Cell. Mol. Life Sci. (2010) 67:569–579
DOI 10.1007/s00018-009-0180-6

Cellular and Molecular Life Sciences

REVIEW

RNA-seq: from technology to biology

Samuel Marguerat · Jürg Bähler

Received: 23 July 2009 / Revised: 11 September 2009 / Accepted: 8 October 2009 / Published online: 27 October 2009

Abstract Next-generation sequencing technologies are now being exploited not only to analyse static genomes, but also dynamic transcriptomes in an approach termed RNA-seq. Although these powerful and rapidly evolving technologies have only been available for a couple of years, they are already making substantial contributions to our understanding of genome expression and regulation. Here, we briefly describe technical issues accompanying RNA-seq data generation and analysis, highlighting differences to array-based approaches. We then review recent biological insight gained from applying RNA-seq and related approaches to deeply sample transcriptomes in different cell types or physiological conditions. These approaches are providing fascinating information about transcriptional and post-transcriptional gene regulation, and they are also giving unique insight into the richness of transcript structures and processing on a global scale and at unprecedented resolution.

Keywords High-throughput sequencing · Transcriptional control · Non-coding RNA · Post-transcriptional control · Gene expression · Splicing · Transcriptome · Genome

Introduction

Regulation of gene expression is fundamental to link genotypes with phenotypes. The synthesis and maturation of RNAs are tightly controlled, and they shape complex gene expression networks that ultimately drive biological processes. These networks need to be robust as well as highly plastic in order to allow rapid adaptation to environmental or genetic perturbations [1]. An in-depth understanding of the principles and mechanisms governing these complex gene expression programmes is important to better understand complex diseases such as cancer. For more than 10 years, microarrays have allowed the simultaneous monitoring of expression levels of all annotated genes in cell populations [2, 3]. The ability to analyse entire gene expression programmes has opened new horizons for our understanding of global processes regulating gene expression. Similarly, with the increasing realisation that RNAs transcribed from non-coding portions of genomes are playing fundamental roles, genome-wide approaches have provided valuable insights into this aspect of transcriptomes. Later generations of microarrays (referred to as "tiling arrays"), which consist of probes designed to interrogate a genome systematically irrespective of any gene annotation, have been instrumental in discovering unknown transcripts [4]. Applying this technique to several different organisms has demonstrated that the complexity of transcriptomes has indeed been vastly underestimated [5]. This is when next-generation sequencers have entered the market. These platforms allow the rapid and cost-effective generation of massive amounts of sequence data. Obviously, this breakthrough provides a huge potential to revolutionise the field of transcriptomics. Even though direct sequencing of cDNA libraries has been achieved before with SAGE [6] and MPSS [7] approaches, next-generation sequencing (NGS) technologies are more straightforward and more affordable. RNA-seq was thus born [8–11].

In this review, we will first provide an overview of the strengths and challenges inherent to RNA-seq and will then

S. Marguerat · J. Bähler (✉)
Department of Genetics, Evolution and Environment,
UCL Cancer Institute, University College London,
Darwin Building, Gower Street, London WC1E 6BT, UK
e-mail: j.bahler@ucl.ac.uk

highlight major biological insights gained from RNA-seq in a wide range of organisms.

RNA-seq data generation and analysis

The NGS market is currently dominated by three different platforms: the FLX pyrosequencing system from 454 Life Sciences (a Roche company), the Illumina Genome Analyser (developed initially by Solexa), and the AB SOLiD system (now Life Technologies). On all three platforms, DNA fragments are sequenced in parallel, producing large numbers of relatively short sequence "reads" or "tags". The throughput varies from hundreds of thousands of reads for the FLX system to hundreds of millions of reads for the Illumina Genome Analyser and AB SOLiD systems. Read lengths range from 30–100 bp for Illumina and SOLiD to 200–500 bp for FLX. It is important to note that these technologies are evolving at a tremendous pace, with ever-increasing numbers and lengths of sequence reads. The three major systems differ significantly in the approaches used to produce massive amounts of sequences. An in-depth discussion of the technical and methodological aspects of these next-generation sequencers is beyond the scope of this review and can be found elsewhere [12, 13]. Despite their technological differences, the three major platforms rely on similar work flows for the production and analysis of sequencing libraries (Fig. 1). First, the sample nucleic acids have to be sheared in order to reach a size compatible with sequencing (typically <500 bp). Second, DNA adapters containing unique sequences are attached at both ends of the sheared DNA molecules. These adapters subsequently allow the DNA fragments to be singled out, either on beads or on a slide ("flowcell"), to then be sequenced in parallel.

The library preparation is a key step of RNA-seq, because it determines how closely the cDNA sequence data reflect the original RNA population. In the classic NGS protocols, which have been developed for the analysis of genomic DNA, adapters are ligated onto shared double-stranded DNA fragments. In order to allow the analysis of transcriptomes by NGS, these protocols have been adapted to the sequencing of cDNA. The most straightforward approach is to simply synthesise double-stranded cDNA, to which the adapter can then be ligated. This robust protocol has been attractive, because it applies the procedures developed by the manufacturer for the analysis of genomic DNA, and it has been widely used in the original RNA-seq studies. A substantial drawback of this approach, however, is the loss of information on transcriptional direction, because the adaptor is ligated to double-stranded cDNA. An elegant study has managed to maintain strand information simply by pre-treating the RNA samples with sodium bisulphate [14]. This chemical triggers the transformation of cytidine into uridine; widespread C–T transition therefore "marks" the coding strand of each transcript. Six additional RNA-seq protocols that maintain strand-specificity have been published. They differ in how the adaptor sequences are inserted into the cDNA, which is achieved (1) by direct ligation of RNA adaptors to the RNA sample before reverse transcription [15, 16], (2) by addition of the adaptor sequences by template switch during reverse transcription [17], (3) by double-random priming coupled to solid phase extraction [18], (4) by direct ligation of the DNA adaptors to single-stranded cDNA [19–21], (5) by reverse transcription of in vitro polyadenylated RNA fragments followed by intramolecular ligation [22], or (6) by incorporation of dUTP during second strand synthesis and digestion with uracil-*N*-glycosylase [23]. These methods are likely to differ in potential biases introduced in the data, and careful comparisons will be highly interesting.

NGS technologies exploit light that is emitted when the correct base (or oligonucleotides in case of SOLiD) matches the template being sequenced and is incorporated into the sequencing reaction. Thus, NGS raw outputs are image records of the light emitted by every single parallel sequencing reaction at every sequencing cycle. These raw image files represent terabytes of data and require substantial storage resources. The images are then processed in order to extract numerical signals for every base at every synthesis event from all the parallel reactions. These signals are used for base calling. Improving the quality and reliability of signal extraction and base calling has led to significant increases in the quality and throughput of NGS data [24–26].

After image and signal processing, NGS data consist of a list of short sequences together with their base call qualities. These data are fundamentally different from microarray data. With hybridisation-based techniques, the scanner returns signal intensities for each probe on the array. In the case of RNA-seq data, the number of reads mapping to any given region of the genome makes up the signal. Besides providing single base pair resolution, sequencing allows the maintaining of total control on which reads are included in the final analysis and hence contribute to the expression signals. Thus, RNA-seq data are countable and digital in nature. The generation of reliable RNA-seq data therefore relies heavily on proper mapping of sequencing reads to corresponding reference genomes or on their efficient de novo assembly. Mapping NGS reads with high efficiency and reliability currently faces several challenges. First, the computing resources required to map huge numbers of small reads within a reasonable time can be limiting. However, tremendous effort has been invested during the last couple of years to

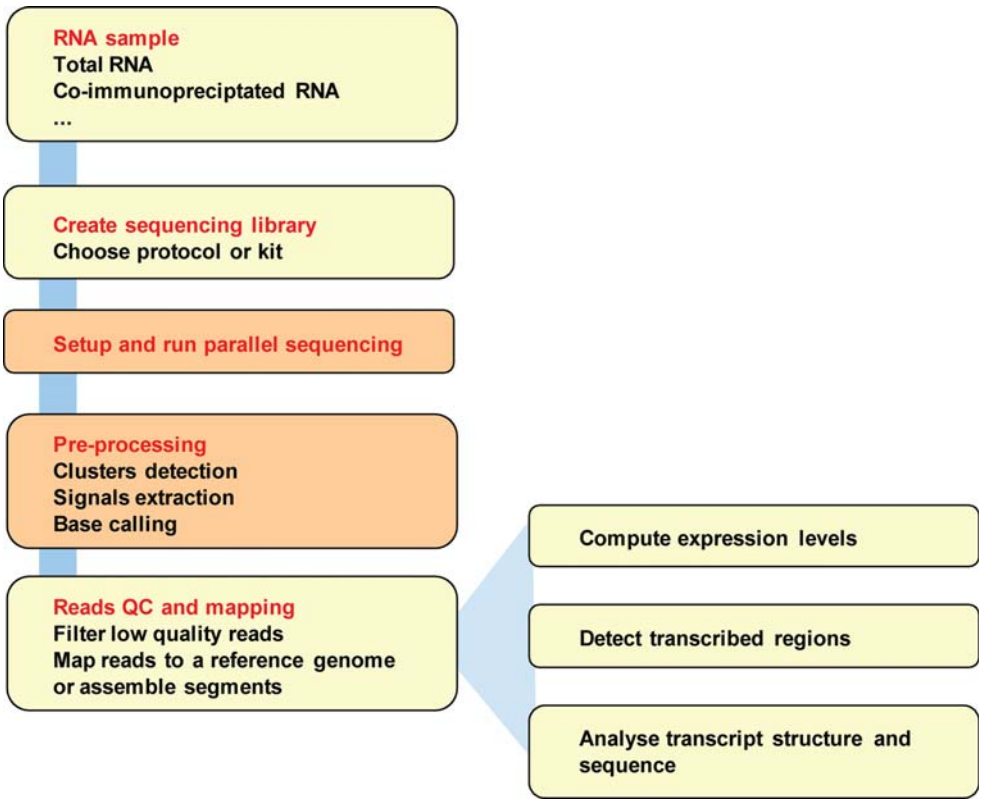

Fig. 1 Flowchart of a typical RNA-seq experiment

develop algorithms that allow mapping of millions of small reads using limited computing resources and time [27–33]. The second challenge arises from the relatively high error rate of NGS data, meaning that non-perfect matches have to be considered when mapping reads back to a genome. This issue is particularly relevant when single nucleotide polymorphisms (SNPs) are of interest to detect allele-specific expression in RNA-seq data. To distinguish sequencing errors from SNPs requires higher sequencing depths such that correct base calls at each position can be made, even in heterozygous samples, because each base is sequenced multiple times. Analysis protocols have been developed for the detection of genetic variation at a reasonable sequencing depth and hence at affordable costs [34]. Library preparation and/or sequencing procedures can also introduce systematic biases and artefacts such as over-amplification of GC-rich regions and generation of duplicate sequences [35]. A third challenge, which is also one of the most exciting feature of RNA-seq data, is to identify reads containing post-transcriptionally modified or rearranged sequences which cannot be mapped directly to the reference genome. This feature will be discussed in more detail below. Finally, for cases when no good quality reference genome is available, direct de novo assembly of RNA-seq data into contigs may be useful. Several assemblers optimised for short sequence reads have been recently developed [36–45].

Once the sequencing reads have been filtered and mapped (or assembled), it is possible to compute an expression score for every base in the genome and thus obtain transcriptome maps at the best possible resolution. The true resolution of this approach, however, depends on the amount of sequence coverage and therefore on the amount of sequences generated. Sequence coverage can be a limiting factor, especially when large genomes are analysed, due to costs and machine time required.

Applying RNA-seq to probe the breadth and depth of genome transcription

The use of NGS technologies for the analysis of RNA has been pioneered by researchers working with small regulatory RNAs, possibly because this field has benefited less from microarrays as the usual size of small RNAs is too short to be captured adequately with the limited resolution provided by microarrays. Sequencing of short regulatory RNAs has resulted in important and exciting papers which has been extensively reviewed elsewhere [46, 47]. Whole transcriptome studies using RNA-seq have emerged soon after. To date, transcriptomes have been sequenced for over a dozen organisms including human [14, 16, 18–20, 48–55], mouse [17, 23, 56–58], budding yeast [22, 23, 59–62], fission yeast [63], worm [64], fruit fly [65],

non-model organisms [66, 67], several plants [15, 68–71] and prokaryotes [21, 72, 73]. Unlike the genome, the transcriptome dynamically changes in response to the environment or to intrinsic programmes, and many studies have reported transcriptome sequences for several cell types or physiological conditions.

The countable, almost digital, nature of RNA-seq data makes them particularly attractive for the quantitative analysis of transcript expression levels. Nearly every RNA-seq study published to date has addressed this question, and they agree that RNA-seq data are highly quantitative and give reliable measurements of transcript levels in one or more conditions. The dynamic range of these data is theoretically only limited by the sequencing depth and has been reported to span at least 5 orders of magnitude [58]. This dynamic range is well beyond the range achieved by microarrays and close to the estimated range of transcript frequencies in the cell. A few studies also looked at the ability of RNA-seq to measure differential gene expression [51, 57, 61]. These studies agree in saying that RNA-seq performs at least as well as microarrays provided an adequate sequencing depth. RNA-seq has the advantage though that, besides differential transcripts levels, levels of different splice variants or of transcripts with different UTR length can be assessed at the same time (see below). Producing enough reads for accurate quantification of lowly expressed transcripts, however, can still be quite expensive for large transcriptomes. In a variant of RNA-seq, only small tags at the 3′ ends of transcripts are sequenced. This assay permits the measurement of even lowly expressed transcripts with a limited amount of sequencing reads [57, 74].

Besides this quantitative aspect, RNA-seq studies are enabling researchers to refine transcript annotation, providing for instance accurate maps of transcript start and end sites. This feature is of particular help for dense prokaryotic genomes, allowing confident discrimination between single gene transcriptional units and operons encompassing several genes [72]. The analysis of transcript structures is also fundamental for the study of complex diseases such as cancer. Genomic re-arrangements or mutations can generate aberrant fusion transcripts which, if stably expressed, can lead to pathologies. Such gene fusions have been shown to be commonly associated with different types of tumours [75]. Direct sequencing of transcriptomes, coupled with analysis pipelines allowing the detection of sequence re-arrangements and abnormal transcript structures, are powerful tools which permit direct detection of such fusion events. Several studies have already provided proofs of principle that this approach is suitable for discovering new aberrant transcripts [19, 50]. Thus, this technological breakthrough will hopefully fuel our understanding of complex diseases.

Another characteristic of RNA-seq data is their high sensitivity, allowing the detection of the expression of substantially more transcripts in a given cell type compared to what could be detected by microarrays. RNA-seq studies also contribute to an increased list of the transcripts expressed in all organisms studied, most of these newly defined transcripts being non-coding. A high coverage RNA-seq study of the fission yeast (*Schizosaccharomyces pombe*) transcriptome during vegetative growth revealed that over 94% of this genome is actively transcribed at some level, including genes required only under specialised physiological conditions [63]. This finding could reflect a small percentage of cells in the population expressing a different transcriptional programme [72], or it could reflect a certain amount of basal background transcription. The latter would be compatible with the suggestion that as much as 90% of all RNA Polymerase II (Pol II) initiation events represent transcriptional noise and raises the question of the biological relevance of an almost ubiquitous noisy transcription [76].

RNA-seq has also been used to dig deep into eukaryotic transcriptomes and reveal an intriguing new feature of eukaryotic transcription at promoters. Cryptic unstable transcripts (CUTs) are small RNA Pol II transcripts found in the budding yeast (*Saccharomyces cerevisiae)* which are targeted for degradation by the exosome complex immediately after synthesis [77]. While the mechanisms regulating their processing have been extensively studied, the prevalence of CUTs in the yeast genome has remained unknown. Two studies have determined the genome-wide distributions and structures of CUTs [78, 79], using NGS to sequence a SAGE library enriched for CUTs or high-density tiling arrays, respectively. Interestingly, CUTs seem to be well-defined transcriptional units arising mostly from nucleosome-free regions (NFRs). NFRs are characteristic of eukaryotic genomes and can be found mainly in the promoters and terminators of genes [80]. A fraction of CUTs are overlapping the 5′ ends of genes, suggesting a potential regulatory function. However, CUTs are most frequently transcribed in divergent orientation from the promoters of genes, suggesting that they could be by-products of Pol II-dependent transcription [78, 79]. These data suggest that bidirectional transcription is a widespread characteristic of eukaryotic promoters. In budding yeast, stable transcripts arising from bidirectional transcription can also be detected, suggesting that this phenomenon is not restricted to cryptic transcripts [79]. Interestingly, these transcripts show extensive overlaps with annotated genes. A possible regulatory role of bidirectional transcription remains to be determined, but some data suggest that divergent transcripts could act as transcriptional "links" between neighbouring genes and potentially regulate their co-expression [79]. Bidirectional transcription seems to be

a conserved characteristic as it can also be detected in multicellular eukaryotes. Transcripts similar to yeast CUTs have been detected after inactivation of the exosome in human cells. These so-called "promoter upstream transcripts" (PROMTs) are mostly transcribed from promoters of active genes in both directions [81]. As in yeast, stable transcripts mapping to both strands of promoters can also be detected in metazoans [16, 82–84]. A similar class of short transcripts, 20–90 nucleotides in length, has been found in mouse ES cells, up- and downstream of the transcription start sites (TSS) [82]. Interestingly, these short divergent transcripts are not enriched in terminator or intergenic regions. Analysis of histone marks around these transcripts has revealed that marks associated with transcription elongation are present on the gene sequences but not in the antisense direction, suggesting that productive elongation occurs mostly downstream of the TSS. In this context, it is possible that these short RNAs mark regions of Pol II pausing [82]. A similar picture could be detected in human fibroblasts where nascent RNAs have been sequenced using NGS technology, providing an overview of the distribution of Pol II engaged in transcription at a given time [16]. This study concludes that a large amount of Pol II is paused shortly after initiation. In addition, engaged Pol II has been detected in divergent direction relative to genes. However, the lack of sequencing reads further upstream indicates that divergent Pol II does not productively elongate transcripts [16]. These findings suggest that regulation of transcript elongation participates in the control of gene expression. In summary, bidirectional transcription at promoters seems to be a widespread phenomenon conserved across evolution. Further investigation will now be required to understand what portion of these divergent transcription events represents useless by-products of transcription initiation and what portion plays regulatory roles.

Applying RNA-seq to interrogate post-transcriptional gene regulation

Post-transcriptional regulation is a fundamental part of gene expression, which may well match transcriptional control in importance and sophistication. It includes the control of alternative splicing and polyadenylation, RNA editing, RNA degradation and translation. With the possible exception of translational control, these processes involve the modification of transcript sequences or structures. The sequences of the processed RNA molecules can therefore differ substantially from the corresponding genome sequences. Our understanding of the sequence motifs governing post-transcriptional control improves steadily but does not yet allow prediction of mRNA processing events based on the genomic sequence alone. Techniques allowing global characterisation of post-transcriptional sequence alterations and rearrangements are therefore required. High-density tiling arrays are only partially suited for the analysis of post-transcriptional structural changes as their probe design is unable to capture sequences that either are not encoded in the genome, as in the case of editing, or are not adjacent in the genome, as in the case of splicing. These limitations could in principle be circumvented by designing additional sets of probes for the array, but this requires high quality annotation. RNA-seq, on the other hand, is particularly well suited for the study of mRNA processing, as it generates transcript sequence data from a library independently of the organism's genome sequence. In case of RNA splicing, for instance, where tiling arrays require the design of special sets of probes, sequencing relies only on an appropriate mapping strategy able to retrieve reads containing non-adjacent sequences (Fig. 2a). Several strategies have been developed for this purpose. In one approach, the set of reads which does not map properly to the reference genome can successively be mapped against a reference sequence library containing all known or predicted exon–exon junctions. Sequencing reads mapping across exon–exon junctions (often called "trans-reads") are diagnostic for post-transcriptional rearrangements. While quite straightforward and flexible, this approach is limited when it comes to discovering new, un-annotated splice junctions. Alternatively, a reference sequence library of all possible splice junctions instead of all known splice junctions could be used for mapping. This approach would permit discovery of new splicing events. In another approach, sequencing reads are either mapped allowing gaps in the alignment or split in two before mapping both halves back separately to the reference genome. The reads, whose two halves do not map next to each other, point to a post-transcriptional rearrangement or splicing event. This approach is potentially extremely powerful as it does not rely on any genome annotation. However, it requires sufficiently long sequencing reads to be confidently mapped even if split in two. In addition to mapping the sites of post-transcriptional rearrangements, trans-reads provide a quantitative measurement of the levels of different transcript isoforms. Furthermore, the amount of trans-reads at a given exon–exon junction relative to the amount of reads spanning the corresponding exon–intron junctions provides a measure of the splicing efficiency at this junction. This feature has been exploited to sample splicing efficiencies across all introns and genes under different conditions in fission yeast [63]. A fourth strategy takes advantage of so-called paired-end sequencing. NGS sequencers have been up-graded to allow sequencing both ends of each DNA fragment in the library. In this case, the data consist of two sequencing reads per DNA fragment. The distance between the two reads is

known as it equals the fragment size of the library. This development has been critical for making it much easier, for example, to map short reads in low complexity regions [85]. For the analysis of post-transcriptional rearrangements by RNA-seq, the paired reads that map much closer or farther apart to each other than the insert size of the library can point to rearrangements. While being compatible with short reads and not relying on any prior knowledge about the regulatory motifs or genome annotation, this fourth approach does not provide direct base pair mapping of the junction. An advantage of the first three strategies described above is that the exact splice junction or rearrangement point coordinates can be identified.

Analysis of alternative splicing by RNA-seq has been performed recently on several human tissues [48, 49, 56] and cell lines [48, 55]. The ability to globally sample every possible splice isoform has uncovered a much larger amount of alternative splicing in human tissues than previously estimated. Considering different tissues, as many as 95% of the human multi-exon genes have been found to undergo alternative splicing, with exon skipping being the most frequent form of regulation [48, 49]. These results considerably increase previous estimates, which have suggested that about two-thirds of human genes are differentially spliced [86]. Importantly, for 92% of genes, the second most frequent isoform has a relative frequency above 15%, indicating that in most cases several isoforms of the same transcript reach substantial levels of expression [48]. Isoforms differ mostly between tissues, while between individual variations are two- to threefold less common [48]. This finding indicates that tissue specific alternative splicing is an almost universal mode of tissue-specific gene regulation. Extreme "switch-like" behaviours, where two isoforms are mutually exclusive in two distinct tissues, have also been detected [48]. In these cases, alternative splicing can produce different proteins in different contexts. Interestingly, "switch-like" exons are characterised by conserved regulatory motifs [48]. Different spliced isoforms can also occur together in the same tissues. An interesting study has applied RNA-seq to analyse the transcriptome of single mouse cells [56]. The authors report 335 genes that display multiple isoforms in a single blastomere, indicating that alternative splicing can also increase the diversity of the transcriptome of a single cell during embryonic development. Similar analyses performed in fission and budding yeasts have provided interesting insights into how simpler unicellular eukaryotes exploit alternative splicing as a mode of post-transcriptional regulation [59, 63]. In fission yeast, intron retention seems to be the main event detected during sexual differentiation. This finding has confirmed and extended observations from smaller-scale studies [87]. In addition, global splicing efficiencies and transcript expression levels seem to be positively correlated during vegetative growth and sexual differentiation, suggesting coordination between transcription and splicing [63]. A recent RNA-seq study in budding yeast has uncovered many alternative isoforms showing differential expression between vegetative growth and response to heat-shock [60]. Interestingly, some of these isoforms are possibly coding for proteins of

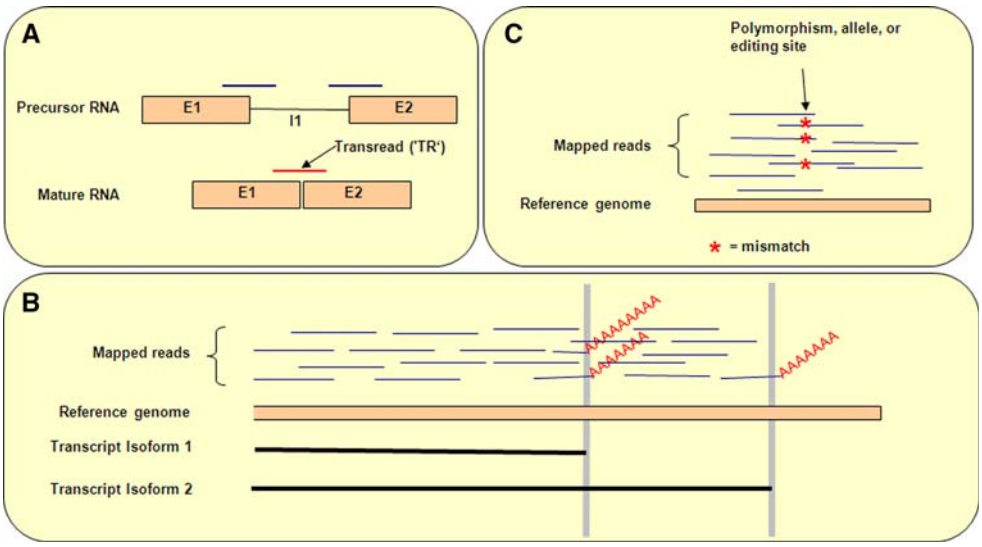

Fig. 2 Detection of post-transcriptional modifications and rearrangements by RNA-seq. **a** Reads spanning exon–exon junctions give positive evidence for splicing events (trans-reads in *red*). Comparing the number of trans-reads for a selected junction to the number of reads spanning its corresponding exon–intron junctions (*blue*) gives a measure of splicing efficiency. **b** Reads containing poly(A) tracts which are not encoded in the reference genome are diagnostic of polyadenylation events. **c** Reads containing sequence polymorphisms compared with the reference genome are potential polymorphisms or editing sites

different lengths. Taken together, these data show that regulation of splicing is also used by unicellular eukaryotes to control and diversify gene expression. Finally, bioinformatics tools helping to extract the respective expression levels of different transcript isoforms from RNA-seq data are becoming available and will help to refine the global picture of alternative splicing in eukaryotes [88, 89].

A related mechanism by which transcript diversity can be increased is the use of alternative polyadenylation sites. RNA-seq is particularly well suited to study polyadenylation as it allows direct sequencing of the junctions between poly(A) tails and the rest of the transcript (Fig. 2b). This approach permits the disentangling of several isoforms with alternative polyadenylation sites in a single sample. For example, human cells show a strong correlation of alternative splicing and alternative polyadenylation between tissues, suggesting coordination between these two processes [48]. Interestingly, alternative introns and 3′ untranslated regions (UTRs) are sharing common regulatory motifs, suggesting that they also share regulatory factors [48].

Transcriptome diversity can also be increased by editing of mRNA transcripts. This process involves deamination of adenosines into inosines, which are then read as guanosines. Editing is critical for brain function in mammals and linked to several diseases [90]. However, the extent of this phenomenon has remained elusive. Direct sequencing of transcriptomes is the method of choice to understand how prevalent is this mode of post-transcriptional regulation (Fig. 2c). Indeed, a pioneering RNA-seq analysis of human brain and other tissues has revealed hundreds of new editing sites, many of which are located in non-coding RNAs [91].

Information about protein–RNA interactions is fundamental for the understanding of regulatory networks governing the different layers of post-transcriptional control. Predicting protein–RNA binding sites is difficult not least due to the relatively low sequence conservation of RNA binding motifs. Protein–RNA interactions can be mapped directly, however, using approaches similar the chromatin immunoprecipitation technique used to identify protein–DNA interactions [92]. This approach is achieved in two ways: (1) RNA-binding proteins are immunoprecipitated together with their intact target transcripts (RIP) [93], or (2) RNA-binding proteins are crosslinked to the RNAs they interact with and treated with RNAse before immunoprecipitation (CLIP for crosslinking immunoprecipitation) [94]. This second approach limits the analysis to RNA fragments protected by the binding protein and is reminiscent of a footprint. The immunoprecipitated RNAs need eventually to be identified using either single-gene [94] or genome-wide methods [95]. NGS technologies have been successfully applied to these approaches. Several CLIP-seq (also called HITS-CLIP, for high-throughput sequencing CLIP) studies have analysed the binding patterns of human splicing regulators in different cell types and tissues [96–98]. For example, analysis of the binding patterns of the neuron-specific splicing factor Nova has demonstrated that its binding to introns determines the outcome of alternative splicing while its binding to 3′-UTRs can regulate alternative polyadenylation [97]. RIP and CLIP-seq have also been used to characterise Ago-RNA complexes in mouse, human and fission yeast [99–101]. The Ago protein binds small RNAs to form a core RNA silencing complex. Sequencing the populations of microRNAs (miRNAs) and mRNAs bound to Ago proteins in the mouse brain has allowed direct identification of in vivo expressed miRNAs and their potential target transcripts [99]. RIP-seq with Ago has led to the discovery of a new class of small RNAs in humans, originating from small nucleolar RNAs (snoRNA) which can function like miRNAs [100].

Ribosomes are riboprotein complexes mediating the translation of RNA transcripts into proteins and are probably the most abundant RNA-binding proteins in the cell. Studying the amount and position of ribosomes bound to transcripts globally can provide important information about regulation of translation. To this end, total cellular RNA is fractionated based on the amount of associated ribosomes (“polysome profiling”) [102]. This technique has provided information on basic properties of the translation process. NGS technologies with their ability to detect the exact sequence of short RNA molecules have now enabled a transition from genome-wide polysome profiling to genome-wide ribosome foot-printing [22]. Similarly to the CLIP method outlined above, this approach is based on the isolation of short RNA fragments occupied by ribosomes and hence protected from degradation by an endonuclease. It permits not only the measurement of the number of ribosomes associated with different transcripts but their exact positions along the RNA molecules. This method, termed “ribosome profiling”, has been applied to budding yeast grown under two different physiological conditions [22]. The ability to detect the distribution of ribosomes on transcripts at maximum resolution has revealed that the density of ribosomes is not uniform across transcripts. All transcripts contain a region of constant length at their 5′ends showing a high density of ribosomes [22]. This observation could explain the previously published phenomenon that short transcripts tend to be much more densely packed with ribosomes than large transcripts [103, 104]. The amount of ribosomes found in introns and 3′-UTRs is less than 1% of the ribosome density seen in open reading frames (ORFs), indicating that retained introns are rarely translationally active. Moreover, many small ORFs (uORFs) are detected in the 5′-UTRs of genes,

but their functional relevance remains elusive. The ribosome density in these uORFs is significantly higher than in other regions of the 5′-UTRs, indicating that pervasive translation occurs upstream of the ORF [22]. Surprisingly, a substantial amount of these uORFs are using non-AUG start codons, thus unexpectedly increasing the scope of peptides that can be translated from a given transcript.

Conclusions and outlook

Next-generation sequencing technologies are revolutionising genomics research and beyond by enabling the much more rapid and cost-effective generation of massive amounts of sequences compared to traditional Sanger sequencing. This technological breakthrough provides an opportunity for regular research institutes and departments to engage in ambitious projects which so far have only been conceivable for large genome centers. The impact of NGS technologies for the analysis of gene regulation is particularly high. Within only two years, RNA-seq has reached a point where recent state-of-the-art technologies such as high-density tiling arrays look almost old fashioned. It looks likely that sequencing-based approaches will largely supersede hybridisation-based approaches within a few years. RNA-seq permits the sequencing and quantifying of transcriptomes at maximal resolution and dynamic range, independently of transcript size, and above all free from any preconception (or even knowledge) of the genomes they are derived from. RNA-seq has started to change the way we think about studying the complexity and dynamics of transcriptomes and genome regulation. Early RNA-seq studies have revealed more extensively expressed genomes and more complex transcriptomes than anticipated, thus giving insight into novel regulatory mechanisms. These pioneering studies have also uncovered rich and extensive post-transcriptional regulation of transcript structures and sequences.

RNA-seq will without doubt drive many more exciting discoveries within the next few years. For example, sequencing of RNA from complex samples containing more than one organism, either collected in the wild [105–108] or created in the laboratory, will ultimately provide information about transcriptome dynamics of living communities and interactions within ecosystems. On the other hand, sequencing of RNA from closely related species or members of a population will give insight into the processes linking transcriptome plasticity to phenotypic diversity and evolution. Given sufficient sequencing depth, RNA-seq analysis of cell populations adapting to changing environmental conditions could also reveal rare changes in transcript sequences that do not necessarily lead to an increase in fitness, thus helping to understand evolutionary mechanisms and dynamics. The main challenge for researchers is to creatively exploit the opportunities provided by those rapidly evolving technologies. Even more powerful sequencing approaches are already on the horizon. For example, “next-next-generation” sequencers such as the Helicos system, which can sequence millions of single molecules in parallel, are entering the market and seem to be suited to analyse RNA [109]. Truly, progress is limited mainly by our imagination, and exciting times are certainly ahead.

Acknowledgments We would like to thank Luis López-Maury, Rachel Imoberdorf, Vera Pancaldi, Martin Převorovský, and Brian Wilhelm for critical reading of the manuscript. Research in our laboratory is funded by Cancer Research UK and by PhenOxiGEn, an EU FP7 research project.

References

1. López-Maury L, Marguerat S, Bähler J (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat Rev Genet 9:583–593
2. Shalon D, Smith SJ, Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639–645
3. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW (1998) Microarrays: biotechnology’s discovery platform for functional genomics. Trends Biotechnol 16:301–306
4. Bertone P, Gerstein M, Snyder M (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res 13:259–274
5. Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487
7. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634
8. Lister R, Gregory BD, Ecker JR (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol 12:107–118
9. Marguerat S, Wilhelm BT, Bähler J (2008) Next-generation sequencing: applications beyond genomes. Biochem Soc Trans 36:1091–1096
10. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
11. Wilhelm BT, Landry J (2009) RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249–257
12. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

13. Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25:195–203
14. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322:1855–1857
15. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536
16. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848
17. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613–619
18. Li H, Lovci MT, Kwon Y, Rosenfeld MG, Fu X, Yeo GW (2008) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci USA 105:20179–20184
19. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458:97–101
20. Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman M, Taillon BE, Du L, Bouffard P, Kingsmore SF, Miller NA, Farmer AD, Jensen RV, Gullans SR, Bueno R (2008) Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci USA 105:3521–3526
21. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. PLoS Genet 5:e1000569
22. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223
23. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A (2009) ranscriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*
24. Quinlan AR, Stewart DA, Strömberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods 5:179–181
25. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F (2008) Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 9:431
26. Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, Zaranek AW, Abnizova I, Brown C (2009) Swift: primary data analysis for the *Illumina Solexa* sequencing platform. Bioinformatics 25:2194–2199
27. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
28. Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24:2431–2437
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
30. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
31. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111
32. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5:e1000386
33. Li R, Yu C, Li Y, Lam T, Yiu S, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079
35. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6:291–295
36. Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501
37. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942–2944
38. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17:1697–1706
39. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18:810–820
40. Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18:324–330
41. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18:802–809
42. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829
43. Bryant DW, Wong W, Mockler TC (2009) QSRA: a quality-value guided de novo short read assembler. BMC Bioinformatics 10:69
44. Birol I, Jackman SD, Nielsen C, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ (2009) De novo Transcriptome Assembly with ABySS. Bioinformatics (in press)
45. Schmidt B, Sinha R, Beresford-Smith B, Puglisi SJ (2009) A fast hybrid short read fragment assembly algorithm. Bioinformatics 25:2279–2280
46. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655
47. Naqvi AR, Islam MN, Choudhury NR, Haq QMR (2009) The fascinating world of RNA interference. Int J Biol Sci 5:97–117
48. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476
49. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415
50. Zhao Q, Caballero OL, Levy S, Stevenson BJ, Iseli C, de Souza SJ, Galante PA, Busam D, Leversha MA, Chadalavada K, Rogers Y, Venter JC, Simpson AJG, Strausberg RL (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci USA 106:1886–1891

51. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517
52. Guffanti A, Iacono M, Pelucchi P, Kim N, Soldà G, Croft LJ, Taft RJ, Rizzi E, Askarian-Amiri M, Bonnal RJ, Callari M, Mignone F, Pesole G, Bertalot G, Bernardi LR, Albertini A, Lee C, Mattick JS, Zucchi I, De Bellis G (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 10:163
53. Wu Q, Kim YC, Lu J, Xuan Z, Chen J, Zheng Y, Zhou T, Zhang MQ, Wu C, Wang SM (2008) Poly A—transcripts expressed in HeLa cells. PLoS ONE 3:e2803
54. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques 45:81–94
55. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960
56. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382
57. 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, Boer JM, van Ommen GB, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36:e141
58. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
59. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349
60. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, Nusbaum C, Thompson D, Friedman N, Regev A (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci USA 106:3264–3269
61. Bloom J, Khan Z, Kruglyak L, Singh M, Caudy A (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 10:221
62. Lee A, Hansen KD, Bullard J, Dudoit S, Sherlock G (2008) Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. PLoS Genet 4:e1000299
63. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239–1243
64. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res 19:657–666
65. Torres TT, Metta M, Ottenwälder B, Schlötterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Res 18:172–177
66. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009) Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. BMC Genomics 10:234
67. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17:1636–1647
68. Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17:69–73
69. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910–918
70. Denoeud F, Aury J, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9:R175
71. Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J 7:334–346
72. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH (2009) Structure and complexity of a bacterial transcriptome. J Bacteriol 191:3203–3211
73. Mao C, Evans C, Jensen RV, Sobral BW (2008) Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system. BMC Microbiol 8:72
74. Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S, Zhao Y, Hirst M & Marra MA (2009) Next-generation tag sequencing for cancer gene expression profiling. *Genome Res*
75. Mitelman F, Johansson B, Mertens F (2007) The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7:233–245
76. Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14:103–105
77. Wyers F, Rougemaille M, Badis G, Rousselle J, Dufour M, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, Libri D, Jacquier A (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121:725–737
78. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature 457:1038–1042
79. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature 457:1033–1037
80. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10:161–172
81. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851–1854
82. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA (2008) Divergent transcription from active promoters. Science 322:1849–1851
83. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488

84. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest ARR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando V, Carninci P, Hayashizaki Y, Mattick JS (2009) Tiny RNAs associated with transcription start sites in animals. Nat Genet 41:572–578
85. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426
86. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–2144
87. Averbeck N, Sunder S, Sample N, Wise JA, Leatherwood J (2005) Negative control contributes to an extensive program of meiotic splicing in fission yeast. Mol Cell 18:491–498
88. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25:1026–1032
89. Zheng S, Chen L (2009) A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res 37:e75
90. Maas S, Kawahara Y, Tamburro KM, Nishikura K (2006) A-to-I RNA editing and human disease. RNA Biol 3:1–9
91. Li JB, Levanon EY, Yoon J, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 324:1210–1213
92. Kuo MH, Allis CD (1999) In vivo cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. Methods 19:425–433
93. Gerber AP, Herschlag D, Brown PO (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol 2:E79
94. Ule J, Jensen K, Mele A, Darnell RB (2005) CLIP: a method for identifying protein–RNA interaction sites in living cells. Methods 37:376–386
95. Wang Z, Tollervey J, Briese M, Turner D, Ule J (2009) CLIP: Construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. Methods 48:287–293
96. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu X, Gage FH (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA–protein interactions in stem cells. Nat Struct Mol Biol 16:130–137
97. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456:464–469
98. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res 19:381–394
99. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 460:479–486
100. Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G (2008) A human snoRNA with microRNA-like functions. Mol Cell 32:519–528
101. Bühler M, Spies N, Bartel DP, Moazed D (2008) TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the *Schizosaccharomyces pombe* siRNA pathway. Nat Struct Mol Biol 15:1015–1023
102. Melamed D, Arava Y (2007) Genome-wide analysis of mRNA polysomal profiles with spotted DNA microarrays. Meth Enzymol 431:177–201
103. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100:3889–3894
104. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, Preiss T, Bähler J (2007) A network of multiple regulatory layers shapes gene expression in fission yeast. Mol Cell 26:145–155
105. Bailly J, Fraissinet-Tachet L, Verner M, Debaud J, Lemaire M, Wésolowski-Louvel M, Marmeisse R (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1:632–642
106. Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3:e3042
107. Gilbert JA, Thomas S, Cooley NA, Kulakova A, Field D, Booth T, McGrath JW, Quinn JP, Joint I (2009) Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. Environ Microbiol 11:111–125
108. Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ Microbiol 11:1358–1375
109. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M (2009) Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol 27:652–658