece independently to the genomes. In addition, approaches that involve paired-end reads now enable sequence information to be obtained from two points in a transcript with an estimated distance between the reads. As a result, it is now possible to search for splicing patterns without a requirement for prior knowledge of transcript annotations[45,46] (FIG. 1). Examination of splicing patterns and transcript connectivity in an unbiased and genome-wide manner requires full-length transcript sequences to be obtained, which may be enabled in the future by emerging technologies[47,48].

Gene fusion detection

RNA-seq combined with computational analyses analogous to the ones described above for splice-site detection can also be used to identify gene fusion events in disease tissues, which has particular importance for cancer research[49]. Genomic DNA can be analysed with single-read and paired-end-read strategies for the detection of translocations and other genomic rearrangements[50]. However, RNA-seq may be preferable for identifying events that produce aberrant RNA species and therefore have a higher likelihood of being functional or causal in biological or disease settings[51,52] (FIG. 2).

Spike pool
Internal controls added to RNA samples, consisting of RNA elements of known sequence and composition.

Paired-end reads
A strategy involving sequencing of two different regions that are located apart from each other on the same DNA fragment. This strategy provides elevated physical coverage and alleviates several limitations of NGS platforms that arise because of their relatively short read length.

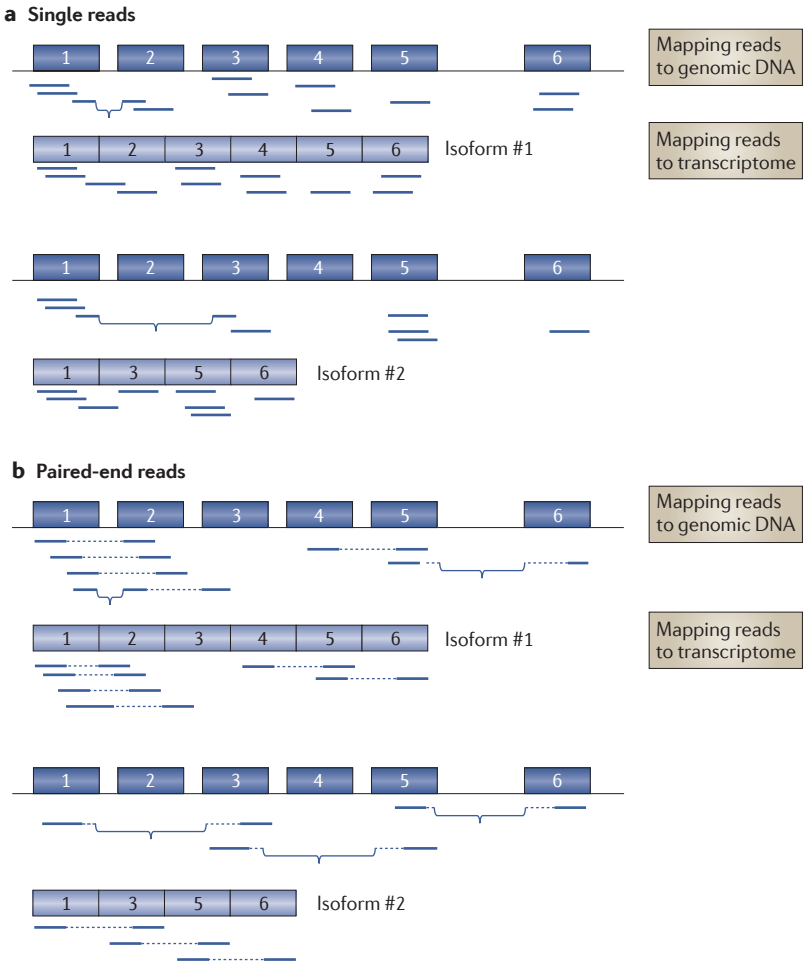


Figure 1 | **RNA-seq for detection of alternative splicing events. a** | Sequence reads are mapped to genomic DNA or to a transcriptome reference to detect alternative isoforms of an RNA transcript. Mapping is based simply on read counts to each exon and reads that span the exonic boundaries. One infers the absence of the genomic exon in the transcript by virtue of no reads mapping to the genomic location. **b** | Paired sequence reads provide additional information about exonic splicing events, as demonstrated by matching the first read in one exon and placing the second read in the downstream exon, creating a map of the transcript structure.

Furthermore, genomic DNA-based approaches cannot identify fusion events that are due to non-genomic factors, such as *trans*-splicing[53] and read-through events between adjacent transcripts[51,54]. Paired-end RNA-seq can be particularly advantageous for fusion identification because of the increased physical coverage it offers. This approach has led to important biological findings in oncology[55,56], offering potential targets for therapeutic modulation.

The challenges faced in fusion detection are generally in parallel with those for alternative splicing detection. In addition, RNA-seq-based analyses cannot detect fusion events that involve the exchange of the promoter of a gene with the coding sequence of another gene. Furthermore, RNA-seq data include chimeric cDNA artefacts that are generated by template switching during reverse transcription and amplification[57] (discussed below), leading to false positives in gene fusion identification. These difficulties may be partly alleviated when long-read RNA sequencing technologies with sufficient throughput and sequencing performance become available[4].

Targeted approaches using RNA-seq

Despite the increasing capabilities of NGS in terms of throughput and decreasing costs per data point, the expenditure necessary to obtain sufficient sequencing coverage for several research and potential clinical applications is still prohibitive. Such applications include the characterization of low-abundance transcripts and genotyping to determine, for example, which alleles of the transcripts might be differentially expressed. In these scenarios, it may be preferable to enrich for the desired subset of transcripts, to minimize the overall cost of sequencing and maximize the number of samples that can be analysed.

Target-enrichment strategies were originally developed for genomic DNA resequencing[4,58]. Many of these technologies have been used to capture the human exome from genomic DNA, given that a large fraction of disease-causing mutations are likely to be located in the protein-coding transcriptome. RNA-seq of poly(A)$^+$ RNA species offers a natural route for exome sequencing without the use of enrichment strategies. The potential suitability of mRNA-seq data for the identification of nucleotide variations has been demonstrated recently by several studies[59–61]. However, these studies also underscored some challenges — for example, the high sequencing depth required to sufficiently cover low-abundance transcripts.

Slight modifications of the genomic DNA-enrichment strategies for cDNA applications have allowed the development of targeted RNA-seq (FIG. 3). Targeted RNA-seq approaches have been used to detect fusion transcripts, allele-specific expression, mutations and RNA-editing events in a subset of transcripts[62–64]. Targeted RNA-seq strategies currently require longer sample preparation steps and higher input RNA and cDNA quantities than do other RNA-seq approaches, owing to the additional probe or microarray preparation and target-selection steps. Furthermore, capture efficiency usually differs between target regions depending on hybridization efficiency and other factors. Simplification of this process and improvements in capture efficiency are desirable for better experimental outcomes.

Small RNA profiling

The impact of NGS technologies on sRNA discovery and characterization has been particularly noteworthy. These studies have been reviewed extensively by others (for example, see REF. 65), so we do not review this topic in depth here but provide a brief summary for completeness.

Most initial sRNA-discovery studies used pyrosequencing[66,67]. Subsequently, the use of other NGS platforms with higher throughput has resulted in genome-wide surveys and the discovery of an ever-growing number of sRNA species[15,68,69]. Because NGS sample preparation strategies for 'longer' RNAs

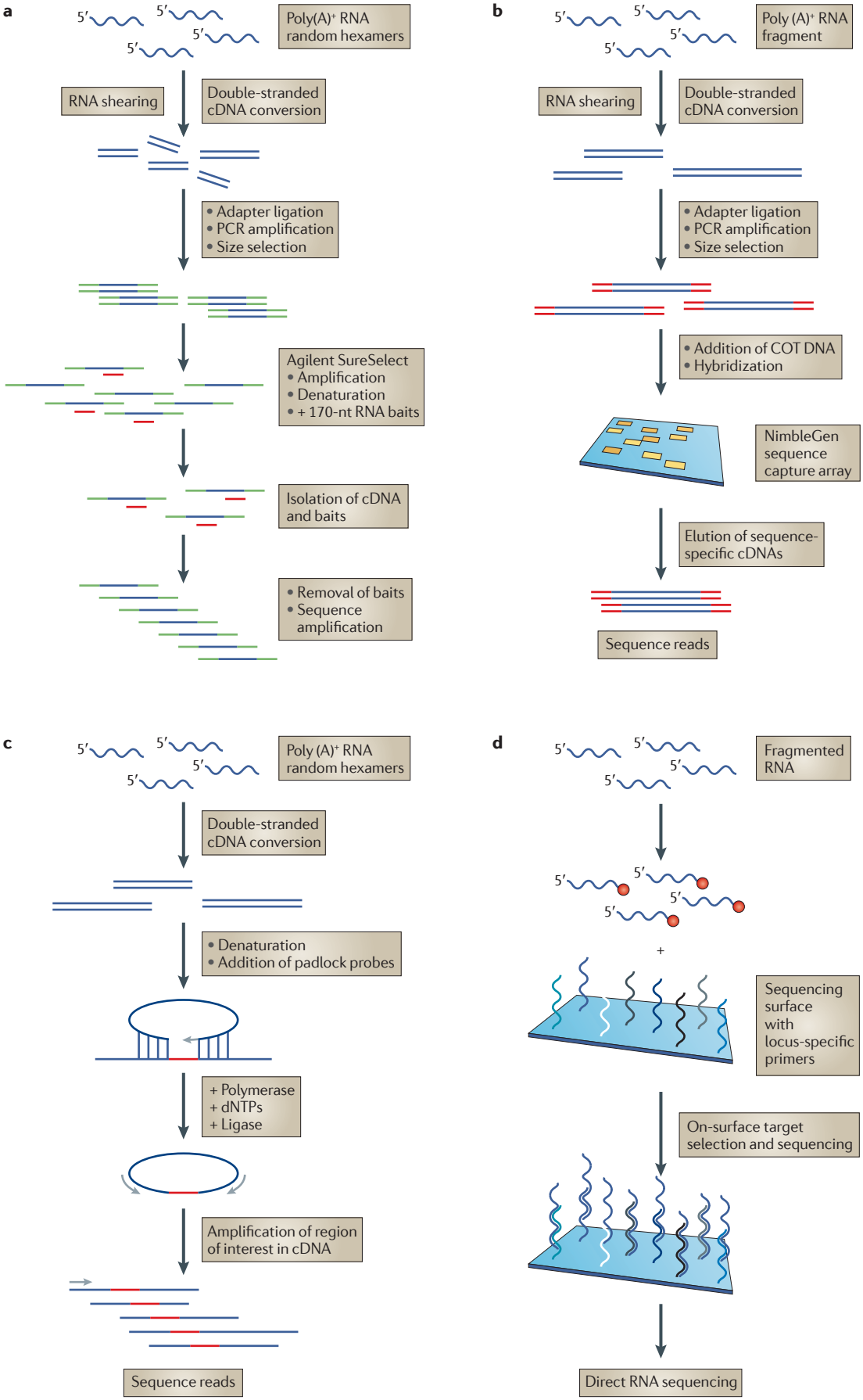
Figure 2 | **Use of RNA-seq for *BCR–ABL* fusion gene detection. a** | Breakpoint cluster region (*BCR*) and *ABL1* gene transcripts. **b** | *BCR–ABL* fusion gene transcript. **c** | Sequence reads mapping across the *BCR–ABL* fusion gene site demonstrating the ability to accurately identify the site of gene fusion. The data were derived from RNA-seq analysis[25] of the K562 transcriptome using the HeliScope (the raw data files are available at the University of California Santa Cruz Genome Browser and the Helicos Technology Center).

(>200 nucleotides) are not suitable for sRNAs, such as reverse transcription with random priming (because this way of priming cDNA synthesis from short RNA species yields even shorter cDNA species that are not long enough for efficient alignment), modified preparation strategies were developed[70–72].

One important limitation of the current RNA-seq-based approaches for studying sRNAs is their inability to provide an absolutely quantitative view of these transcripts. It has recently become clear that, although the NGS-based sRNA-profiling approaches can be used for differential expression analyses, the number of reads obtained per sRNA does not necessarily correlate with their actual abundance[73,74]. This discrepancy seems to be due to biases that are introduced during the sample preparation and sequencing steps. Whether emerging technologies can improve sRNA quantification remains to be seen.

Direct RNA sequencing

cDNA synthesis and other RNA manipulations limit some RNA-seq applications. As noted above, most current RNA-seq methods rely on cDNA synthesis and a range of subsequent manipulation steps, which places limitations on the current approaches for some applications. For example, as we have discussed, the generation of spurious second-strand cDNAs can present difficulties for strand-specific RNA-seq. Strand-specific libraries can also be prepared to avoid this problem (discussed above), but the approaches that use RNA–RNA ligation are laborious to construct. Another limitation imposed by cDNA synthesis is template switching[75–77]. During the process of reverse transcription, the nascent cDNA that is being synthesized can sometimes dissociate from the template RNA and re-anneal to a different stretch of RNA with a sequence similar to the initial template, generating artefactual



◀ Figure 3 | **Alternative methods for targeted RNA-seq. a** | Using poly(A)$^+$ RNA converted to double-stranded cDNA the Agilent SureSelect method uses RNA probes to enrich selected cDNA[62]. **b** | A custom NimbleGen array may allow selection of cDNAs of interest. **c** | The generation of DNA molecules with sequence-specific complementary targeting sites allows the targeting of cDNAs[63,64]. **d** | Helicos sequencing surfaces containing target-specific oligonucleotides can be used to select desired RNA, DNA and cDNA species and sequence regions of interest in a single step. nt, nucleotide.

chimeric cDNAs. Template switching may cause problems in the identification of exon–intron boundaries and true chimeric transcripts. Reverse transcriptases can also synthesize cDNA in a primer-independent manner, which is thought to be caused by self priming arising from the RNA secondary structure. This results in the generation of random cDNA synthesis. Furthermore, reverse transcriptases have lower fidelity compared to other polymerases owing to their lack of proofreading mechanisms[78,79], and they have variable RNA to cDNA conversion efficiency depending on the experimental conditions.

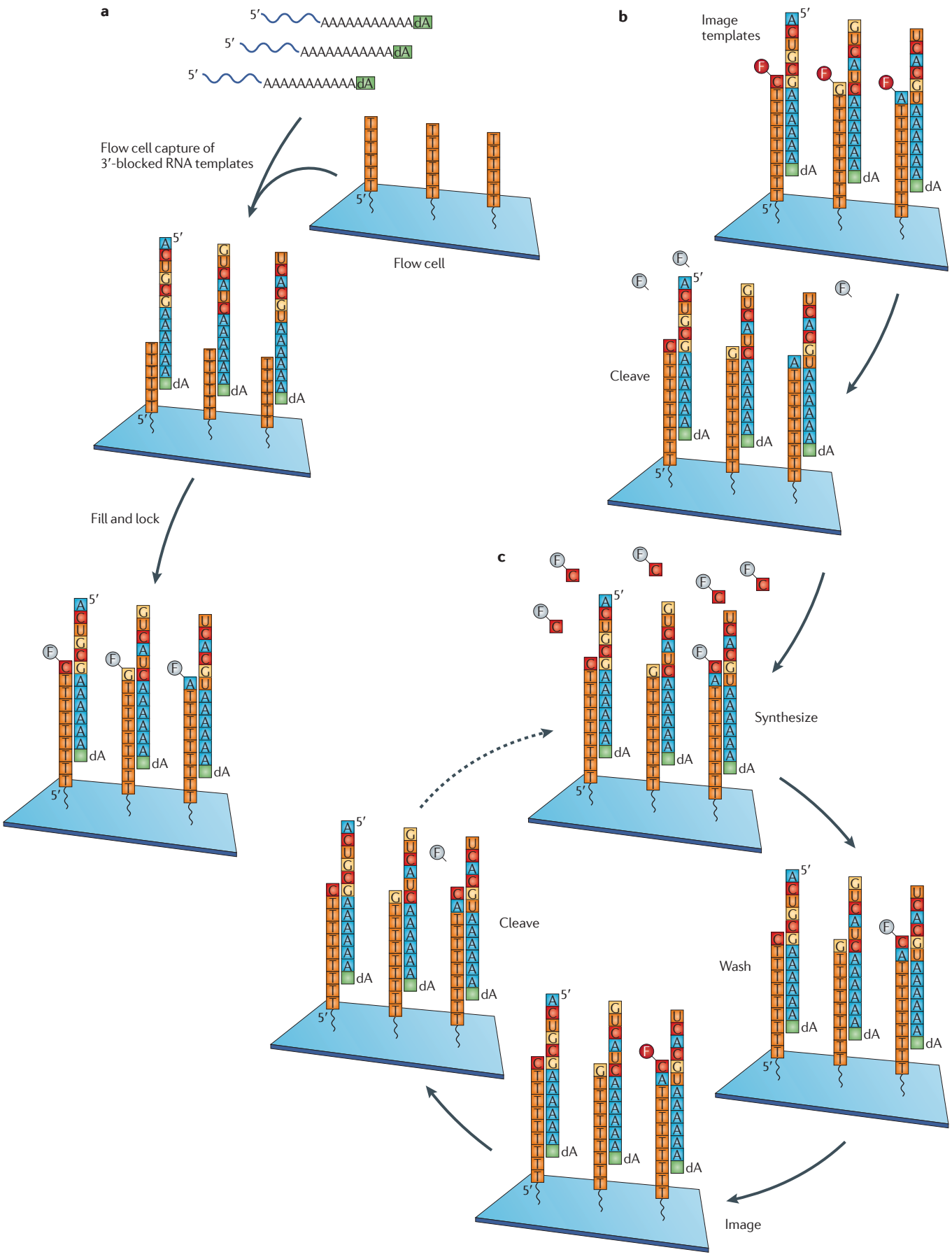
In addition to their requirement for cDNA synthesis, current RNA-seq approaches can present other difficulties. First, the RNA-seq signal across transcripts tends to show non-uniformity of coverage, which may be a result of biases introduced during various steps, such as priming with random hexamers[80,81], cDNA synthesis, ligation[31,32], amplification[35] and sequencing[33–35,82]. Second, commonly used RNA-seq strategies can result in transcript-length bias because of the multiple fragmentation and RNA or cDNA size-selection steps they use[83]. This bias may result in complications for downstream analyses[84]. Third, quantification of transcripts with RNA-seq requires consideration of read mapping uncertainty (owing to sequencing error rates, repetitive elements, incomplete genome sequence and inaccuracies in transcript annotations)[85] and normalization of the number of reads mapping to each transcript, based on transcript length. Despite improvements in sequencing methods and bioinformatics advances allowing *de novo* construction of transcriptomes[86,87], the existing approaches are often not sufficient to detect certain transcripts and/or cover their entire length. Together with the uncertainty regarding transcript boundaries and length because of events such as alternative splicing, polyadenylation sites and promoter usage, the required length-normalization step is a potential source of errors for quantitative applications. Fourth, RNA-seq strategies often involve a poly(A)$^+$ mRNA-enrichment step. Polyadenylation of transcripts also takes place during transcript degradation steps, and thus poly(A)$^+$-enrichment steps may also enrich for RNA degradation products of RNA polymerase I transcripts and other RNAs[88].

Direct sequencing of RNAs. The limitations of current RNA-seq approaches discussed above might be at least partly alleviated by emerging RNA analysis technologies, including DRS, that substantially alter the method of RNA characterization. DRS currently requires single-molecule sequencing capabilities, as the amplification of RNA molecules directly without cDNA conversion has not been examined. Although RNA-dependent RNA polymerases do exist[89], the extent to which they can be adapted to the amplification-based next-generation sequencing technologies is unknown at present.

The first massively parallel DRS approach was recently developed using the Helicos single-molecule sequencing platform[7,90,91] (FIG. 4). It relies on hybridization of several femtomoles of 3′-polyadenylated RNA templates to single channels of poly(dT)-coated sequencing surfaces, followed by sequencing by synthesis. This approach can select and sequence poly(A)$^+$ RNA from total RNA or cellular lysates, with sequence data being derived from regions immediately upstream of the polyadenylation sites[7]. Thus, the technology offers a path to obtain gene expression profiles and map polyadenylation sites in a quantitative and genome-wide manner. RNA species that lack natural poly(A) tails can be polyadenylated *in vitro* and analysed with DRS.

The development of DRS approaches that are free from cDNA synthesis artefacts such as template switching and spurious second-strand synthesis provides potential improvements for applications such as the surveying of strand-specific transcription. Furthermore, DRS requires only femtomole or attomole levels of input RNA, depending on the application, and involves relatively simple sample preparation. DRS-type technologies may therefore be advantageous for applications that are challenging for current cDNA-based methodologies, such as experiments that yield subnanogram-level RNA (discussed below), archival specimens or short RNA species, which cannot be easily converted to cDNA. Furthermore, unlike cDNA-based approaches, which require different strategies for the analysis of short and longer RNA species, DRS sample preparation involving polyadenylation can be applied to any RNA species, thus allowing both short and long RNAs to be observed in a single experiment. DRS may in the future also simplify targeted RNA-seq by enabling the integration of target selection and sequencing steps (FIG. 3d). Such integration may reduce the sample preparation steps to only nucleic acid fragmentation, and may minimize costs as well as the quantity of input nucleic acid required.

A key challenge for DRS is to generate the multi-million-level read quantities that are required for many RNA applications, particularly quantification, and to further reduce error rates and input RNA quantities through alterations to the sequencing chemistry and template-capture steps. DRS may also not solve all of the RNA-seq limitations listed above — including, for example, the issues of degradation products being captured during poly(A)$^+$ RNA selection. Furthermore, the combination of paired-end approaches with DRS and longer read lengths is needed for various applications discussed above, including studies focusing on the identification of 5′ (for example, CAGE-type TSS mapping) and 3′ boundaries of RNA species.



◀ Figure 4 | **Direct RNA sequencing using the Helicos approach. a** | RNA that is polyadenylated and 3′ deoxy-blocked with poly(A) polymerase is captured on poly(dT)-coated surfaces. A 'fill-and-lock' step is performed, in which the 'fill' step is performed with natural thymidine and polymerase, and the 'lock' step is performed with fluorescently labelled A, C and G Virtual Terminator (VT) nucleotides[104] and polymerase. This step corrects for any misalignments that may be present in poly(A) and poly(T) duplexes, and ensures that the sequencing starts in the RNA template rather than the polyadenylated tail. **b** | Imaging is performed to locate the positions of the templates. Then, chemical cleavage of the dye–nucleotide linker is performed to release the dye and prepare the templates for nucleotide incorporation. **c** | Incubation of this surface with one labelled nucleotide (C-VT is shown as an example) and a polymerase mixture is carried out. After this step, imaging is performed to locate the templates that have incorporated the nucleotide. Chemical cleavage of the dye allows the surface and DNA templates to be ready for the next nucleotide-addition cycle. Nucleotides are added in the C, T, A, G order for 120 total cycles (30 additions of each nucleotide).

Profiling low-quantity RNA samples

Biological specimens (such as tissue and body fluids) are generally heterogeneous, being a complex mixture of multiple cell types. The need to specifically select and study particular cells is clear, but the implementation of this task is not straightforward. Several tools now allow selection of specific cell types, such as flow-assisted cell sorting (FACS), laser-capture microdissection (LCM)[92], serial dilution, specialized microfluidic devices[93] and micromanipulation. In addition, methods for high-quality RNA isolation from small quantities of cells are also available. The main limitation preventing reliable, global profiling of minute RNA quantities has been the incompatibility of high-throughput RNA profiling approaches with low-quantity RNA samples. The absence of such methods has slowed our progress in a range of areas, such as forensics, stem cell biology, metagenomics and plant biology. The effects of this limitation are perhaps most acutely felt in research into cancer and other diseases, as samples obtained from patients are generally limited in quantity; the transition between findings from molecular profiling studies and technologies for use in clinical research and molecular diagnostics is being held back, slowing our progress towards personalized medicine. Strategies that can provide a comprehensive and bias-free view of transcriptomes using picogram quantities of input RNA would therefore stimulate great advances in a range of areas.

Methods for small quantities of RNA. The analysis of low-quantity RNA samples with global microarray and sequencing technologies has traditionally required one or more amplification step(s) to obtain sufficient nucleic acid material for subsequent detection. Since the early 1990s, several nucleic acid amplification strategies for low-quantity RNA applications have been developed, such as ligation-mediated PCR[94], multiple displacement amplification (MDA)[95], single-primer isothermal amplification[96] and *in vitro* transcription (IVT)-based linear amplification[97]. The ideal amplification method should provide accurate sequences with a low or zero error rate, be reproducible, produce high levels of amplification to provide the quantities of nucleic acid needed, be applicable for nucleic acids from a wide array of species, and preserve the representation of the distinct RNA molecules in the original sample. To what extent the current methods meet these criteria is not clear. Studies performed with microarray-based measurements suggest that amplification introduces variability and discrepancies, especially for middle- and low-abundance transcripts and as input RNA quantity is lowered further[98].

Sequencing-based low-quantity RNA profiling is relatively new. A recently reported mRNA-seq method relies on double PCR amplification steps and can be used to profile the transcriptomes of single oocytes[40]. It was observed, however, that the reproducibility of such low-quantity RNA-seq approaches may be negatively affected owing to stochastic amplification biases that may result in the drop-out of some RNA species and preferential amplification of others[23]. Such outcomes can lead to, for instance, duplicate reads and reduced quantification power.

Emerging technologies. A number of both hybridization- and sequencing-based technologies are now emerging that may allow reliable transcriptome profiles to be obtained from minute cell quantities. On the sequencing side, nanoCAGE[12] now allows TSS mapping from 10 nanograms of total RNA through the use of various amplification strategies. Amplification-free RNA-seq approaches have recently been developed that minimize the quantity of input RNA required. One approach involves the sequencing of first-strand cDNA products from as little as ~500 picograms of RNA, with priming carried out in solution with oligo-dT or random hexamers[24,25]. Another approach involves the use of poly(dT) primers on sequencing surfaces to select for poly(A)$^+$ mRNA from cellular lysates, followed by on-surface first-strand cDNA synthesis and sequencing[26]. This approach allows reproducible gene expression profiles to be obtained from ~1,000 cells and eliminates RNA loss during the RNA isolation steps, which may be particularly important as the input cell quantity is reduced. As described above, DRS eliminates the cDNA synthesis stage and requires only a few femtomoles of RNAs containing natural poly(A) tails or RNAs polyadenylated *in vitro*. It is also conceivable that microfluidic capabilities could be combined with DRS for single-cell applications (FIG. 5a).

Hybridization-based methodologies are also providing promise for working with very small quantities of RNA. The NanoString nCounter System provides an alternative method for RNA quantification without the requirement for cDNA synthesis, and it relies on the generation of target-specific probes (FIG. 5b). The probe mixture is hybridized to RNA samples in solution, followed by the immobilization of probe–RNA duplexes on surfaces and single-molecule imaging to identify and count individual transcripts[99]. In principle, the system can detect up to 16,384 transcripts simultaneously. This approach requires ~100 nanograms of RNA or 2000–5,000 cells[100], but optimization of the probe hybridization and surface immobilization steps may further reduce input RNA quantity.

Laser capture microdissection
(Often abbreviated to LCM.) A method allowing cells of interest that are chosen by the operator using a microscope to be specifically captured from heterogeneous tissue samples. The isolated cells can be used for various analyses including of protein and nucleic acid.

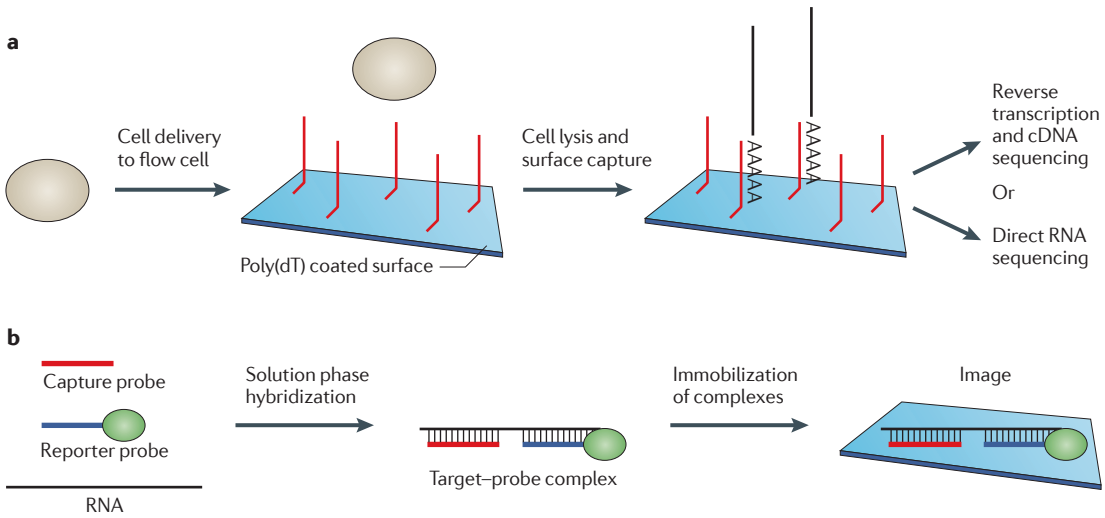


Figure 5 | **Emerging technologies for single-cell or low-quantity-cell gene expression profiling.** **a** | Single-molecule DNA and RNA sequencing technologies could be modified for single-cell applications. Cells can be delivered to flow cells using fluidics systems, followed by cell lysis and capture of mRNA species on the poly(dT)-coated sequencing surfaces by hybridization. Standard sequencing runs could take place on channels with a 127.5 mm^2 surface area, requiring 2,750 images to be taken per cycle to image the entire channel area. The surface area needed to accommodate ~350,000 mRNA molecules contained in a single cell is ~0.4 mm^2; thus, only eight images per cycle would be needed. Sequence analysis can be done with direct RNA sequencing (DRS)[7] or on-surface cDNA synthesis followed by single-molecule DNA sequencing[26]. **b** | Counter system workflow. Two probes are used for each target site: the capture probe (shown in red) contains a target-specific sequence and a modification that allows the immobilization of the molecules on a surface; the reporter probe contains a different target-specific sequence (shown in blue) and a fluorescent barcode (shown by a green circle) that is unique to each target being examined. After hybridization of the capture and reporter probe mixture to RNA samples in solution, excess probes are removed. The hybridized RNA duplexes are then immobilized on a surface and imaged to identify and count each transcript with the unique fluorescent signals on the capture and reporter probes.

Fluidigm offers a microfluidics platform that can perform quantitative real-time polymerase chain reaction (qRT-PCR) experiments on gene panels in a multiplexed manner and has been used to profile single cells. Commercial kits allowing one-step cDNA synthesis and amplification are used for cell lysis, cDNA synthesis and PCR amplification of the transcript region of interest. Pre-amplified cDNAs are then introduced to the Fluidigm Dynamic Array for qRT-PCR analysis. This approach may be useful for the determination of the expression levels of a subset of transcripts across cells of interest[101,102].

None of the approaches described above is mature, and none so far fully addresses our need for reliable, genome-wide and in-depth transcriptome profiles from minute cell quantities. For example, both the Fluidigm and NanoString technologies interrogate only a selected subset of transcripts and do not provide comprehensive analyses. However, it is hoped that future advances that will arise from the foundation formed by these technologies will enable such capabilities.

Future perspectives

Recent advances in RNA-seq have provided researchers with a powerful toolbox for the characterization and quantification of the transcriptome. Emerging sequencing technologies promise to at least partly alleviate the difficulties of current RNA-seq methods and equip scientists with better tools. Using these technological advances, we can build a complete catalogue of transcripts that are derived from genomes ranging from those of simple unicellular organisms to complex mammalian cells, as well as in tissues in normal and disease states. Furthermore, with our increasing ability to work with minute RNA quantities from fresh and formalin-fixed paraffin-embedded tissues and cells, and to provide quantification of RNA species from even single cells, we have the opportunity to define complex biological networks in a wide range of biological specimens. With these networks in hand, we can use data-driven RNA network models of cells and tissues in an attempt to fully understand the biological pathways that are active in various physiological conditions. In addition, these technologies are bringing us closer to the ability to use RNA measurements for clinical diagnostics. For example, analysis of circulating extracellular nucleic acid[103] and cells, such as fetal RNA and circulating tumour cells, with these new technologies may allow for earlier assessment of health, disease recurrence or mutational status. Thus, these technologies will continue to help us realize the full potential of genomic information as it relates to basic biological questions of differentiation and diversity, as well as its growing impact on the personalization of healthcare.

Quantitative real-time polymerase chain reaction
A PCR application that enables the measurement of nucleic acid quantities in samples. Nucleic acid of interest is amplified with PCR. The level of the amplified product accumulation during PCR cycles are measured in real time. This data is used to infer starting nucleic acid quantities.

Circulating extracellular nucleic acid
Extracellular DNA or RNA molecules in plasma and serum.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Berretta, J. & Morillon, A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**, 973–982 (2009).
3. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
4. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
This Review provides a comprehensive overview of currently available and in-development NGS technologies.
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
6. van Vliet, A. H. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol. Lett.* **302**, 1–7 (2010).
7. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
The first technology for high-throughput direct sequencing of RNA molecules without prior reverse transcription.
8. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
9. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
10. Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255–265 (2009).
11. Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods* **7**, 521–527 (2010).
12. Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Methods* **7**, 528–534 (2010).
13. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).
14. Ozsolak, F. *et al.* Chromatin structure analyses identify miRNA promoters. *Genes Dev.* **22**, 3172–3183 (2008).
15. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
This paper raises the possibility of 5′-cap addition during post-transcriptional processing steps.
16. Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nature Rev. Mol. Cell Biol.* **10**, 637–643 (2009).
An excellent review of the literature on sense and antisense transcription.
17. Gubler, U. Second-strand cDNA synthesis: mRNA fragments as primers. *Meth. Enzymol.* **152**, 330–335 (1987).
18. Perocchi, F., Xu, Z., Clauder-Munster, S. & Steinmetz, L. M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* **35**, e128 (2007).
19. Spiegelman, S. *et al.* DNA-directed DNA polymerase activity in oncogenic RNA viruses. *Nature* **227**, 1029–1031 (1970).
20. Wu, J. Q. *et al.* Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol.* **9**, R3 (2008).
21. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613–619 (2008).
22. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
23. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature Methods* **7**, 130–132 (2010).
24. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnol.* **27**, 652–658 (2009).
25. Ozsolak, F. *et al.* Digital transcriptome profiling from attomole-level RNA samples. *Genome Res.* **20**, 519–525 (2010).
26. Ozsolak, F. *et al.* Amplification-free digital gene expression profiling from minute cell quantities. *Nature Methods* **7**, 619–621 (2010).
27. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
28. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
29. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
30. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**, 709–715 (2010).
31. Faulhammer, D., Lipton, R. J. & Landweber, L. F. Fidelity of enzymatic ligation for DNA computing. *J. Comput. Biol.* **7**, 839–848 (2000).
32. Housby, J. N. & Southern, E. M. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.* **26**, 4259–4266 (1998).
33. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
34. Goren, A. *et al.* Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nature Methods* **7**, 47–49 (2010).
35. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. *Nature Methods* **6**, 291–295 (2009).
36. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
37. Wang, G. S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Rev. Genet.* **8**, 749–761 (2007).
38. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621–628 (2008).
39. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
40. Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).
41. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
42. Carninci, P. Is sequencing enlightenment ending the dark age of the transcriptome? *Nature Methods* **6**, 711–713 (2009).
43. Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-seq. *Bioinformatics* **25**, 1026–1032 (2009).
44. Richard, H. *et al.* Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Res.* **38**, e112 (2010).
45. Ameur, A., Wetterbom, A., Feuk, L. & Gyllensten, U. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.* **11**, R34 (2010).
46. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
47. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
48. Olasagasti, F. *et al.* Replication of individual DNA molecules under electronic control using a protein nanopore. *Nature Nanotech.* **5**, 798–806 (2010).
49. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature Rev. Cancer* **7**, 233–245 (2007).
50. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
51. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
52. Zhao, Q. *et al.* Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc. Natl Acad. Sci. USA* **106**, 1886–1891 (2009).
53. Li, H., Wang, J., Mor, G. & Sklar, J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* **321**, 1357–1361 (2008).
54. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **106**, 12353–12358 (2009).
55. Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).
56. Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nature Med.* **16**, 793–798 (2010).
57. McManus, C. J., Duff, M. O., Eipper-Mains, J. & Graveley, B. R. Global analysis of trans-splicing in *Drosophila*. *Proc. Natl Acad. Sci. USA* **107**, 12975–12979 (2010).
58. Garber, K. Fixing the front end. *Nature Biotech.*. **26**, 1101–1104 (2008).
59. Chepelev, I., Wei, G., Tang, Q. & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-seq. *Nucleic Acids Res.* **37**, e106 (2009).
60. Cirulli, E. T. *et al.* Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.* **11** (2010).
61. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
62. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10**, R115 (2009).
63. Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
64. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods* **6**, 613–618 (2009).
65. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature Rev. Genet.* **10**, 94–108 (2009).
66. Rajagopalan, R., Vaucheret, H., Trejo, J. & Bartel, D. P. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* **20**, 3407–3425 (2006).
67. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
68. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
69. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41**, 572–578 (2009).
70. Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nature Genet.* **38**, 1375–1377 (2006).
71. Kapranov, P. *et al.* New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* **466**, 642–646 (2010).
72. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
73. Kawaji, H. & Hayashizaki, Y. Exploration of small RNAs. *PLoS Genet.* **4**, e22 (2008).
74. Linsen, S. E. *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods* **6**, 474–476 (2009).
The authors describe the difficulties associated with the analysis and quantification of short RNA species using current NGS platforms.
75. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127–131 (2006).
76. Mader, R. M. *et al.* Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J. Lab. Clin. Med.* **137**, 422–428 (2001).
77. Roy, S. W. & Irimia, M. When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *Bioessays* **30**, 601–605 (2008).
78. Chen, D. & Patton, J. T. Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5′-RACE and primer extension. *Biotechniques* **30**, 574–582 (2001).
79. Roberts, J. D. *et al.* Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis *in vitro*. *Mol. Cell. Biol.* **9**, 469–476 (1989).
80. Armour, C. D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods* **6**, 647–649 (2009).

81. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131 (2010).
82. Rosenkranz, R., Borodina, T., Lehrach, H. & Himmelbauer, H. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics* **92**, 187–194 (2008).
83. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4**, 14 (2009).
84. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
85. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
86. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotech.* **28**, 503–510 (2010).
87. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech.* **28**, 511–515 (2010).
88. Shcherbik, N., Wang, M., Lapik, Y. R., Srivastava, L. & Pestov, D. G. Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells. *EMBO Rep.* **11**, 106–111 (2010).
89. Makeyev, E. V. & Bamford, D. H. Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage phi6. *EMBO J.* **19**, 124–133 (2000).
90. Gurumurthy, S. *et al.* The Lkb1 metabolic sensor maintains haematopoietic stem cell survival. *Nature* **468**, 659–63 (2010).
91. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018–1029 (2010).
92. Simone, N. L., Bonner, R. F., Gillespie, J. W., Emmert-Buck, M. R. & Liotta, L. A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).
93. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci.* USA **104**, 11889–11894 (2007).
94. Pfeifer, G. P., Steigerwald, S. D., Mueller, P. R., Wold, B. & Riggs, A. D. Genomic sequencing and methylation analysis by ligation mediated PCR. *Science* **246**, 810–813 (1989).
95. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
96. Dafforn, A. *et al.* Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis. *Biotechniques* **37**, 854–857 (2004).
97. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA* **89**, 3010–3014 (1992).
98. Nygaard, V. & Hovig, E. Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling. *Nucleic Acids Res.* **34**, 996–1014 (2006).
This review provides a good overview of the current low-quantity RNA applications and the complications associated with them.
99. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotech.* **26**, 317–325 (2008).
100. Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263 (2009).
101. Byrne, J. A., Nguyen, H. N. & Reijo Pera, R. A. Enhanced generation of induced pluripotent stem cells from a subpopulation of human fibroblasts. *PLoS ONE* **4**, e7118 (2009).
102. Helzer, K. T. *et al.* Circulating tumor cells are transcriptionally similar to the primary tumor in a murine prostate model. *Cancer Res.* **69**, 7860–7866 (2009).
103. Lo, Y. M. *et al.* Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. *Nature Med.* **13**, 218–223 (2007).
This paper describes the quantification of extracellular circulating RNA in mother's plasma during pregnancy to detect fetal aneuploidy.
104. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593–595 (2009).

Acknowledgements
We apologize to authors whose work could not be cited owing to space constraints. We are grateful to the US National Human Genome Research Institute for their support (grants R01 HG005230 and R44 HG005279).

Competing interests statement
The authors declare competing financial interests; see Web version for details.

FURTHER INFORMATION
Fatih Ozsolak and Patrice M. Milos's homepage (Helicos BioSciences website): www.helicosbio.com
Helicos Technology Center: http://open.helicosbio.com
The University of California Santa Cruz Genome Browser: http://genome.ucsc.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF