RNA sequencing: the teenage years

Rory Stark[1], *Marta Grzelak*[1] *and James Hadfield*[2]*

Abstract | Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

Differential gene expression
(DGE). The analysis methods that together allow users to determine the quantitative changes in expression levels between experimental groups.

Read depth
The total number of sequencing reads obtained for a sample. This should not be confused with coverage, or sequencing depth, in genome sequencing, which refers to how many times individual nucleotides are sequenced.

Short-read
Sequencing technologies that generate reads of up to 500 bp, more commonly 100–300 bp, that represent fragmented or degraded mRNAs.

[1]*Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK.*

[2]*Precision Medicine, Oncology R&D, AstraZeneca, Cambridge, UK.*

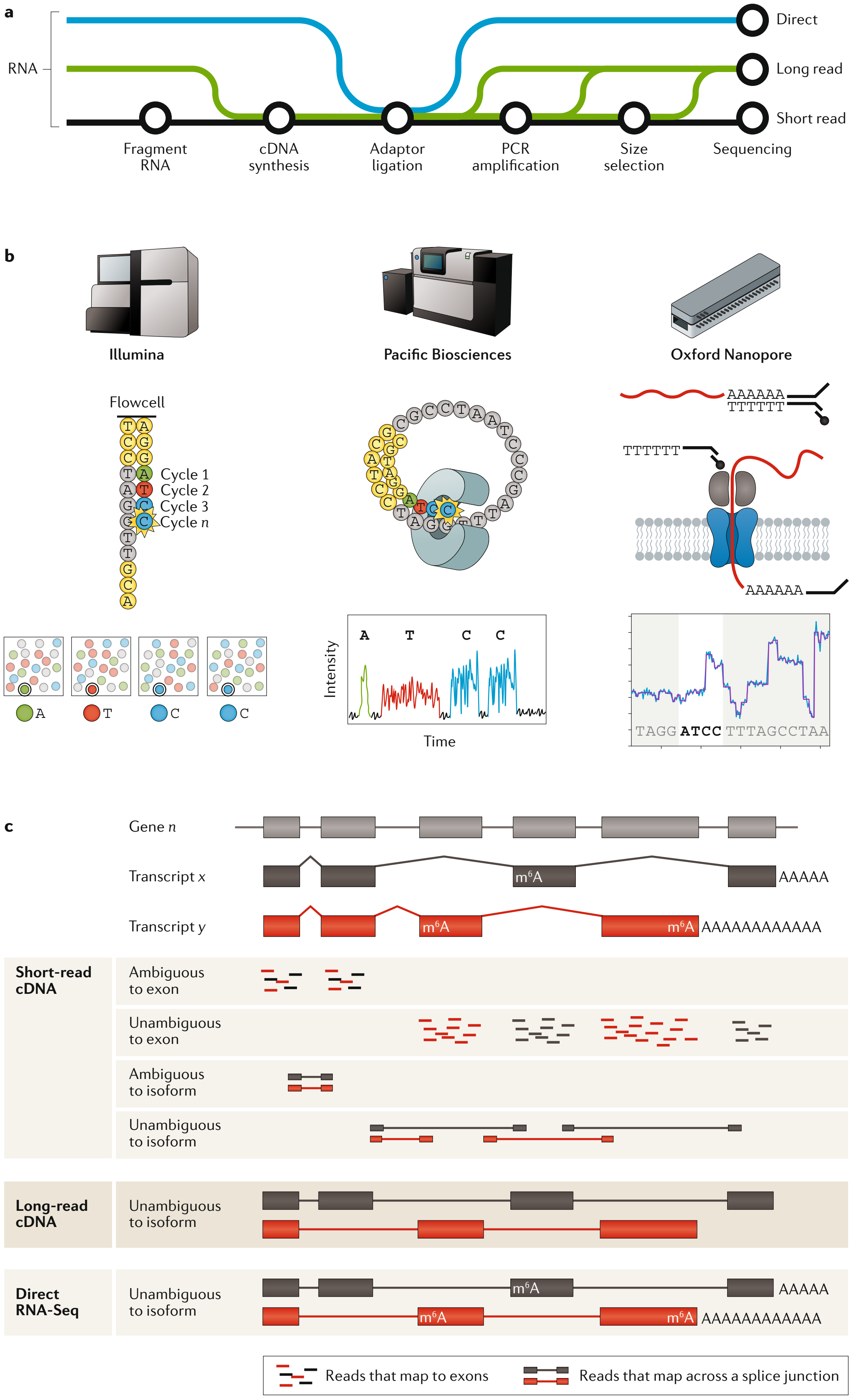
**e-mail: James.hadfield@astrazeneca.com*

https://doi.org/10.1038/s41576-019-0150-2

RNA sequencing (RNA-seq) was developed more than a decade ago[1,2] and since then has become a ubiquitous tool in molecular biology that is shaping nearly every aspect of our understanding of genomic function. RNA-seq is most often used for analysing differential gene expression (DGE). The essential stages of a DGE assay have not changed substantially from the earliest publications. The standard workflow begins in the laboratory, with RNA extraction, followed by mRNA enrichment or ribosomal RNA depletion, cDNA synthesis and preparation of an adaptor-ligated sequencing library. The library is then sequenced to a read depth of 10–30 million reads per sample on a high-throughput platform (usually Illumina). The final steps are computational: aligning and/or assembling the sequencing reads to a transcriptome, quantifying reads that overlap transcripts, filtering and normalizing between samples, and statistical modelling of significant changes in the expression levels of individual genes and/or transcripts between sample groups. Early RNA-seq experiments generated DGE data from bulk tissue and demonstrated its use across a wide range of organisms and systems, including *Zea mays*[1], *Arabiodopsis thaliana*[2], *Saccharomyces cerevisae*[3], *Mus musculus*[4] and *Homo sapiens*[5,6]. While the term RNA-seq is often used as a catch-all for very different methodological approaches and/or biological applications, DGE analysis remains the primary application of RNA-seq (Supplementary Table 1) and is considered a routine research tool.

Broader applications of RNA-seq have shaped our understanding of many aspects of biology, such as by revealing the extent of mRNA splicing[7] and the regulation of gene expression by non-coding RNAs[8,9] and enhancer RNAs[10]. The adaptation and evolution of RNA-seq has been driven by technological developments (both wet-lab and computational) and has enabled a richer and less biased view of RNA biology and the transcriptome than was possible with previous microarray-based methods. To date, almost 100 distinct methods have been derived from the standard RNA-seq protocol[11]. Much of this method development has been achieved on Illumina short-read sequencing instruments, but recent advances in long-read RNA-seq and direct RNA sequencing (dRNA-seq)[12–14] methods are enabling users to ask questions not answerable with Illumina short-read technologies.

In this Review, we begin by establishing a 'baseline' short-read RNA-seq assay for DGE before comparing and contrasting standard short-read approaches with the emerging long-read RNA-seq[15,16] and dRNA-seq technologies[12–14]. We describe the developments in library preparation for short-read sequencing protocols, practices in experimental design and computational workflows that have made DGE analysis so pervasive. We then look at developments that go beyond bulk RNA-seq for DGE, including single-cell and spatially resolved transcriptome analysis. We provide examples of how RNA-seq has been adapted to investigate key aspects of RNA biology, including analysis of transcriptional and translational dynamics, RNA structure, and RNA–RNA and RNA–protein interactions. We finish by briefly discussing the likely future for RNA-seq, whether single-cell and spatial RNA-seq methods will become as routine as DGE analysis, and in what niches long reads might replace short reads for RNA-seq analysis. Space limitations prevent us from covering all RNA-seq methods; notable omissions include analysis of non-coding transcriptomes[17,18], prokaryotic transcriptomes[19,20] and epitranscriptomes[21,22].


a
RNA
Direct
Long read
Short read
Fragment RNA
cDNA synthesis
Adaptor ligation
PCR amplification
Size selection
Sequencing
b
Illumina
Pacific Biosciences
Oxford Nanopore
Flowcell
Cycle 1
Cycle 2
Cycle 3
Cycle n
A
T
C
C
Intensity
Time
A T C C
AAAAAA
TTTTTT
TTTTTT
AAAAAA
TAGG ATCC TTTAGCCTAA
c
Gene n
Transcript x
m6A
AAAAA
Transcript y
m6A
m6A
AAAAAAAAAAAA
Short-read cDNA
Ambiguous to exon
Unambiguous to exon
Ambiguous to isoform
Unambiguous to isoform
Long-read cDNA
Unambiguous to isoform
Direct RNA-Seq
Unambiguous to isoform
m6A
AAAAA
m6A
m6A
AAAAAAAAAAAA
Reads that map to exons
Reads that map across a splice junction

◀ Fig. 1 | **Short-read, long-read and direct RNA-seq technologies and workflows.** **a** | Shown is an overview of library preparation methods for different RNA-sequencing (RNA-seq) methods, which can be categorized as short-read sequencing (black), long-read cDNA sequencing (green) or long-read direct RNA-seq (blue). The complexity and bias of library preparation varies according to the specific approach used. The short-read and long-read cDNA methods share many of the same steps in their protocols, but all methods require an adaptor ligation step and all are affected by sample quality and computational issues upstream and downstream of library preparation. **b** | An overview is shown of the three main sequencing technologies for RNA-seq. The Illumina workflow (left panel): after library preparation, individual cDNA molecules are clustered on a flowcell for sequencing by synthesis using 3′ blocked fluorescently labelled nucleotides. In each round of sequencing, the growing DNA strand is imaged to detect which of the four fluorophores has been incorporated, and reads of 50–500 bp can be generated. The Pacific Biosciences workflow (middle panel): after library preparation, individual molecules are loaded into a sequencing chip, where they bind to a polymerase immobilized at the bottom of a nanowell. As each of the fluorescently labelled nucleotides is incorporated into the growing strand, they fluoresce and are detected, and reads of up to 50 kb can be generated. The Oxford Nanopore workflow (right panel): after library preparation, individual molecules are loaded into a flowcell, where motor proteins, which are attached during adaptor ligation, dock with nanopores. The motor protein controls the translocation of the RNA strand through the nanopore, causing a change in current that is processed to generate sequencing reads of 1–10 kb. **c** | Comparison of short-read, long-read and direct RNA-seq analysis. Over 90% of human genes (gene *n*) are alternatively spliced to form two or more distinct and expressed isoforms (transcripts *x* and *y*). The complexity of information captured increases from short-read cDNA sequencing, where isoform detection can be compromised by reads that cannot be mapped unambiguously, to long-read methods that directly sequence isoforms. In short-read cDNA sequencing, a significant proportion of reads map ambiguously when an exon is shared between isoforms; reads that span exon–exon junctions can be used to improve the isoform analysis but can also be mapped ambiguously when a junction is shared between isoforms. These issues complicate analysis and the interpretation of results. Long-read cDNA methods can generate full-length isoform reads that remove, or substantially reduce, these artefacts and improve differential isoform expression analysis. However, these methods rely on cDNA conversion, which removes information about RNA base modifications and can only make crude estimates of polyadenylation (poly(A)) tail length. Direct RNA-seq enables full-length isoform analysis, base modification detection (such as N6-methyladenosine (m^6A)) and poly(A) tail length estimation.

Advances in RNA-seq technologies

The Illumina short-read sequencing technology has been used to generate more than 95% of the published RNA-seq data available on the Short Read Archive (SRA)[23] (Supplementary Table 2). As short-read sequencing of cDNA comprises nearly all publicly available mRNA-seq data, we consider this the baseline RNA-seq technology and discuss the primary workflow and its limitations. However, long-read cDNA sequencing and, most recently, dRNA-seq methods may soon present a challenge to its dominance, as users seek out methods that can deliver improved isoform-level data (FIG. 1; TABLE 1).

Short-read cDNA sequencing for DGE. Short-read sequencing has become the de facto method to detect and quantify transcriptome-wide gene expression, partly because it is cheaper and easier to implement than microarrays but primarily because it generates comprehensive, high-quality data that capture quantitative expression levels across the transcriptome. The core steps of a DGE assay using the Illumina short-read sequencing platform include RNA extraction, cDNA synthesis, adaptor ligation, PCR amplification, sequencing and analysis (FIG. 1). This protocol results in cDNA fragments that are usually under 200 bp in length, due to mRNA fragmentation and size selection of 150- to 200-bp fragments during bead-based library purification. The RNA-seq library is sequenced to an average of 20–30 million reads per sample, and the data are computationally processed to determine the fraction of reads associated with individual genes or transcripts before being subject to a statistical analysis (see RNA-seq data analysis). Short-read RNA-seq is robust, and large-scale comparisons of short-read sequencing technologies for RNA-seq have reported high intra-platform and inter-platform correlations[24,25]. However, there are a number of points in the process, during both the sample preparation and computational analysis phases, at which imperfections and biases may be introduced. These limitations may affect the ability of the experiment to address specific biological questions, such as correctly identifying and quantifying which of multiple isoforms are expressed from a gene[8]. This example is particularly relevant to very long, or highly variable, transcript isoforms such as those found in the human transcriptome; 50% of transcripts are >2,500 bp long in humans[26], with a range from 186 bp to 109 kb (REF.[27]). Although short-read RNA-seq allows detailed analysis of even the longest transcripts, the required protocols do not scale to whole-transcriptome analysis[28,29]. Other biases and limitations can result from the myriad computational methods that can be applied to RNA-seq data, such as differences in how ambiguous or multi-mapped reads are handled (see RNA-seq data analysis). A novel approach to generating synthetic long reads enables full-length mRNA sequencing and attempts to address some of these limitations[30]. It does so by tagging full-length cDNAs with unique molecular identifiers (UMIs)[31–34], which are copied across the length of individual cDNA molecules before preparation of a short-read RNA-seq library. Transcript isoforms can be reconstructed in contigs of up 4 kb for isoform discovery and expression analysis. However, the greatest potential for fundamentally addressing the inherent limitations of short-read cDNA sequencing lies with long-read cDNA sequencing and dRNA-seq methods.

Long-read cDNA sequencing. Although Illumina sequencing is currently the dominant RNA-seq platform, both Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) provide alternative long-read technologies that enable single-molecule sequencing of complete individual RNA molecules after conversion to cDNA[15,16,35]. By removing the need for the assembly of short RNA-seq reads, these approaches overcome some of the issues associated with short-read approaches. For example, ambiguity in the mapping of sequence reads is reduced, and longer transcripts can be identified, which leads to a more complete capture of isoform diversity. Also reduced is the high rate of false-positive splice-junction detection by many short-read RNA-seq computational tools[36].

The development of Iso-Seq for the PacBio technology enabled the generation of full-length cDNA reads for transcripts up to 15 kb, which facilitated the discovery of large numbers of previously unannotated transcripts and confirmed earlier gene predictions by detecting full-length homologous sequences across species[15,16,37].

Long-read
Sequencing technologies that generate reads of over 1,000 bp that represent either full-length or near-full-length mRNAs.

Direct RNA sequencing
(dRNA-seq). Sequencing technologies that generate reads by directly sequencing RNA without modification or reverse transcription, usually with the aim of sequencing full-length or near-full-length mRNAs.

Multi-mapped reads
Sequencing reads from homologous regions of the transcriptome that cannot be unambiguously mapped to the transcriptome or genome.

Synthetic long reads
A method for generating long reads from multiple short reads by assembly.

Table 1 | **Comparison of short-read and long-read RNA-seq platforms**

Sequencing technology	Platform	Advantages	Disadvantages	Key applications
Short-read cDNA	Illumina, Ion Torrent	• Technology features very high throughput: currently 100–1,000 times more reads per run than long-read platforms • Biases and error profiles are well understood (homopolymers are still an issue for Ion Torrent) • A huge catalogue of compatible methods and computational workflows are available • Analysis works with degraded RNA	• Sample preparation includes reverse transcription, PCR and size selection adding biases to all methods • Isoform detection and quantitation can be limited • Transcript discovery methods require a de novo transcriptome alignment and/or assembly step	Nearly all RNA-seq methods have been developed for short-read cDNA sequencing: DGE, WTA, small RNA, single-cell, spatialomics, nascent RNA, translatome, structural and RNA–protein interaction analysis, and more are all possible
Long-read cDNA	PacBio, ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified	• Technology features low-to-medium throughput: currently only 500,000 to 10 million reads per run • Sample preparation includes reverse transcription, PCR and size selection (for some protocols), adding biases to many methods • Degraded RNA analysis is not recommended	Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis
Long-read RNA	ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified • Sample preparation does not require reverse transcription or PCR-reducing biases • RNA base modifications can be detected • Poly(A) tail lengths can be directly estimated from single-molecule sequencing	• Technology features low throughput: currently only 500,000 to 1 million reads per run • Sample preparation and sequencing biases are not well understood • Degraded RNA analysis is not recommended	• Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis • Ribonucelotide modifications can be detected

The table provides a high-level overview of the advantages and disadvantages of the three major sequencing technologies for RNA sequencing (RNA-seq). DGE, differential gene expression; HLA, human leukocyte antigen; MHC, major histocompatibility complex; ONT, Oxford Nanopore; PacBio, Pacific Biosciences; poly(A), polyadenylation; WTA, whole-transcriptome analysis.

In the standard Iso-Seq protocol, high-quality RNA is converted to full-length cDNA for sequencing using a template-switching reverse transcriptase[38,39]. The resulting cDNAs are PCR amplified and used as the input for PacBio single-molecule, real-time (SMRT) library preparation. Owing to a bias in the sequencing of short transcripts, which diffuse more quickly to the active surface of the sequencing chip, size selection is recommended for transcripts from 1 to 4 kb, to allow more equal sampling of long and short transcripts in this size range. Due to the large amount of template required for PacBio sequencing, a large-volume PCR is performed, which requires optimization in order to reduce the impact of overamplification. After PCR end-repair and PacBio SMRT-adaptor ligation, long-read sequencing is performed; size selection bias can be further controlled at this step by modifying the loading conditions of the sequencing chip[40].

ONT cDNA sequencing also generates full-length transcript reads[35,41], even from single cells[14]. Template-switching reverse transcription is again used to prepare full-length cDNAs, which can be optionally amplified by PCR, before adaptors are attached in order to create a sequencing library. Direct cDNA sequencing removes PCR bias, leading to higher-quality results; however, the sequencing yields (numbers of reads) are higher for PCR-amplified cDNA libraries, which enables users to start with much smaller amounts of input RNA. The size selection bias observed for PacBio instruments has not been reported for ONT cDNA sequencing.

Both of these long-read cDNA methods are limited by the use of the standard template-switching reverse transcriptase, which generates cDNA from full-length RNAs as well as truncated RNAs. Reverse transcriptases are available that convert only 5′-capped mRNAs to cDNA[16], which improves data quality by reducing the amount of cDNA generated from transcripts truncated by RNA degradation, RNA shearing or incomplete cDNA synthesis. However, these reverse transcriptases have been shown to negatively affect read length on the ONT platform[42].

Long-read direct RNA sequencing. The long-read methods discussed above, like the baseline short-read platform, rely on converting mRNA to cDNA before sequencing. Oxford Nanopore recently demonstrated that their nanopore sequencing technology[43,44] can be used to sequence RNA directly[12,45] — that is, without modification, cDNA synthesis and/or PCR amplification during library preparation. This approach, termed dRNA-seq, removes the biases generated by these processes and enables epigenetic information to be retained. Library preparation from RNA involves sequential ligation of two adaptors. First, a duplex adaptor bearing an oligo(dT) overhang is annealed and ligated to the RNA polyadenylation (poly(A)) tail, which is followed by an optional (but recommended) reverse-transcription step that improves the sequencing throughput. The second ligation step attaches the sequencing adaptors, which are pre-loaded with the motor protein that drives sequencing. The library is then ready for MinION sequencing, in which RNA is sequenced directly from the 3′ poly(A) tail to the 5′ cap. Initial studies demonstrated that dRNA-seq

Unique molecular identifiers (UMIs). Short sequences or barcodes usually added during RNA sequencing (RNA-seq) library preparation (but also by direct RNA ligation), before amplification, that mark a sequence read as coming from a specific starting molecule. The approach is used to reduce the quantitative biases of RNA-seq and is particularly useful in low-input or single-cell experiments.

Read length The length of the individual sequencing reads, which is usually 50–150 bp for short-read RNA sequencing.

generates read lengths of around 1,000 bp, with maximum lengths exceeding 10 kb (REFS[12,45,46]). These long reads have several advantages over short reads: they can improve isoform detection compared with short reads, and they can also be used to estimate poly(A) tail length, which is important for alternative poly(A) analysis (see Improved RNA-seq library preparation). The nanopolish-polya tool uses nanopore data to measure poly(A) tail lengths, both between genes and between transcript isoforms. This analysis has confirmed that transcripts that retain introns have marginally longer poly(A) tails than completely spliced transcripts[45]. Finally, although still in its infancy, dRNA-seq has the potential to detect RNA base modifications and therefore has huge potential to enable new insights into the epitranscriptome, in particular[12,22,45,47].

Comparing long-read and short-read technologies. Although long-read technologies have some clear advantages over short-read sequencing for evaluating transcriptomes, they have some distinct limitations, as well. In particular, long-read platforms suffer from much lower throughput and much higher error rates than more mature short-read platforms. The main advantages of long reads — their ability to capture more of individual transcripts — are additionally dependent on having high-quality RNA libraries as input. Together, these limitations can affect the sensitivity and specificity of experiments that rely exclusively on long reads.

The major limitation of long-read sequencing methods is currently their throughput. While a single RNA-seq run on an Illumina platform can generate 10^9–10^{10} short reads, experiments performed with the PacBio and ONT platforms will generate 10^6–10^7 reads per run. This low throughput limits the size of experiments that can be undertaken with long-read sequencing and reduces the sensitivity of differential gene expression. However, high read depth is not necessary for all applications. Users primarily interested in isoform discovery and characterization will consider read length to be of greater importance than read depth. Obtaining a long read from highly expressed genes of >1 kb is almost guaranteed with 1 million PacBio circular consensus-sequencing (CCS) reads[48], and this situation is likely to be the same using ONT technology. As such, read depth is primarily an issue for genes expressed at low to medium levels. The limitation of lower throughput is most obvious when performing the large-scale DGE experiments required for contemporary functional genomics analysis. In these studies, multiple sample groups, each consisting of multiple replicates (see Designing better RNA-seq experiments), must be profiled in order to attain sufficient statistical power to have confidence that changes in expression across the transcriptome are being characterized accurately. For these applications, long-read technologies are unlikely to supplant short-read platforms until their throughput can be improved by at least two orders of magnitude. As the number of full-length RNA-seq reads increases, transcript detection sensitivity will increase to levels similar to those seen on Illumina, but with even higher specificity. In the meantime, by combining Illumina short-read RNA-seq with PacBio long-read Iso-Seq (and presumably also with ONT methods), it is possible to increase the number, sensitivity and specificity of full-length RefSeq-annotated isoform detection, while maintaining the quality of transcript quantification[48,49]. Although long-read RNA-seq methods currently have higher experimental costs, they can detect isoforms that are missed by short-read methods, particularly in regions that are difficult to sequence yet clinically relevant, such as the highly polymorphic human major histocompatibility complex[50] or the androgen receptor[51].

The next important limitation of long-read sequencing platforms is their higher error rates, which are one or two orders of magnitude higher than those seen with mature Illumina machines[46]. Data generated from the platforms also contain more insertion–deletion errors[52]. While these error rates are of concern for variant calling, in RNA-seq it is less crucial that every base be called correctly, as the goal is only to disambiguate transcripts and isoforms. For applications in which the error rate is a concern, there are potential mitigations. The random errors that typically occur on the PacBio SMRT sequencing platform can be mitigated by increasing the read depth[53] using CCS. In this approach, cDNAs are size-selected and circularized using adaptors so that each molecule is sequenced multiple times, generating continuous long reads that vary from >10–60 kb in length and that contain many copies of the original cDNA. These long reads are processed computationally into individual cDNA subreads, which are combined in order to generate the consensus sequence. The more times the molecule is sequenced, the lower the resulting error rate; CCS has been shown to reduce error rates to short-read levels, or even lower[15,54]. However, devoting more of the sequencing power of this platform to re-reading the same molecule exacerbates the throughput issues, as even fewer unique transcripts can be read.

The sensitivity of long-read RNA-seq methods is also limited by several other factors. First, they depend on long RNA molecules being present as full-length transcripts, which is not always possible, because of RNA degradation or shearing during sample handling and RNA extraction. Although this leads to a controllable 3′ bias in short-read RNA-seq data, even low levels of RNA degradation will limit long-read RNA-seq for users interested primarily in full-length transcriptome analysis. As such, prospective users need to carefully control the quality of the samples used after RNA extraction. Second, median read lengths are further constrained by technical issues and biases in library preparation, such as truncation of cDNA synthesis or synthesis of degraded mRNAs[16]. These processes may be improved by the recent development of highly processive reverse transcriptases, which generate better strand specificity and more even 3′–5′ transcript coverage[55,56]. Although not yet widely adopted, these highly processive reverse transcriptases also improve coverage of structurally stable RNAs (such as tRNAs), which the reverse transcriptases commonly used in the oligo-dT and whole-transcriptome analysis (WTA) methods struggle to process. Third, biases inherent to sequencing platforms

Sensitivity
A measure of the proportion of transcripts present in the sample that are detected. It is affected by sample handling, library preparation, sequencing and computational biases.

Specificity
A measure of the proportion of differentially expressed transcripts that are correctly identified. It is affected by sample handling, library preparation, sequencing and computational biases.

(such as low diffusion of long library molecules onto the surface of the sequencing chip) can reduce the coverage of longer transcripts.

Long-read methods (using either cDNA or dRNA-seq) address the basic limitation of short-read methods for isoform analysis — that is, their read length. Long-read methods can generate full-length transcript reads spanning isoforms from the poly(A) tail to the 5′ cap. As such, these methods make it possible to analyse transcripts and their isoforms without reconstructing them, or inferring their existence, from short reads; each sequence read simply represents its starting RNA molecule. Future application of full-length cDNA-seq or dRNA-seq to DGE analysis will depend on higher yields from the PacBio and ONT technologies. Long-read RNA-seq analysis is being adopted rapidly by users and combined with deep short-read RNA-seq data for more comprehensive analysis — very similar to the hybrid approach taken for genome assembly[57]. In time, the long-read and dRNA-seq methods are likely to demonstrate that the list of identified genes and transcripts, even in well-characterized organisms, is far from comprehensive[58]. As the methods mature, and as sequencing yields increase, differential isoform analysis will become routine. It remains to be seen what impact synthetic long-read RNA-seq or other developments will have on the field. However, for now, Illumina short-read RNA-seq continues to dominate, and the rest of this Review will focus on short-read sequencing.

Improved RNA-seq library preparation

RNA-seq was initially developed to analyse polyadenylated transcripts, using methods derived from earlier expressed-sequence tag and microarray studies. However, the use of next-generation sequencing revealed limitations in these methods that were not clearly evident in microarray data. As such, a number of major advances in library preparation methods were reported with, or soon after, the initial publication of RNA-seq. For instance, fragmentation of RNA before cDNA synthesis was shown to reduce 3′:5′ bias[4], and strand-specific library preparation methods, which allow sense and antisense transcripts to be differentiated, were shown to provide a more accurate estimate of transcript abundance[2,59] (reviewed and compared by Levin et al.[60]). RNA fragmentation and strand-specific library preparation quickly became standard in most RNA-seq kits. Here we briefly describe some of the other protocol modifications that RNA-seq users should be aware of when choosing the method most suited to their biological questions and the samples available. These include alternative methods to oligo-dT enrichment when selecting RNAs for sequencing, methods to specifically select for the 3′ or 5′ ends of transcripts, the use of UMIs to differentiate technical from biological duplication and improved library preparation for degraded input RNA. Combinations of these methods (and/or the use of dRNA-seq and/or the methods described in Beyond steady-state RNA analysis) allow users to unravel the transcriptome complexity produced by alternative poly(A) (APA), alternative promoter usage and alternative splicing.

Moving beyond poly(A) enrichment. The majority of published RNA-seq data have been generated from oligo-dT-enriched mRNA, which selects for transcripts containing a poly(A) tail and focuses sequencing on the protein-coding regions of the transcriptome. However, in addition to this method being 3′ biased, many non-coding RNAs, such as microRNAs (miRNAs) and enhancer RNAs, are not polyadenylated and therefore cannot be studied using this approach. Removing selection entirely is not an option, as such a procedure results in up to 95% of reads coming from ribosomal RNAs (rRNAs)[61]. As such, users have a choice of using oligo-dT for mRNA-seq or rRNA depletion for WTA. Short non-coding RNAs that are not captured by oligo-dT methods and are poorly represented by WTA approaches require specific small-RNA methods, which primarily use sequential RNA ligation[62] (reviewed elsewhere[17]).

WTA generates RNA-seq data from coding and some non-coding RNAs. It is also compatible with degraded samples in which fragmentation of the RNA leads to separation of the poly(A) tail from the rest of the transcript. Ribosomal RNA removal is achieved either by separating rRNAs from other RNA species (so-called pull-out) or by selective degradation of rRNA by RNase H. Both approaches use sequence- and species-specific oligonucleotide probes that are complementary to both cytoplasmic rRNAs (5S rRNA, 5.8S rRNA, 18S rRNA and 28S rRNA) and mitochondrial rRNAs (12S rRNA and 16S rRNA). Oligos, often pre-mixed in order to simplify the processing of human, rat, mouse or bacterial (16S and 23S rRNA) samples, are added to RNA and hybridize with rRNA for subsequent depletion. Other high-abundance transcripts, such as globin or mitochondrial RNA, can similarly be depleted. Pull-out methods incorporate biotinylated probes and streptavidin-coated magnetic beads, which are used to remove the oligo-rRNA complexes from solution, leaving other RNAs for library preparation[63] (for example, Ribo-Zero (Illumina, USA) and RiboMinus (Thermo Fisher, USA)). RNase H methods degrade the resulting oligo-DNA:RNA hybrid using RNase H[61] (for example, NEBnext RNA depletion (NEB, USA) and RiboErase (Kapa Biosystems, USA)). A recent comparison of these methods shows that, in high-quality RNA, both can reduce rRNA to under 20% of the subsequent RNA-seq reads[64]. However, the authors also report that RNase H methods were much less variable than pull-out approaches and that some length bias was evident when comparing DGE across the different kits. The comparison also describes one other method, similar to RNase H, that performed well but has not previously been reported. The ZapR method (Takara Bio Inc., Japan) is a proprietary technology that enzymatically degrades RNA-seq library fragments derived from rRNAs. One limitation of rRNA depletion approaches is that they generally require a higher read depth per sample than oligo-dT RNA-seq[65,66] does, primarily because of carry-over of rRNAs.

Both the oligo-dT and rRNA depletion methods can be used for DGE experiments, and users will probably default to the method that has previously been used in their laboratory or that is most easily available to them.

However, some consideration should be given to which method to use, particularly for degraded samples, as WTA approaches will detect more transcripts, but at higher experimental cost, than oligo-dT methods.

Enriching RNA 3′ ends for Tag RNA-seq and alternative polyadenylation analysis. The standard short-read Illumina method requires 10–30 million reads per sample for high-quality DGE analysis. For users focused on gene-level expression and working on large or highly replicated experiments, or who are resource constrained, 3′-tag counting should be considered as an option. As sequencing is focused on the 3′ ends of transcripts, fewer reads are required, which reduces costs and allows higher numbers of samples to be run. Enriching 3′ ends also enables determination of the poly(A) sites on individual transcripts, which can vary because of APA of pre-mRNAs[67].

The 3′ mRNA-seq methods[68–70] generate a single fragment per transcript (a tag read), usually from the 3′ end, and tag abundance is assumed to be proportional to RNA concentration. Tag-sequencing protocols, such as QuantSeq (Lexogen, Austria)[70], are generally shorter than standard RNA-seq protocols. They have been optimized in order to remove the need for poly(A) enrichment and/or rRNA depletion by the use of random or anchored oligo-dT-primed cDNA synthesis and to replace adaptor-ligation steps with PCR immediately after cDNA synthesis. This approach can achieve sensitivity levels similar to that of standard RNA-seq, but at much lower read depths, which allows many more libraries to be multiplexed for sequencing. Data analysis is also simplified, because exon junction detection and the normalization of reads to gene length are not required[71]. However, 3′ mRNA-seq methods can be affected by internal priming on homopolymeric regions of transcripts, which leads to erroneous tags; they also offer very limited isoform analysis, which can offset any cost benefits from their lower read-depth requirements, especially for single-use samples.

APA of mRNAs generates isoforms with substantially different 3′ untranslated region (UTR) lengths. Not only does this generate multiple isoforms of a specific gene but it affects regulation of that transcript, owing to the *cis*-regulatory elements located in the 3′ UTR. Methods that allow users to investigate APA enable a more detailed understanding of miRNA regulation, mRNA stability and localization, and the translation[72] of mRNAs. APA methods aim to enrich the 3′ ends of transcripts in order to boost signal and sensitivity, and the tag-sequencing methods described above are well-suited to this end. Other methods include polyadenylation site sequencing (PAS-seq)[73], which fragments mRNA to around 150 bp and then uses oligo-dT-primed template switching to generate cDNAs for sequencing, with 80% of reads coming from 3′ UTRs. TAIL-seq[74] avoids the use of oligo-dT altogether, by first depleting rRNA and ligating 3′-RNA adaptors to the end of the poly(A) tail before fragmenting the RNA. After fragmentation, the RNA-seq library is completed by ligation of the 5′-RNA adaptor. APA can also be assessed by RNA–protein analysis methods, such as cross-linking immunoprecipitation (CLIP)[75] (see Beyond analysis of gene expression) and dRNA-seq.

Enriching RNA 5′ ends for transcription start-site mapping. Analysis of DGE can be complemented by the use of methods that enrich for 7-methylguanosine 5′-capped RNAs, to identify promoters and transcription start sites (TSSs). Several methods exist for this task, but only a few are in routine use. In cap analysis of gene expression (CAGE)[76] and RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE)[77], the 5′ cap of mRNAs is biotinylated after random-primed first-strand cDNA synthesis, which allows 5′ cDNA fragments to be enriched by streptavidin pull-down. The CAGE protocol produces short cDNA tags by using type II restriction enzymes that cut 21–27 bp downstream from 5′-ligated adaptors. By contrast, the RAMPAGE protocol makes use of template switching to produce slightly longer cDNAs, which are then enriched for sequencing. Single-cell-tagged reverse transcription sequencing (STRT-seq)[78] was developed to allow TSS mapping in single cells. The method uses biotinylated template-switching oligos to produce cDNAs, which are captured on beads and fragmented near the 5′ end to produce short cDNA tags. The 5′-end-capping technology that underpins CAGE was developed at the Riken Institute as a means to maximize the number of full-length cDNA clones in early functional genomics experiments. The Riken-led Functional Annotation of the Mouse (FANTOM) consortium demonstrated the power of CAGE by characterizing TSSs in over 1,300 human and mouse primary cells, tissues and cell lines[79]. CAGE also performed best in a recent method comparison[80]. However, the authors reported that 5′-end sequencing alone generates high numbers of false-positive TSS peaks, and they recommended confirmation of true positives with orthogonal methods, such as DNase I mapping or H3K4me3 chromatin immunoprecipitation followed by sequencing (ChIP–seq).

Use of unique molecular identifiers to detect PCR duplicates. RNA-seq data sets generally have high duplication rates, with many sequence reads mapping to the same location in the transcriptome. As opposed to whole-genome sequencing, where duplicate reads are assumed to be due to technical biases in the PCR step and are removed, in RNA-seq they are considered to be indicative of a true biological signal and are retained. Highly expressed transcripts may be represented by millions of starting RNA molecules in a sample, and, when sequenced as cDNA, many fragments will be identical. As such, the computational removal of duplicates identified during alignment is not necessarily recommended[81], as many of those duplicates are true biological signals. This is more likely when using single-end sequencing, as only one end of a pair of fragments need be the same for the pair to be identified as a duplicate; with paired-end sequencing, both ends must have been fragmented at the same position, which is less likely[81]. However, when preparing a cDNA library there will be some degree of technical duplication due to PCR biases, and it can be difficult to know the degree

Tag read
A read that is unique to a transcript, usually from the 3′ end of mRNA, for differential gene expression analysis, or the 5′ end, for analysis of transcription start sites and promoters.

Duplication rates
The frequencies at which sequencing reads for an RNA sequencing (RNA-seq) sample map to the same location in the transcriptome. In RNA-seq libraries, duplication rates can seem high for some transcripts because they are present at wildly different levels in the sample. Highly expressed genes will have high duplication rates, while low expressors may have minimal duplication. RNA-seq presents a particular challenge, as much of the duplication may be genuine signal from highly expressed transcripts, while some may be attributable to amplification and sequencing biases.

Single-end sequencing
Short-read sequencing performed from one end of the cDNA fragment, commonly used for differential gene expression experiments, due to its low cost.

Paired-end sequencing
Short-read sequencing performed from both ends of the cDNA fragment, often used for differential gene expression experiments, where maximum sensitivity to splicing is required because more bases of the individual cDNAs will be sequenced.

of technical versus biological duplication and to control for this in cases where PCR duplication bias represents a quality control issue that can compromise the results of an RNA-seq experiment.

UMIs[31–34] have been proposed as a way to account for amplification biases. The addition of random UMIs to cDNA molecules before amplification enables PCR duplicates to be identified and computationally removed from analysis, while retaining the true biological duplicates, thereby improving the quantification of gene expression and the estimation of allele frequency[81,82]. For a pair of sequence reads to be identified as a technical duplicate, they need to include the identical UMI and to map to the same place in the transcriptome (one or both ends, depending on the use of single-end or paired-end sequencing).

UMIs have been shown to improve the statistical analysis of RNA-seq data for DGE by reducing variance and false-discovery rates[81,83] and to be vital in the analysis of single-cell data, where amplification biases can be more problematic[84]. UMIs can also be useful when attempting to perform variant calling in RNA-seq data sets. Although highly expressed transcripts can yield high coverage rates suitable for such variant calling, especially if duplicate reads are included, UMIs can be used to remove amplification artefacts that can lead to erroneous calculations of allele frequency. UMIs are becoming standard in single-cell RNA-seq (scRNA-seq) library preparation kits and are also being used more frequently for bulk RNA-seq.

Improving the analysis of degraded RNA. Developments in RNA-seq library preparation methods have also improved the analysis of low-quality or degraded RNA, such as that obtained from clinically relevant archival material stored as formalin-fixed paraffin-embedded (FFPE) blocks. Low-quality RNA results in uneven gene coverage, higher DGE false-positive rates and higher duplication rates[85], and it negatively correlates with library complexity[86] in RNA-seq experiments. However, library preparation methods have been adapted to reduce the effect of RNA degradation. These methods are likely to be particularly important in the development of RNA-seq-based diagnostics, such as future assays similar to OncotypeDX (not currently a sequencing assay), which predicts breast cancer recurrence based on a 21-gene RNA signature. Although several methods are available, two (RNase H[61] and RNA exome[87]) have performed well in comparison studies[88,89]. As described above, the RNase H method uses a nuclease to digest rRNA in RNA:DNA hybrids but preserves degraded mRNA for RNA-seq analysis. The RNA exome method uses oligonucleotide probes to capture RNA-seq library molecules in a manner very similar to exome sequencing[90]. Both methods are simple to implement and generate high-quality and highly concordant gene expression data by reducing the impact of contaminating rRNA without losing degraded and fragmented mRNAs. The 3′-end tag-sequencing (see above)[91] and amplicon-sequencing (in which PCR amplifies over 20,000 exonic amplicons) methods[92] can also be used for analysis of degraded RNA but are not currently used as widely as the RNase H method.

Biological replicates
Parallel measurements of biologically distinct samples, such as tissue from three subjects, that capture natural biological variation, which may itself be either a subject of study or a source of noise. By contrast, technical replicates are repeated measurements of the same sample — for example, the same tissue processed three times.

Designing better RNA-seq experiments

Careful design of bulk DGE RNA-seq experiments is essential to obtaining high-quality and biologically meaningful data. Particular consideration needs to be given to the level of replication, the sequencing read depth and the use of single- or paired-end sequencing reads.

Replication and experimental power. It is essential that enough biological replicates be included in an experiment to capture the biological variability between samples; confidence in a quantitative analysis depends on this aspect more than on read depth or read length[93]. Although RNA-seq affords lower technical variability than microarray platforms, the stochastic variance inherent to biological systems requires any bulk RNA experiment to be carried out in replicate[94]. The use of additional replicates allows outlier samples to be identified and, if necessary, removed or down-weighted before performing biological analysis[95]. Determining the optimal number of replicates requires careful consideration of several factors, including effect size, within-group variation, acceptable false-positive and false-negative rates and maximum sample size[96,97], and can be aided by the use of RNA-seq experimental design tools[96] or power calculation tools[98,99].

Determining the correct number of replicates appropriate to a given experiment is not always straightforward. A 48-replicate yeast study showed that many of the tools available for DGE analysis detected only 20–40% of differentially expressed genes when only 3 replicates were included in the analysis[100]. The study suggested that a minimum of six biological replicates should be used, which is substantially more than the three or four replicates generally reported in the RNA-seq literature. A more recent study suggests that four replicates may be adequate, but it emphasizes the necessity of measuring biological variance — for example, in a pilot study — before settling on an appropriate number of replicates[93]. For highly diverse samples, such as clinical tissue from cancer patient tumours, many more replicates are likely to be required in order to pinpoint changes with confidence.

Determining the optimal read depth. Once RNA-seq libraries have been prepared, a decision needs to be made regarding how deeply to sequence them. Read depth refers to the target number of sequence reads obtained for each sample. For bulk RNA DGE experiments in eukaryotic genomes, it is generally accepted that read depths of around 10–30 million reads per sample are required[93,101–103]. However, it has been shown across multiple species that depths of as few as 1 million reads per sample provide transcript abundance estimates similar to those from 30 million reads for the most highly expressed half of the transcriptome[104]. If only relatively large changes in the expression of the most highly expressed genes are important, and if there are adequate biological replicates, less sequencing may be sufficient to address the hypothesis driving the experiment. After sequencing has been completed, the estimate of read depth can be validated by checking the distribution of reads among the samples and checking saturation curves to assess whether further sequencing is likely to increase the sensitivity of the experiment[103,105]. As sequencing yields have increased, it has become standard practice to multiplex all the samples comprising an experiment into a single 'pooled' sequencing library, in order to control for technical effects. The total number of reads required is determined by multiplying the number of samples by the desired read depth; the pooled library is then run as many times as is required to generate the desired total number of reads. This pooling requires that the concentration of each RNA-seq library be carefully measured and assumes that the amount of cDNA in the multiplexed samples is relatively even (low variance), so that the total number of reads is evenly distributed among the individual samples. Running a single lane to verify low variance between the samples is often worthwhile before committing to an expensive, multi-lane sequencing run.

Choosing parameters: read length and single-end or paired-end sequencing. The final sequencing parameters to be determined include the read length and whether to generate single-end or paired-end reads.

In many sequencing applications, the length of the sequencing reads has a great impact on the usefulness of the data, as longer reads give more coverage of the sequenced DNA. This is less applicable when using RNA-seq to examine DGE, where the important factor is the ability to determine where in the transcriptome each read came from (see Phase 1 — alignment and assembly of sequencing reads). Once a read's position can be mapped unambiguously, longer reads do not add much value in a quantification-based analysis[106]. For more qualitative RNA-seq assays, such as the identification of specific isoforms, longer reads may be more helpful.

The issue of single-end versus paired-end reads is similar. In single-end sequencing, only one end (3′ or 5′) of each cDNA fragment is used to generate a sequence read, while paired-end sequencing generates two reads for each fragment (one 3′ and one 5′). In assays where coverage of as many nucleotides as possible is desired, long-read paired-end sequencing is preferred. However, sequencing every base of a transcript fragment is not required for DGE analysis, where users need only count the reads mapping to a transcript after alignment. For example, a comparison of 'short', 50-bp single-end sequencing to 'long', 100-bp paired-end sequencing confirms that DGE results are not affected by the use of single-end sequencing[106]. This is because single-ended reads are sufficient to identify the source gene of most of the sequenced fragments. The same study showed that using short single-end reads compromised the ability to detect isoforms, as fewer reads were seen that spanned a splice junction. Paired-end sequencing can additionally help disambiguate read mappings and is preferred for alternative-exon quantification, fusion transcript detection and de novo transcript discovery, particularly when working with poorly annotated transcriptomes[107,108].

In practice, the choice between single-end or paired-end sequencing is often based on cost or on the sequencing technology available to the user. Before release of the Illumina NovaSeq, in most cases single-end sequencing cost less per million reads than paired-end sequencing and therefore allowed higher replication or read depth for the same experimental cost. Given a choice between obtaining a greater number of shorter single-end reads and generating longer and/or paired-end reads, an increase in read depth will have more impact on increasing the sensitivity of a DGE experiment.

RNA-seq data analysis

The number of computational approaches for analysing sequence reads to determine differential expression have multiplied considerably over the past 10 years, and, even for straightforward RNA-seq DGE, there is substantial divergence in analytical practice at each stage[50,103]. However, differences in the approaches used at each stage and differing combinations of techniques in a pipeline can have substantial effects on the biological conclusions that may be drawn from the data[50,103,109,110]. The optimal set of tools to use will depend on the specific biological question being explored, as well as the available computational resources[50]. Although multiple end points are possible, our emphasis is on surveying the tools and techniques commonly used in assessment, for every gene, of the likelihood that it is differentially expressed between sample groups. To achieve this, at least four distinct phases of analysis are required (FIG. 2; TABLE 2). The first phase takes the raw sequence reads generated by a sequencing platform and maps them to the transcriptome. Phase 2 quantifies the number of reads associated with each gene or transcript (an expression matrix). This process may involve one or more distinct sub-stages of alignment, assembly and quantification, or it may holistically generate the expression matrix from read counts in a single step. Usually there is a third phase where the expression matrix is altered by filtering lowly expressed features, as well as the crucial step of normalizing the raw counts to account for technical differences between the samples. The final phase in DGE is statistical modelling of the sample groups and covariates, to calculate confidence statistics related to differential expression.

Phase 1 — alignment and assembly of sequencing reads. After sequencing has been completed, the starting point for analysis is the data files, which contain base-called sequencing reads, usually in the form of FASTQ files[111]. The most common first step in processing these files is to map sequence reads to a known transcriptome (or annotated genome), converting each sequence read to one or more genomic coordinates. This process has traditionally been accomplished using distinct alignment tools, such as TopHat[112], STAR[113] or HISAT[114], which rely on a reference genome. Because the sequenced cDNA is derived from RNA, which may span exon boundaries, these tools perform a spliced alignment allowing for gaps in the reads when compared to the reference genome (which contains introns as well as exons).

If no high-quality genome annotation containing known exon boundaries is available, or if it is desirable to associate reads with transcripts (rather than genes), aligned reads can be used in a transcriptome assembly step[115]. Assembly tools such as StringTie[116] and SOAPdenovo-Trans[117] use the gaps identified in the alignments to derive exon boundaries and possible

Expression matrix
Matrix of values capturing the essential data for a differential-expression RNA-seq experiment. Rows are RNA features, such as genes or transcripts, with one column per sequenced sample. Values are generally counts of the number of reads associated with each RNA feature; these may be estimated for isoform features and are often transformed via normalization before subsequent analysis.

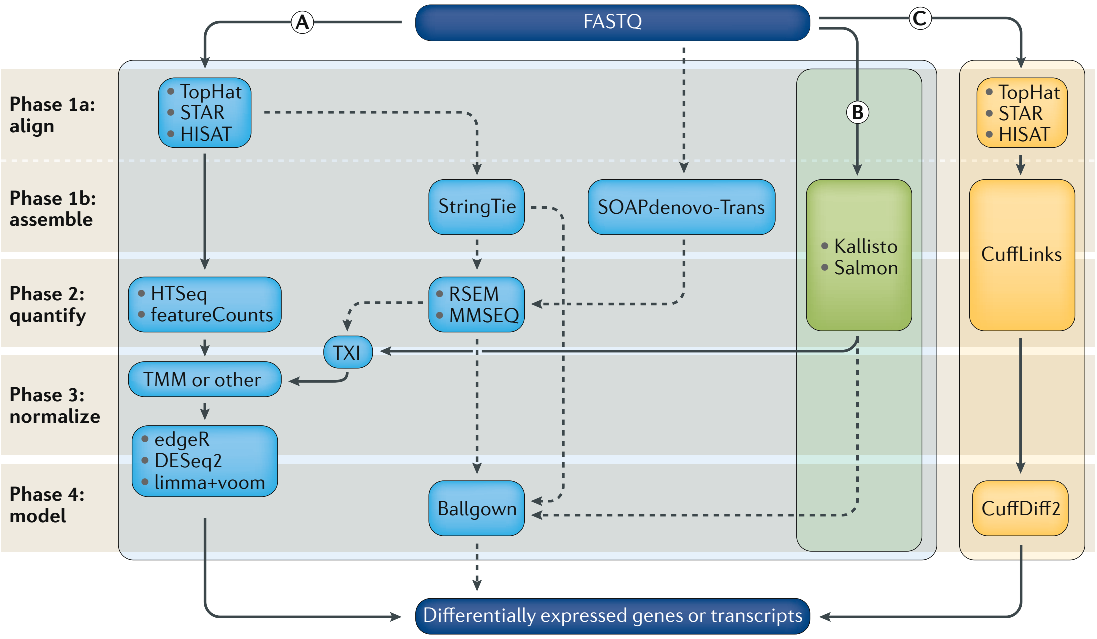


Fig. 2 | **RNA-seq data analysis workflow for differential gene expression.** Computational analysis for differential gene expression (DGE) begins with raw RNA sequencing (RNA-seq) reads in FASTQ format and can follow a number of paths. Three popular workflows (A, B and C, represented by the solid lines) are given as examples, and some of the more common alternative tools (represented by the dashed lines) are indicated. In workflow A, aligners such as TopHat[112], STAR[113] or HISAT2 (REF.[114]) use a reference genome to map reads to genomic locations, and then quantification tools, such as HTSeq[133] and featureCounts[134], assign reads to features. After normalization (usually using methods embedded in the quantification or expression modelling tools, such as trimmed mean of *M*-values (TMM)[142]), gene expression is modelled using tools such as edgeR[143], DESeq2 (REF.[155]) and limma+voom[156], and a list of differentially expressed genes or transcripts is generated for further visualization and interpretation. In workflow B, newer, alignment-free tools, such as Kallisto[119] and Salmon[120], assemble a transcriptome and quantify abundance in one step. The output from these tools is usually converted to count estimates (using tximport[130] (TXI)) and run through the same normalization and modelling used in workflow A, to output a list of differentially expressed genes or transcripts. Alternatively, workflow C begins by aligning the reads (typically performed with TopHat[112], although STAR[113] and HISAT[114] can also be used), followed by the use of CuffLinks[131] to process raw reads and the CuffDiff2 package to output transcript abundance estimates and a list of differentially expressed genes or transcripts. Other tools in common use include StringTie[116], which assembles a transcriptome model from TopHat[112] (or similar tools) before the results are passed through to RSEM[105] or MMSEQ[132] to estimate transcript abundance, and then to Ballgown[157] to identify differentially expressed genes or transcripts, and SOAPdenovo-trans[117], which simultaneously aligns and assembles reads for analysis via the path of choice.

splice sites. These de novo transcript assembly tools are particularly useful when the reference genome annotation may be missing or incomplete, or where aberrant transcripts (for example, in tumour tissue) are of interest. Transcriptome assembly methods may benefit from the use of paired-end reads and/or longer reads that have a greater likelihood of spanning splice junctions. However, complete de novo assembly of a transcriptome from RNA-seq data is not generally required for determining DGE.

More recently, computationally efficient 'alignment-free' tools, such as Sailfish[118], Kallisto[119] and Salmon[120] have been developed that associate sequencing reads directly with transcripts, without a separate quantification step (see phase 2 below). These tools have demonstrated good performance in characterizing more highly abundant (as well as longer) transcripts; however, they are less accurate in quantifying low-abundance or short transcripts[121].

The different tools for mapping sequence reads to transcripts have meaningful differences in how they allocate a subset of the reads, and this can affect the resulting expression estimates[50,121,122]. These effects are particularly noticeable for multi-mapped reads that could have come from more than one distinct gene, pseudo-gene or transcript. A comparison of 12 gene expression estimation methods revealed that some alignment methods underestimate the expression of many clinically relevant genes[123], primarily driven by the treatment of ambiguously mapping reads. Modelling how to properly allocate multi-mapped reads remains an open area of research in the computational analysis of RNA-seq data. It is common practice to exclude these reads from further analysis, which can bias results (see Phase 2 — quantification of transcript abundance)[122]. Other approaches include generating 'merged' expression features that encompass overlapping areas of shared mapping[124] and computing per-gene estimates of mapping uncertainty to be used in subsequent confidence calculations[125].

Phase 2 — quantification of transcript abundance. Once reads have been mapped to genomic or transcriptomic locations, the next step in the analysis process is to assign them to genes or transcripts, to determine abundance measures. Diverse comparative studies have shown that

Table 2 | **Common software tools in use for differential gene expression analysis using RNA-seq data**

Tool name	Alignment and/or assembly	Quantification	Normalization	Differential expression	Ref.
TopHat	Reference genome + annotation	NA	NA	NA	112
STAR		NA	NA	NA	113
HISAT		NA	NA	NA	114
SOAPdenovo-Trans	De novo assembly	NA	NA	NA	117
StringTie	De novo assembly	Transcript estimates	NA	NA	116
Kallisto	Alignment-free assembly	Transcript estimates	NA	NA	119
Salmon		Transcript estimates	NA	NA	120
Cufflinks	Transcript assembly	Transcript estimates	NA	NA	131
RSEM	NA	Transcript estimates	NA	NA	105
MMSeq	NA	Transcript estimates	NA	NA	132
HTSeq	NA	Read counts from non-overlapping annotated features	NA	NA	133
featureCounts	NA	Read counts from non-overlapping annotated features	NA	NA	134
tximport	NA	Transcript estimates converted to read counts	NA	NA	130
edgeR	NA	NA	TMM	Negative binomial distribution + GLM	143
limma+voom	NA	NA	TMM	Mean–variance transform + GLM	156
DESeq2	NA	NA	Various	Negative binomial distribution + GLM	155
Ballgown	NA	NA	NA	Input from StringTie, RSEM or alignment-free quantification, + GLM	157
CuffDiff	NA	NA	NA	DE from Cufflinks estimates	131

Some tools are used for multiple phases, such as combining transcript assembly and quantification, or normalization and differential expression modelling. See also Fig. 2. DE, differential expression; GLM, generalized linear modelling; NA, not applicable; RNA-seq, RNA sequencing; TMM, trimmed mean of *M*-values.

the approach taken at the quantification step has perhaps the largest impact on the ultimate results[110,123], even greater than the choice of aligner[122,126–128]. The quantification of read abundances for individual genes (that is, all transcript isoforms for that gene) relies on counting sequence reads that overlap known genes, using a transcriptome annotation. However, allocating reads to specific isoforms using short reads requires an estimation step, as many reads will not span splice junctions and therefore cannot be unambiguously assigned to a specific isoform[129]. Even in the case in which only gene-level differential expression is being studied, quantifying differences in isoforms may result in more accurate results in the case of a gene shifting its primary expression between isoforms of different lengths[130]. For example, if the primary isoform in one sample group has half the length of that in another sample group but is expressed at double the rate, a purely gene-based quantification will be unable to detect the differential expression of this feature.

Quantification tools in common use include RSEM[105], CuffLinks[131], MMSeq[132] and HTSeq[133], as well as the alignment-free direct quantification tools mentioned above. A read-count-based tool such as HTSeq (or the R equivalent, featureCounts[134]) will generally discard many aligned reads, including those that are multi-mapped or that overlap multiple expression features. As a result, homologous and overlapping transcripts may be eliminated from subsequent analysis. RSEM[105] allocates ambiguous reads using expectation maximization, whereas reference-free alignment methods such as Kallisto[119] include these reads in their transcript count estimates, which can bias results[121]. Transcript abundance estimates can be converted to read count equivalents, which some of the tools below require, using a package such as tximport[130]. The results of the quantification step are usually combined into an expression matrix, with a row for each expression feature (gene or transcript) and a column for each sample, with the values being either actual read counts or estimated abundances.

Phase 3 — filtering and normalization. Generally, quantified gene or transcript counts are also filtered and normalized, to account for differences in read depth, expression patterns and technical biases[135–137]. Filtering to remove features with uniformly low read abundance is straightforward and has been shown to improve the detection of true differential expression[138]. Methods for normalizing an expression matrix can be more complex. Straightforward transformations can adjust abundance quantities in order to account for differences in GC content[135] and read depth[136]. Early methods for accomplishing this, such as RPKM[4], are now recognized to be insufficient[136] and have been replaced by methods that correct for more subtle differences between samples, such as quartile or median normalization[139,140].

Comparative studies have shown empirically that the choice of normalization method can have a major impact on the ultimate results and biological conclusions[50,127,141].

Spike-in control

A pool of exogenous nucleic acids added at known concentration to a sample before processing. They are usually synthetic RNAs pre-pooled at varying concentrations and used to monitor reaction efficiency and to identify methodological bias and false-negative results.

Most computational normalization methods rely on two key assumptions: first, that the expression levels of most genes remain the same across replicate groups[137]; and second, that different sample groups do not exhibit a meaningful difference in overall mRNA levels. It is particularly important to carefully consider whether and how to perform normalization when these basic assumptions may not hold true. For example, if a certain group of genes is highly expressed in one sample group, while the same genes plus an additional group of genes are expressed in another sample group, then simply normalizing for read depth is inadequate, as the same number of sequence reads will be distributed over a greater number of expressed genes in the second sample group. Normalization procedures such as the trimmed mean of *M*-values (TMM) method[142] (incorporated into the edgeR[143] DGE analysis package) can compensate for these cases. Determining the appropriateness of a chosen normalization can be difficult; one option is to attempt an analysis using multiple methods, then to compare the consistency of the outcomes. If the results are highly sensitive to the normalization method, further exploration of the data should be conducted, to determine the origin of the discrepancies. Care must be taken to ensure that such a comparison not be used to select the normalization method that yields results that are the most compatible with the original hypothesis.

One approach to dealing with such issues is the use of spike-in control RNAs[144,145] — that is, introducing a priori alien RNA sequences at pre-defined concentrations, usually during library preparation. Spike-ins for RNA-seq include the External RNA Controls Consortium mix (ERCCs)[146], spike-in RNA variants (SIRVs)[147] and sequencing spike-ins (Sequins)[148]. As the RNA concentration of spike-in is known in advance, and as the concentration is directly related to the number of reads generated, it is possible to calibrate the expression levels of the transcripts from the sample. It has been argued that experiments with large fold-changes in overall expression cannot be properly analysed without spike-in controls[24,144,149]. However, in practice it can be difficult to consistently incorporate control spike-ins at preset levels[137,150,151], and they are more reliable at normalizing read counts at the gene level than at the transcript level[152], because individual isoforms can be expressed at markedly different concentrations within a sample. Currently, spike-in controls are not in wide use in published RNA-seq DGE experiments, although this is likely to change as more users encounter them in single-cell experiments (where they are used more widely[153]), and if the techniques can be refined to be more consistent.

Phase 4 — differential expression modelling. Once sequence reads have been processed into an expression matrix, the experiment can be modelled to determine which transcript features are likely to have changed their level of expression. Several tools are commonly used to accomplish this; some model read counts of gene-level expression, whereas others rely on transcript-level estimates. Gene-level tools typically rely on aligned read counts and use generalized linear models that enable complex experimental set-ups to be evaluated[154]. These include tools such as edgeR[143], DESeq2 (REF.[155]) and limma+voom[156], which are computationally efficient and provide comparable results[50,103]. Tools that model differential isoform expression, such as CuffDiff[131], MMSEQ[132] and Ballgown[157], tend to require more computational power and to vary more in their results[50,110,158]. However, the choices made before implementing these differential expression tools — that is, regarding alignment, quantification, or filtering and normalization — have a greater impact on the overall variance of the final results.

Beyond bulk RNA analysis

RNA-seq from bulk tissue and/or cells has revolutionized our understanding of biology, but it cannot easily resolve specific cell types and it fails to preserve spatial information, both of which are critical to understanding the complexity of biological systems. The methods that enable users to move beyond bulk RNA are very similar to standard RNA-seq protocols, but they enable very different questions to be asked. Single-cell sequencing has revealed cell types that were unknown in what were considered well-studied diseases, such as the discovery of ionocyte cells, which could be relevant to the pathology of cystic fibrosis[159]. Spatially resolved RNA-seq promises similar revelations in our understanding of cell-to-cell interactions in solid tissues, such as revealing the extent of fetal marker gene expression in minor populations of adult heart tissue[160]. Bulk RNA-seq will remain a dominant and valuable tool for the foreseeable future. But single-cell laboratory and analysis methods are rapidly being adopted by researchers, and, as spatial RNA-seq methods mature, they are also likely to become part of the routine RNA-seq toolkit. Both types of method will improve our ability to interrogate the complexity of multicellular organisms, and both are likely to be used in combination with bulk RNA-seq methods. Here we briefly describe the major single-cell and spatially resolved transcriptome methods, how they differ from bulk RNA-seq and what new users need to consider.

Single-cell analysis. scRNA-seq was first reported in 2009 (REF.[161]) by isolating individual oocytes in Eppendorf tubes containing a lysis buffer. Its application to novel biological questions, and the laboratory and computational methods available, continues to advance at such a rapid pace that even recent reviews[162,163] are rapidly becoming outdated. Each scRNA-seq method requires solid tissues to be dissociated, single cells to be separated (using very different approaches) and their RNA to be labelled and amplified for sequencing, and all methods use steps borrowed from earlier bulk RNA-seq protocols.

Mechanical disaggregation and enzymatic dissociation with collagenase and DNase produces the highest yields of viable cells in a single-cell suspension[164], but yields are highly tissue-specific and are best determined empirically — and very carefully[165]. Once a single-cell suspension is prepared, individual cells can be separated by various methods (FIG. 3a); as most laboratories have access to flow-cytometry instrumentation, the most easily accessible method is to flow-sort cells directly into microtitre plates containing lysis buffer[39,166]. For higher-throughput experiments, a wide number of techniques

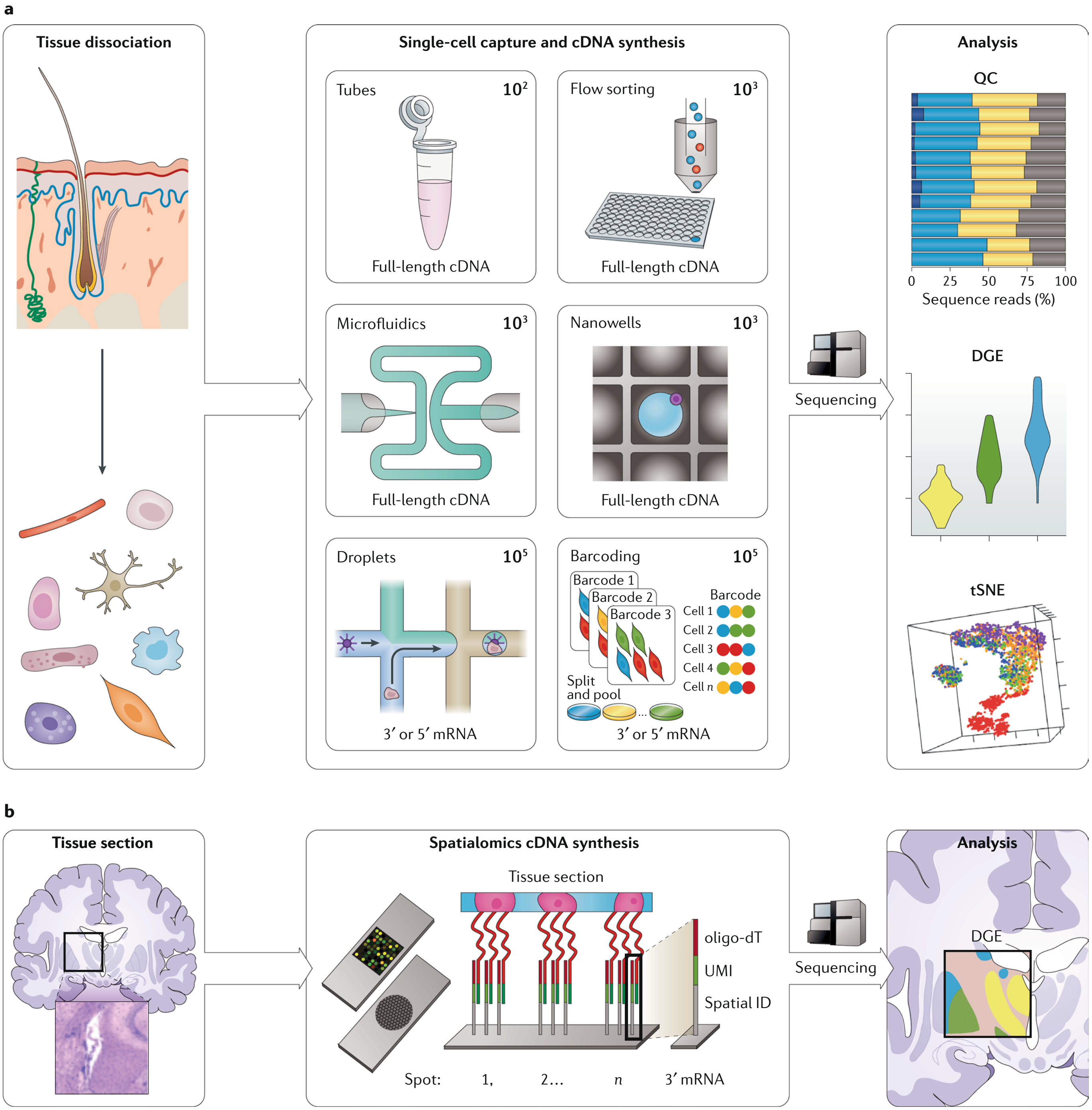


Fig. 3 | **The key concepts of single-cell and spatial RNA-seq. a** | An overview of the single-cell RNA sequencing (RNA-seq) workflow. Single-cell sequencing begins with the isolation of single cells from a sample, such as dissociated skin tissue, by any one of a number of methods, including micropipetting into individual microfuge tubes[161] or flow sorting into 96 or 384 well plates[39,166] containing a lysis buffer, capture in a microfluidic chip[167], distribution in nanowells[168], microfluidic isolation in reagent-filled droplets[169,170] or marking cells with in situ barcodes[171,172]. Cells are reverse transcribed in order to produce cDNA (usually tagged with unique molecular identifiers (UMIs)) for RNA-seq library preparation and sequencing. Quality control (QC), differential gene expression (DGE) and 2D visualization (t-distributed stochastic neighbour embedding (tSNE)), along with unsupervised clustering and network analysis, of the single-cell RNA-seq data are used to determine discrete cell populations. The number of cells usually profiled is indicated alongside each technology, as is the RNA-seq strategy — for example, 3′ or 5′ mRNA or full-length cDNA. **b** | An overview of the spatialomics workflow. Spatial encoding requires a frozen tissue section to be applied to oligo-arrayed microarray slides[184] or to ‘pucks’ of densely packed oligo-coated beads[185]. The mRNA diffuses to the slide surface and hybridizes to oligo-dT cDNA synthesis primers that encode UMIs and spatial barcodes. It is then reverse transcribed to produce cDNA, which is pooled for library preparation and sequencing. Computational analysis of the spatialomics data maps sequence reads back to their spatial coordinates after DGE analysis and allows differential spatial expression to be visualized. Single-cell and spatialomics RNA-seq data are usually generated on short-read sequencers. Part **a** is adapted from REF.[163], Springer Nature Limited.

for isolating cells exist that require specific single-cell instrumentation to be built or purchased. Individual cells can be physically captured in microfluidic chips[167] or loaded into nanowell[168] devices by Poisson distribution, they can be isolated and merged into reagent-filled droplets by droplet-microfluidic isolation (such as in Drop-Seq[169] and InDrop[170]), or they can be labelled with in situ sequence barcodes (such as in single-cell combinatorial indexing RNA sequencing (sci-RNA-seq)[171] and split-pool ligation-based transcriptome sequencing (SPLiT-seq)[172]). After single cells are isolated, they are lysed in order to release RNA into solution for cDNA synthesis, and the cDNA is used as the input for RNA-seq library preparation. The RNA from individual cells is generally amplified by PCR during library preparation. This amplification introduces PCR bias, which can be corrected by the use of UMIs[33,173,174]. Although only 10–20% of transcripts will be reverse transcribed, due to Poisson sampling[33], which limits transcript detection sensitivity, the various methods all generate usable data. Outside the wet lab, computational methods are also rapidly developing, and guidelines on scRNA-seq experimental design[162,175] have recently emerged. This rapid development in methodology means that technical comparisons of scRNA-seq methods are quickly outdated. Nevertheless, Ziegenhain et al.[84] provide a detailed overview of scRNA-seq methods, highlight the importance of UMIs in data analysis and report on which of the six methods profiled was most sensitive. However, their study does not include the widely adopted 10X Genomics (10XGenomics, USA) technology.

The major factors users will consider when choosing an scRNA-seq method include whether they require reads along the full length of transcripts, trade-offs between profiling more cells (breadth) or more transcripts per cell (depth) and the overall experimental costs. Full-length scRNA-seq systems[39,168,173,176] usually have lower throughput, since each cell needs to be processed independently up to the final scRNA-seq library. However, such systems allow users to interrogate alternative-splicing and allele-specific expression. Non-full-length systems generate sequences from the 3′ or 5′ ends of transcripts, which limits their ability to infer isoform expression, but as cells can be pooled after cDNA synthesis, the number of cells that can be processed is 2–3 orders of magnitude higher. The breadth of single-cell sequencing relates to the number of cells, tissues or samples that can be profiled, whereas depth relates to how much of the transcriptome is profiled for a given number of sequencing reads. Although the number of cells sequenced in an experiment is driven by the choice of method, it does allow some flexibility, but as the number of cells profiled rises, the increased sequencing cost usually restricts the depth of transcriptome profiling. Thus, different scRNA-seq systems can be viewed in terms of the two dimensions of breadth and depth. Typically, plate-based or microfluidic methods often capture the fewest cells but detect more genes per cell[168,173,176,177], whereas droplet-based systems can be used to profile the greatest number of cells[169,170] and have been used to generate individual data sets from more than one million cells[178].

The power of scRNA-seq is driving large-scale cell atlas projects, which aspire to determine the full complement of cell types in an organism or tissue. The Human Cell Atlas[179] and NIH Brain Initiative[180] projects intend to sequence all cell types present in the human body and brain, respectively. The Human Cell Atlas aims to sequence 30 to 100 million cells in phase 1 and will increase in breadth and depth as technologies develop. Recent results from this project include the discovery of ionocyte cells[159] and the finding that kidney cancer develops from different cell types in children and adults[181]. However, scRNA-seq users should be aware that the technologies can be applied to almost any organism. Recently, the analysis of *A. thaliana* by root cell protoplasting[182] demonstrated that even the tough cell wall of plant cells is an obstacle that can be overcome in order to generate single cells for sequence analysis. scRNA-seq is rapidly becoming a standard part of the biologist's toolkit and may be as widely used in 10 years' time as bulk RNA-seq is today.

Spatially resolved RNA-seq methods. Current bulk and scRNA-seq methods provide users with highly detailed data regarding tissues or cell populations but do not capture spatial information, which reduces the ability to determine how cellular context relates to gene expression. Two approaches to spatialomics methods are 'spatial encoding' and 'in situ transcriptomics'. Spatial-encoding methods record spatial information during RNA-seq library preparation, either by isolating spatially restricted cells (for example, by laser-capture microdissection (LCM)[183]) or by barcoding RNAs according to their location before isolation (via direct mRNA capture from tissue sections[184,185]) (FIG. 3b). In situ transcriptomics methods generate data within tissue sections by sequencing or imaging RNA in cells. We refer interested readers to recent in-depth reviews for a more comprehensive analysis of the field than is provided below[186–188].

LCM has been successfully used to isolate and profile individual cells or specific regions from tissue sections by RNA-seq[183,189–191]. Despite its requiring specialized equipment, LCM is widely available in many institutions. However, although it can achieve high spatial resolution, it is laborious and therefore difficult to scale. In both the Spatial Transcriptomics[184] (10X Genomics, USA) and Slide-seq[185] methods, mRNAs are directly captured for RNA-seq from frozen tissue sections applied directly to oligo-arrayed microarray slides or to 'pucks' of densely packed oligo-coated beads. The oligos comprise a spatial barcode, UMI and oligo-dT primers, which uniquely identify each transcript and its location. Sequence reads are mapped back to slide coordinates to generate spatial gene expression information. The Spatial Transcriptomics approach has been shown to work across tissues from a range of species, including mouse brain and human breast cancer tissue[184], human heart tissue[160] and *A. thaliana* inflorescence tissue[192]. Slide-seq is a recently developed technology that has been shown to work on frozen sections of mouse brain[185]. These direct mRNA capture methods do not require specialized equipment, have relatively simple analysis methods and are likely to be applicable at scale to many tissues.

Spatialomics
Transcriptome analysis methods that preserve the spatial information of individual transcripts within a given sample, usually a tissue section.

However, two important limitations remain to be solved. First, the technology can only be applied to fresh frozen tissue. Second, the resolution is limited by both the size of the array and the spacing of capture oligo spots or beads; the current arrays measure 6.5 × 7 mm and 3 × 3 mm, respectively, limiting the size of a tissue section that can be applied. The Spatial Transcriptomics spots are 100 µm in diameter and are spaced 100 µm apart, meaning they are not small enough or densely enough packed to achieve single-cell resolution. Slide-seq beads are much smaller, at just 10 µm in diameter, and are very densely packed, giving tenfold higher spatial resolution, and around half of the beads profiled appear to generate data from single cells. Computational methods that combine scRNA-seq from disaggregated tissue and spatial-encoding data improve resolution[185,193], but further developments in the underlying technology will be required in order to make this a more routine RNA-seq tool.

Alternatives to the spatially resolved RNA-seq methods described above include in situ sequencing[194] and imaging-based approaches that use single-molecule fluorescence in situ hybridization[195,196]. These methods generate narrower transcriptome profiles than do RNA-seq approaches, but they directly detect RNA, and targeted methods allow low-abundance transcripts to be profiled[197]. At the same time, they provide information on the tissue architecture and microenvironment and can generate subcellular data[198]. Substantial progress is being made[199,200], but the major limitations of imaging methods are the requirements for high- or super-resolution microscopy combined with automated fluidics, as well as the time taken for imaging, which can be many hours, or even days. Compared to sequencing costs, which have dropped faster than Moore's law predicts, the opportunities to scale imaging systems for high-throughput processing appear more limited.

All of the spatialomics methods described above are currently limited by an inability to generate deep transcriptome data, by cellular resolution and/or by very high costs (in time and/or money), but the methods are being rapidly improved and are already being applied to clinical samples[201]. Specific computational methods for spatialomics analysis are beginning to emerge[202,203]. Furthermore, advances in in situ RNA sequencing and imaging methods have already made it possible to generate transcriptome data for 10^3 to 10^5 cells, which is similar to the amounts of data available from droplet-based single-cell methods. Future development is likely to make spatialomics accessible to the more general user. However, truly single-cell, or subcellular, resolution is unlikely to be required by the majority of users. As such, the breadth of transcriptome profile and applicability to a wide range of tissues or samples may drive the development of these technologies in specific niches. Spatialomics is likely to be widely adopted if its technical limitations can be overcome.

Beyond steady-state RNA analysis

DGE studies use RNA-seq to measure steady-state mRNA levels, which are maintained by balancing the rates of mRNA transcription, processing and degradation. However, RNA-seq can also be used to study the processes and dynamics involved in transcription and translation, and these studies are providing new insights into gene expression.

Measuring active transcription with nascent RNA approaches. Gene expression is an inherently dynamic process, and measurement of DGE is limited in its ability to detect the subtle and rapid changes in complex transcriptional responses or to identify unstable noncoding RNAs such as enhancer RNAs. RNA-seq can be used to map TSSs and quantify newly transcribed nascent RNA, which enables the investigation of RNA dynamics. However, compared to DGE analysis, the cataloguing of nascent RNAs is challenging, because of their short half-lives and low abundance. As such, the importance of understanding these dynamics has led to multiple methods being developed to analyse nascent RNA; these methods have revealed the extent of divergent transcription at promoters[204], shown that promoter-proximal pausing of transcriptionally active RNA polymerase II (Pol II) is a key regulatory step for gene expression[205], demonstrated that nascent RNA has a direct role in regulating transcription and shown that its sequence and structure affect transcription elongation, pausing and stalling, as well as the binding of chromatin modifiers and enhancer RNAs[206]. Nascent RNA-seq methods that aim to distinguish between newly transcribed RNA and other RNAs can be broadly split into three categories: 'run-on' methods, Pol II immunoprecipitation (IP)-based methods and metabolic-labelling approaches (FIG. 4).

Run-on methods rely on the incorporation of nucleotide analogues that enable nascent RNA to be enriched from the total RNA pool and that allow measurement of transient RNA transcription (FIG. 4a). Global run-on sequencing (GRO-seq)[204,207,208] and precision nuclear run-on sequencing (PRO-seq)[209] achieve this by incorporating 5-bromouridine 5′-triphosphate (BrU) or biotin-modified nucleotides, respectively, into nascent RNA during transcription. Nuclei are isolated and endogenous nucleotides are removed by washing, before the exogenous biotin-tagged nucleotides are added and transcription is resumed. Immunoprecipitation or affinity purification and sequencing of the enriched newly transcribed RNA allows the position and activity of transcriptionally engaged RNA polymerases to be determined transcriptome-wide. Owing to the number of nucleotides labelled during run-on, GRO-seq can only achieve 10–50 bp (REF.[210]) resolution, which reduces the precision of TSS mapping. PRO-seq achieves base-pair resolution because transcription is stopped upon biotin-nucleotide incorporation, allowing identification of the incorporation site. Run-on methods are conceptually simple — only RNA molecules incorporating the modified nucleotides should be enriched for sequencing, but in practice the presence of background non-nascent RNA increases the read depth required. Use of these methods has revealed the extent of divergent or bidirectional transcript initiation at promoters and has identified the role of enhancer RNAs in modulating gene expression[211]. By incorporating specific enrichment for 5′-capped RNAs, GRO-cap[212], PRO-cap[209] or small 5′-capped RNA

Nascent RNA
RNA that has just been transcribed, as opposed to RNA that has been processed and transported to the cytoplasm.

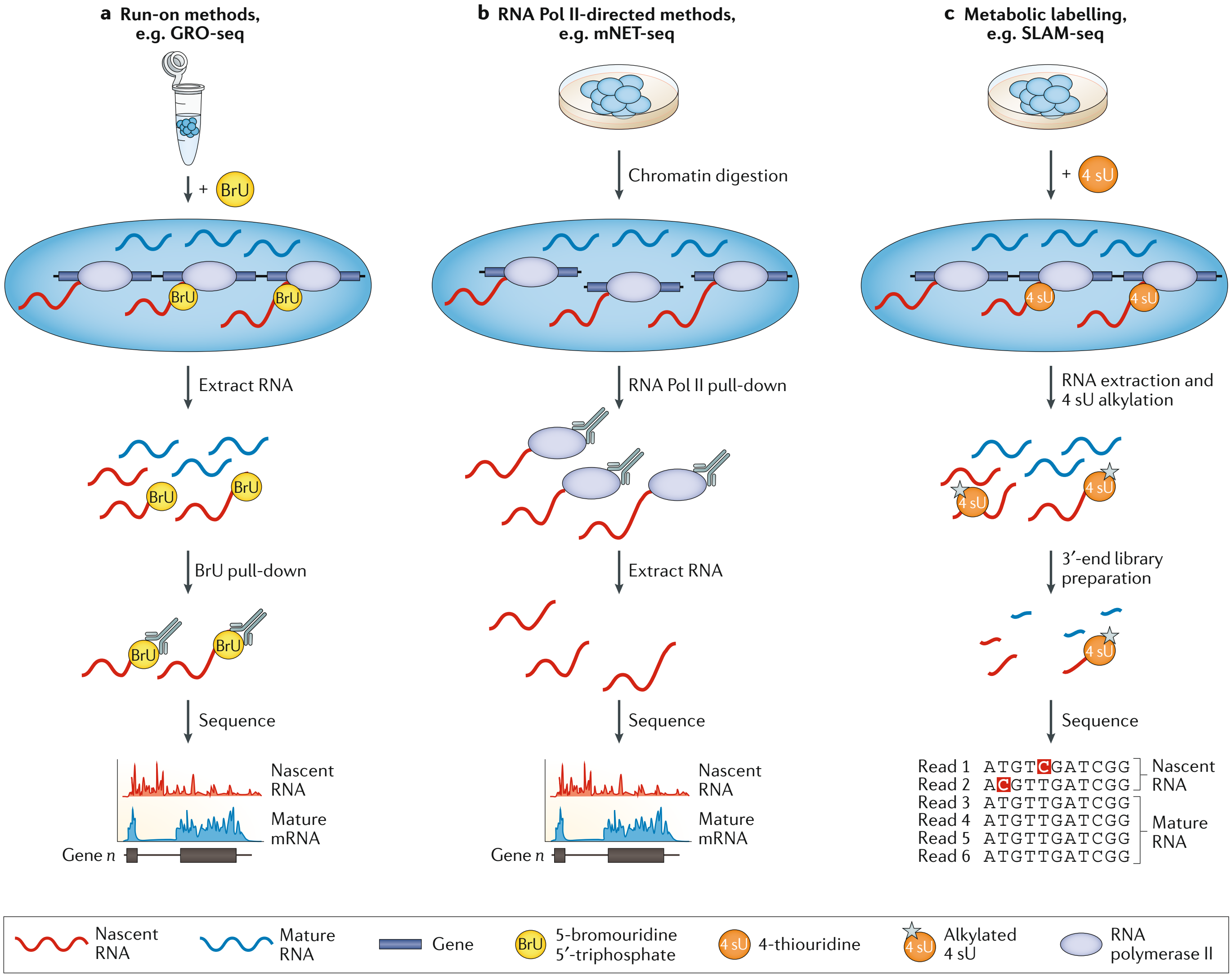


Fig. 4 | **The key concepts of nascent RNA and translatome analysis.** Nascent RNA analysis methods enrich newly transcribed RNAs from the other RNA in a cell and compare this to an unenriched (mature RNA) control, by one of three primary methods. **a** | Run-on methods label RNA by adding a time-limited pulse of modified ribonucleotides into cell media; various modified nucleotides can be used, but global run-on sequencing (GRO-seq)[203] and its corresponding 5-bromouridine 5′-triphosphate (BrU) nucleotide are shown. After incorporation of the label, nascent-RNA strands are enriched by immunoprecipitation (IP) with antibodies specific to the modified nucleotide used and are prepared for RNA-sequencing (RNA-seq) analysis. **b** | RNA polymerase II (Pol II) IP methods pull down Pol II-associated RNAs after chromatin digestion with micrococcal nuclease. During chromatin digestion, the nascent RNA is protected from nuclease activity by its Pol II footprint. The protected RNA is extracted and processed for RNA-seq analysis. **c** | Metabolic labelling methods label RNA similarly to run-on methods, but they use the nucleotide analogue 4-thiouridine (4 sU). Alkylation of 4 sU after RNA extraction prompts misincorporation of G nucleotides during reverse transcription, allowing 4 sU incorporation sites to be directly determined by mutational analysis with base-pair resolution. Preparation of a 3′-end RNA-seq library increases the signal by reducing the amount of unlabelled RNA carried through to sequencing. Figure adapted from REF.[214], Springer Nature Limited. Part **a** adapted with permission from REF.[222], AAAS.

sequencing (START-seq)[213] increase sensitivity and specificity for detecting transcription initiation and capture RNAs that would be removed by co-transcriptional processing, as well as reducing the background signal from post-transcriptionally capped RNAs.

Pol II IP methods, such as native elongating transcription sequencing (NET-seq)[214] and native elongating transcript sequencing for mammalian chromatin (mNET-seq)[215], pull down any Pol II-associated RNA using anti-FLAG (for FLAG-tagged Pol II) or various antibodies directed against the Pol II C terminal domain (CTD) (FIG. 4b). RNA-seq of the nascent RNA associated with these chromatin complexes is used to map TSSs, although non-nascent Pol II-associated RNA and background mRNA negatively affect read depth and confound the analysis. NET-seq can lack specificity, in that any RNA strongly associated with Pol II can contaminate the nascent-RNA enrichment, as evidenced by the presence of tRNA and small nucleolar RNA in NET-seq data[216]. The use of multiple CTD antibodies

4-Thiouridine
(4 sU). A thio-substituted nucleoside not naturally found in eukaryotic mRNAs, which is easily incorporated into nucleic acids and is used in nascent RNA analysis.

Translatome
The complete set of proteins translated from mRNA in a cell, tissue or organism.

in mNET-seq reveals how CTD modifications affect transcription, detects RNA-processing intermediates and enables locating specific Pol II nascent RNAs to TSSs. However, these abilities come at the price of more complex experiments, the need for larger numbers of cells and higher overall sequencing costs.

Metabolic pulse-labelling with the nucleotide analogue 4-thiouridine (4 sU) allows the identification of nascent RNA[217] (FIG. 4c). However, in methods requiring long labelling times, most of the transcript will be labelled, which limits sensitivity. By specifically targeting the 3′ end of RNAs (that is, only the newly transcribed RNA closest to the RNA polymerase), both transient transcriptome sequencing (TT-seq)[218] and thiol(SH)-linked alkylation for metabolic sequencing of RNA (SLAM-seq)[71] reduce the signal from 5′ RNA. TT-seq restricts the labelling time to 5 minutes, so that only the 3′ end of new transcripts is labelled, and it includes an RNA fragmentation step before biotin affinity purification, to enrich for labelled RNA. SLAM-seq incorporates a 3′ mRNA-seq library preparation (although it can be used with other library preparations, such as for miRNA[219]), directing sequencing only to labelled newly transcribed RNA instead of to the whole transcript. Additionally, in SLAM-seq, iodoacetamide is added after RNA extraction, to alkylate 4 sU residues that have been incorporated into the growing nascent-RNA strand. This modification induces reverse-transcription-dependent thymine-to-cytosine (T>C) nucleotide substitutions, which are detected as 'mutations' in a sequencing analysis, thereby directly identifying the 4 sU incorporation sites. However, a low incorporation rate means that only a small number of 4 sU sites are available to be converted to cytosines[71], which limits sensitivity. Two methods, TUC-seq[220] and TimeLapse-seq[221], also use T>C mutational analysis but do not enrich for 3′ ends. They have been used to interrogate the transcriptional responses to cellular perturbation[222] and to measure RNA half-lives[71].

Methods for nascent-RNA analysis have not yet been directly compared. Nascent-RNA methods are all negatively affected by the enrichment of nonspecific background and/or degraded RNA, which can impact read-depth requirements[210]. By focusing sequencing to the 3′ ends only, the effects of non-nascent RNA are reduced in PRO-seq, TT-seq and SLAM-seq, but there is little evidence to suggest whether any approach outperforms the others. Affinity pull-down is laborious and requires higher quantities of starting material than metabolic labelling does, but determining the timing of pulse labelling is complex, and short pulses generate very little RNA for analysis, which limits sensitivity. Recent developments that enable tissue-specific RNA labelling[223], as well as new computational methods for 'mutational' analysis[224], may persuade users to switch biochemical (biotin-based) enrichment for bioinformatic enrichment of nascent and other RNAs. Further development of nascent-RNA methods and their combination with other methods, such as spatialomics[225] or RNA–RNA and RNA–protein interaction methods, will improve our understanding of the processes involved in transcription.

Measuring active translation with ribosome-profiling methods. The primary emphasis of RNA-seq is on the species and quantities of mRNAs that are extant in a sample, but the presence of mRNAs does not correspond straightforwardly to protein production. Two methods move beyond transcription and allow us to understand the translatome: polysomal profiling[226,227] and ribosome footprinting by RNA-seq (Ribo-seq[228]). Translation of mRNAs by ribosomes is highly regulated, and protein levels are predominantly defined by translation activity. Polysome profiling and Ribo-seq allow users to interrogate how many ribosomes occupy a transcript and their distribution along the transcript (FIG. 5). This allows users to infer which transcripts are being actively translated at a particular time or cell state. Both methods make the assumption that mRNA ribosome density correlates to the protein synthesis level. Comparing samples reveals ribosomal dynamics under treatment, over time, in development or in a disease[229] in which translation dysregulation is implicated, such as fibrosis[230], prion disease[231] or cancer[232,233].

Polysome profiling uses sucrose gradient ultracentrifugation to separate mRNAs bound by multiple ribosomes (the polysomal fraction) from those bound by a single, or no, ribosome (the monosomal fraction) for RNA-seq library preparation[228] (FIG. 5a). The mRNAs detected at higher abundance in the polysomal fraction are presumed to be more highly translated than those in the monosomal fraction. The method allows the translational status of individual mRNAs to be inferred and also generates high-resolution maps of ribosome occupancy and density (although it does not allow the locations of ribosomes to be determined). Several improvements have been made to the original method. For example, the use of nonlinear sucrose gradients improves the ease of polysomal mRNA collection at the interface of the different-concentration sucrose solutions[234], the application of Smart-seq[235] library preparation enables analysis of just 10 ng of polysomal mRNA[234] and the use of higher-resolution sucrose gradients and deep sequencing allows transcript-isoform-specific translation to be measured[236,237]. Nevertheless, polysome profiling generates a relatively low-resolution translation profile and is a laborious method that requires specialized equipment, which places limits on replication studies.

Ribo-seq is based on RNA footprinting[238] and was initially developed in yeast[228]. It uses cyclohexamide to inhibit translation elongation and cause ribosomes to stall on mRNAs. Digestion of mRNA with RNase I leaves ribosome-protected footprints of 20–30 nucleotides, which are processed to generate an RNA-seq library (FIG. 5b). Ribo-seq generates a high-resolution translation profile[239], mapping both ribosome abundance and locations on individual transcripts. Mapping ribosome location, which is not possible with polysome profiling, means it is possible to detect translation pausing, which can regulate protein expression. Protocol modifications include buffer and enzyme optimization, which more clearly reveals the 3-bp periodicity of Ribo-seq data[239], as well as barcoding and the use of UMIs[240], which allow individual molecular events to be determined.

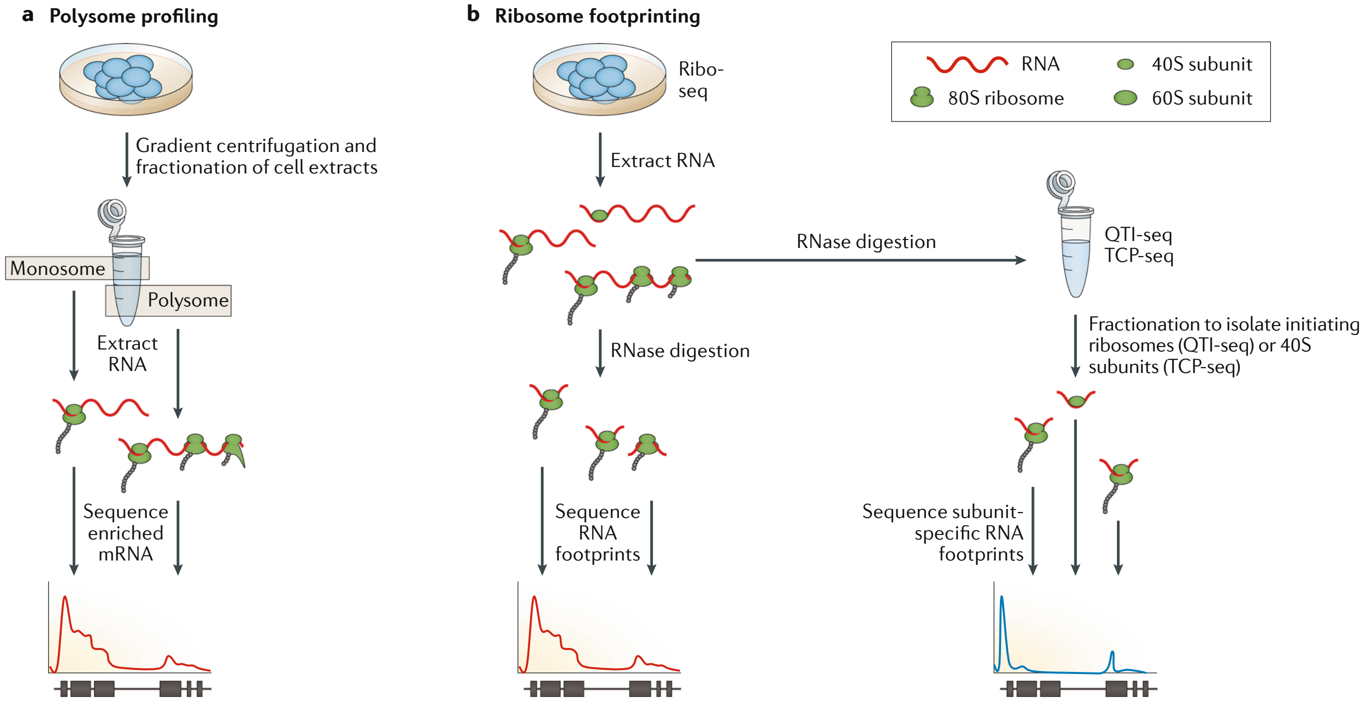


Fig. 5 | **The key concepts of translatome analysis.** Translatome analysis methods generate RNA-sequencing (RNA-seq) data from ribosomally bound RNA, with an assumption that mRNA ribosome density correlates with the protein synthesis level. **a** | Polysome profiling[238] separates RNA molecules by centrifugation into polysomal fractions, which are compared by RNA-seq. RNAs more highly expressed in the polysomal fractions are presumed to be more actively transcribed. **b** | Ribosome footprinting (Ribo-seq) methods use RNase to digest exposed RNA, while leaving ribosome-protected RNA undigested. Sequencing of the protected RNA reveals both the density and location of ribosomes. By modifying the standard Ribo-seq protocol, quantitative translation initiation sequencing (QTI-seq)[245] or translation complex profile sequencing (TCP-seq)[246] makes it possible to specifically enrich for initiating ribosomes, or for their subunits, while depleting elongating ribosomes, thereby allowing a more detailed analysis of translation dynamics. Computational analysis of translatome RNA-seq data identifies the relative translation of individual mRNAs and can determine translation initiation, elongation and termination dynamics.

Standard RNA-seq tools can be used for computational analysis, although specific tools have recently been developed for finding open reading frames[241,242], for differential[243,244] or isoform-level translational analysis[240] and for investigating codon bias[245]. The main limitations of Ribo-seq are the requirement for ultracentrifugation and the need to empirically determine the RNase I digestion conditions, due to batch-to-batch variability of nucleases[240].

These methods average the signal from translation initiation, elongation and termination, but modifications to Ribo-seq enable translational dynamics to be investigated. Quantitative translation initiation sequencing (QTI-seq[246]) maps transcription initiation sites by chemically ‘freezing’ and enriching the initiating ribosomes while removing elongating ribosomes from the associated mRNA. Translation complex profile sequencing (TCP-seq[247]) also maps initiation sites by enriching RNA associated with the 40S ribosomal small subunit before the mature ribosome is assembled. However, as ribosome integrity is preserved in this method, the 80S ribosomal fraction can also be analysed and compared, allowing a fuller picture of translational dynamics to be obtained (FIG. 5b).

All translatome methods are conceptually similar; they make the assumption that mRNA ribosome density correlates to protein synthesis level. Although their sample preparation protocols differ, all require quite large numbers of input cells. Ultimately, their combination with RNA-seq to understand gene expression levels, and with proteomics to determine protein levels, is likely to be required for a comprehensive view of mRNA translation. For a more detailed overview of translatome analysis, we direct readers to recent review articles[241,248].

Beyond analysis of gene expression

RNAs play an important role in the regulation of other biomolecules and of biological processes, such as splicing and translation, that involve the interaction of RNA with various proteins and/or other RNA molecules. RNA-seq can be used to interrogate intramolecular and intermolecular RNA–RNA interactions (RRIs), which can reveal insights into the structurome, or interactions with proteins, enabling deeper insights into transcription and translation (FIG. 6). The various methods developed for interactome analysis share a common theme: enrichment of RNA that is interacting relative to RNA that is not. Some methods make use of native biological interactions, others generate transient or covalent bonds between molecules of interest; most use antibody pull-down, affinity purification or probe hybridization to enrich RNA for sequencing. Here we briefly describe

Structurome
The complete set of secondary and tertiary RNA structures in a cell, tissue or organism.

Interactome
The complete set of molecular interactions in a cell, tissue or organism, including RNA–RNA or RNA–protein interactions.

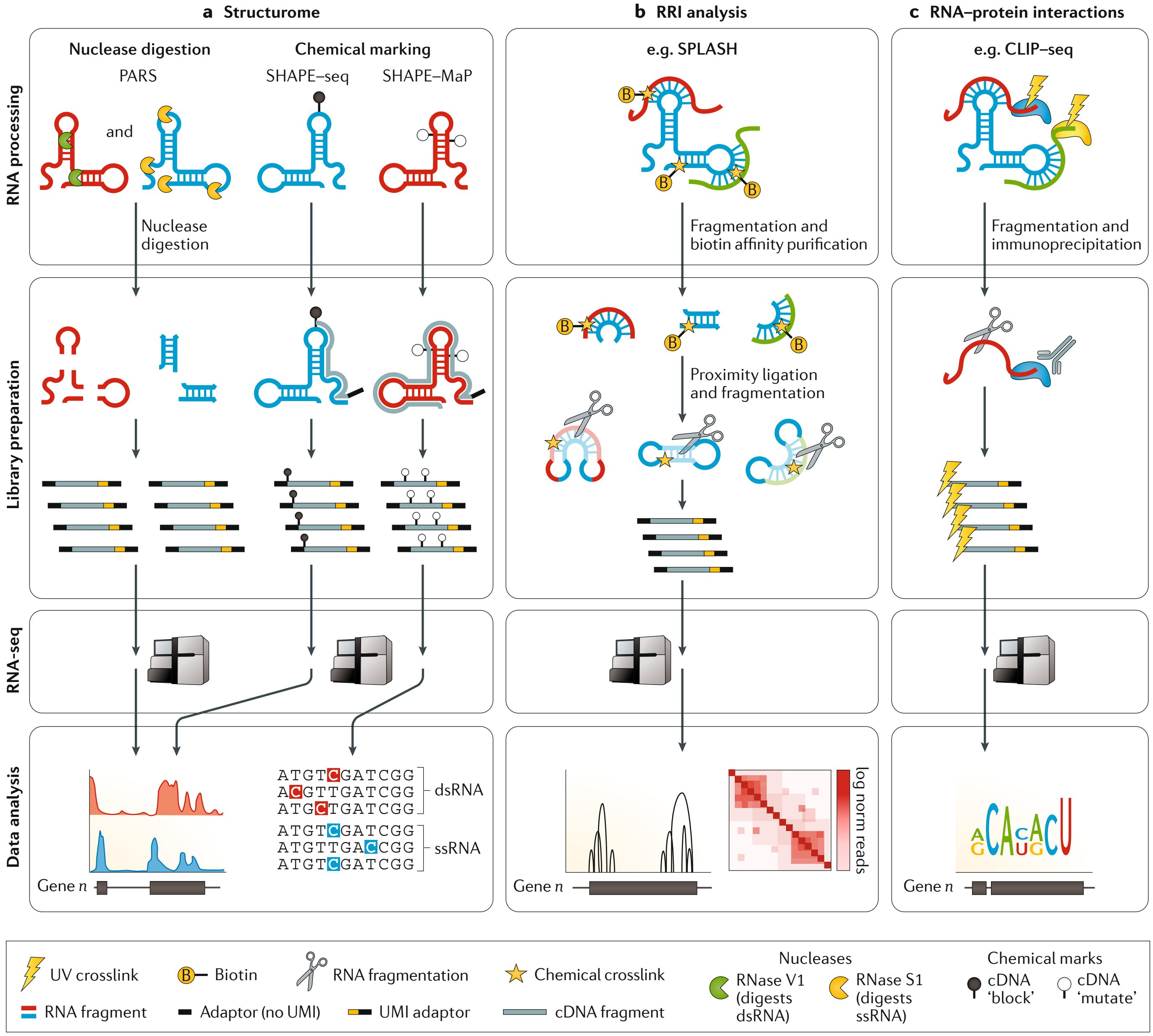


Fig. 6 | **The key concepts of RNA structure and RNA–protein interaction analysis. a** | Structurome analysis uses nucleases or chemical-marking reagents to probe structured (that is, double-stranded RNA (dsRNA)) and unstructured (that is, single-stranded RNA (ssRNA)) RNA in a transcriptome-wide manner. In most experiments, both ssRNA and dsRNA are probed in separate reactions, and the results are combined in a reactivity analysis to reveal the structural features. Nuclease digestion methods probe RNA structure using one or more nucleases specific to dsRNA and/or ssRNA. For example, in parallel analysis of RNA structure (PARS)[253], parallel samples are cleaved in vitro with either RNase V1 (a dsRNA-specific nuclease) or S1 nuclease (an ssRNA-specific nuclease). The RNA that remains after digestion is converted to cDNA and sequenced, with the read depth being proportional to the reactivity of the aligned regions. Overlay and comparison of RNA-sequencing (RNA-seq) data allow structures to be inferred. Chemical-mapping methods, such as selective 2′-hydroxyl acylation analysed by primer extension and followed by sequencing or mutational profiling (SHAPE–seq[255] or SHAPE–MaP, respectively)[258], modify in vitro or in vivo ribonucleotides in double-stranded or single-stranded regions in a structure-dependent manner. Marks can either block reverse transcription, leading to truncated cDNAs, or cause misincorporation mutations at the modified sites. The RNA is converted to cDNA and sequenced, with the read depth or mutation rate being proportional to the reactivity of the aligned regions, allowing structures to be inferred. **b** | RNA–RNA interaction (RRI) analysis methods, such as SPLASH[270], begin by crosslinking interacting RNA molecules with biotinylated psoralen, which are then enriched by streptavidin pull-down before proximity ligation joins the free ends of the interacting RNAs. Further fragmentation is followed by RNA adaptor ligation and circularization to prepare an RNA-seq library for analysis, which reveals sites of both intramolecular (that is, structural) RNA interactions and intermolecular RNA–RNA interactions. **c** | RNA–protein interaction methods, such as crosslinking immunoprecipitation of RNA followed by sequencing (CLIP–seq)[75], generate covalent crosslinks between interacting RNA and proteins using ultraviolet (UV) radiation. After RNA fragmentation, antibody pull-down of the target protein co-purifies the bound RNA, which is 3′-adaptor ligated and extracted for cDNA synthesis. Reverse transcription from the adaptor generates cDNAs from the protein-bound RNA, which is prepared for RNA-seq analysis. UMI, unique molecular identifier.

the major RNA-seq-based methods for investigating the structurome and interactome.

Probing RNA structure via intramolecular RNA interactions. Ribosomal RNA and tRNAs form most of a cell's RNA. Together with other structured non-coding RNAs, they perform various roles in the cell, from gene regulation to translation[249]. Two primary approaches exist that allow users to interrogate RNA structure: nuclease-based and chemical-probing methods. Ribonuclease digestion was first used to determine RNA structure (of tRNA^{Ala})[250] in 1965. Chemical methods, such as selective 2′-hydroxyl acylation analysed by primer extension (SHAPE) chemistry, were developed over the next 40 years and used to determine tRNA^{Asp} structure at base-pair resolution[251]. But only combination of the various nuclease and chemical methods with RNA-seq has enabled methods to move from single-RNA to transcriptome-wide analysis, which is transforming our understanding of the complexity and importance of the structurome. Here we focus on the major differences between the nuclease and chemical-mapping approaches (FIG. 6a), and we direct readers to Strobel et al.[252] for a comprehensive review of the field.

Nuclease methods, such as parallel analysis of RNA structure (PARS)[253] and fragmentation sequencing (FRAG-seq)[254], use enzymes that digest either single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA). The RNA remaining after nuclease digestion is used as the input for RNA-seq library preparation. The structured (double-stranded) and unstructured (single-stranded) regions are subsequently identified by computational analysis of the resulting RNA-seq data. Nucleases are easy to use and allow interrogation of both ssRNA and dsRNA, but they have lower resolution than chemical mapping[254], due to the random nature of nuclease digestion. Furthermore, their large size restricts entry to the cell, making them unsuitable for in vivo studies.

Chemical-mapping methods use chemical probes that react with RNA molecules and mark structured or unstructured nucleotides. These marks either block reverse transcription or result in cDNA misincorporation, allowing the mapping and analysis of RNA-seq reads to reveal the structurome. SHAPE followed by sequencing (SHAPE–seq)[255] marks unpaired ssRNA by reacting with the ribose 2′-hydroxyl of the RNA backbone, although base-stacking in hairpin loops can reduce its efficiency[252]. Structure–seq[256] and dimethyl sulfate sequencing (DMS-seq)[257] mark adenine and cytosine residues with DMS, blocking reverse transcription and enabling RNA structure to be inferred from analysis of the resulting truncated cDNAs. SHAPE and mutational profiling (SHAPE–MaP)[258] and DMS mutational profiling with sequencing (DMS–MaPseq)[259] both modify the experimental conditions to improve reverse-transcriptase processivity and prevent cDNA truncation. Instead, the chemical marks result in misincorporation events, and these 'mutations' can be detected during RNA-seq data analysis to reveal RNA structure. The chemical probes are small molecules, enabling a more biologically meaningful structurome to be determined in vivo, although the data can be more variable due to the dynamic intracellular environment. They can also be used to perform structural analysis of nascent RNAs and reveal the ordering of cotranscriptional RNA folding[260].

Nuclease and reverse-transcription blocking methods generally produce short RNA fragments and only report on a single digestion site or chemical mark, whereas misincorporation and mutation detection methods can report on multiple chemical marks per read. None of the methods is without bias; reverse-transcription blocking is never 100% efficient, chemical marks that should induce mutations can block cDNA synthesis and both of these factors can impact data interpretation. Spike-in controls are likely to improve the quality of structurome analysis[261] but are not yet widely used. A comparison of the SHAPE methods reveals differences in efficiency that are apparent only in in vivo experiments[262], highlighting the care needed in comparing such complex methodologies.

These methods are generating novel insights into how RNA structure plays a role in gene and protein regulation. For example, analysis of DMS mapping data suggests that RNA structure may regulate APA[256] and may slow translation in catalytically active regions, allowing more time for protein folding and thereby reducing misfolding events[263]. A combination of structural RNA-seq methods is likely to be necessary to generate a complete picture of the structurome. As the field expands, we are likely to discover links between RNA structure and development or disease states; recent results have suggested a potential role for aberrant RNA structures in repeat expansion diseases[264]. Ultimately, structurome analysis may enable small-molecule targeting of well-characterized RNA structures, opening up a new area of therapeutic development[265].

Probing intermolecular RNA–RNA interactions. Intermolecular RRIs play important roles in post-transcriptional regulation, such as miRNA targeting of 3′ UTRs. Tools for investigating intermolecular RRIs have been developed for both targeted[266–268] and transcriptome-wide[269–271] analysis. These methods share a common workflow, in which RNA molecules are crosslinked so as to preserve interactions, before fragmentation and proximity ligation (FIG. 6b). Most, but not all, of the chimeric cDNAs generated by the different methods are derived from ligation of stably base-paired (that is, interacting) RNA molecules. Targeted methods such as crosslinking, ligation and sequencing of hybrids (CLASH)[266], RNA interactome analysis and sequencing (RIA–seq)[267], and RNA antisense purification followed by RNA sequencing (RAP–RNA)[268] can generate high-depth interaction maps of a single RNA species, or family of RNAs. CLASH enriches for RRIs mediated by specific protein complexes using IP, whereas RIA–seq[267] uses antisense oligonucleotides to pull down RNAs interacting with the target; neither method distinguishes between direct and indirect RRIs, which complicates biological interpretation. To increase the resolution of RRI analysis, RAP–RNA[268] uses psoralen and other crosslinking agents, followed by RNA capture with antisense oligonucleotides, and high-throughput RNA-seq to detect both direct and indirect RRIs. Although the method does allow for a more specific analysis, it requires the preparation of multiple libraries (one for each crosslinking agent).

Transcriptome-wide methods are fundamentally similar to targeted methods: interacting RNAs are crosslinked in vivo and enriched. Enrichment improves specificity by reducing the amount of non-interacting RNA carried through into the ligation reaction and can be achieved by 2D-gel purification (as in psoralen analysis of RNA interactions and structures (PARIS)[269]) or biotin affinity purification of the crosslinked RNAs (as in sequencing of psoralen crosslinked, ligated and selected hybrids (SPLASH)[270]), or by depletion of non-crosslinked RNAs by RNase R digestion (as in ligation of interacting RNA followed by RNA-seq (LIGR–seq)[271]). After ligation, crosslinks are reversed before RNA-seq library preparation and sequencing. PARIS generates the highest number of interactions of any method[272], but it requires 75 million reads per sample, more than any other RRI method and more than twice the average read depth of DGE experiments.

Analysis of collated RNA interaction data allows multiple interactions to be visualized[273] and has revealed the variation in distribution of RRIs by RNA species[272]. In all, 90% of RRIs involve mRNAs. Nearly half involve miRNAs or long non-coding RNAs, for which most interactions are with an mRNA target. Comparison of these collated data reveals the biases in the different methods for specific RNA species, which results in very little overlap between methods. Hence, a complete picture of RRIs is likely to require the use of more than one method. However, there are several limitations of RRI methods. Perhaps the most challenging is that RRIs are dynamic and are affected by structural conformation and other intermolecular interactions[269], making interpretation difficult without replication[274]. Intramolecular interactions add noise to intermolecular RRI analysis, which requires highly structured RNAs, such as rRNAs, to be filtered and removed[271]. Other issues include the disruption of interactions during RNA extraction, requiring stable crosslinking methods, but the most commonly used RRI crosslinking reagents — psoralen and 4′-amino-methyltrioxsalen (AMT) — only crosslink pyrimidines with low efficiency, reducing sensitivity. Additionally, the proximity ligation step is inefficient and ligates both interacting and non-interacting RNAs, further reducing sensitivity[269–271].

Probing RNA–protein interactions. ChIP–seq[275] has become an indispensable tool for mapping and understanding DNA–protein interactions; a similar IP approach is used to interrogate RNA–protein interactions. RNA–protein interaction methods rely on IP, utilizing an antibody against the RNA-binding protein of interest to capture its bound RNA for analysis (first demonstrated with microarrays[276]) (FIG. 6c). The most obvious difference between the various RNA–protein interaction methods relates to whether and how the interacting RNA and proteins are crosslinked: some methods avoid crosslinking (native IP), others use formaldehyde for crosslinking, and some use ultraviolet (UV) light for crosslinking. The simplest method, RNA immunoprecipitation and sequencing (RIP–seq)[277], often, but not always, uses native IP and does not include RNA fragmentation. This simplicity makes the method easy to adopt. The method generates useful biological insights, but it has two important drawbacks. First, the mild washing conditions used to preserve RNA–protein interactions mean that a relatively high level of nonspecifically bound fragments is enriched. Second, the absence of RNA fragmentation reduces binding site resolution. Thus, RIP–seq is highly variable and dependent on the natural stability of RNA–protein binding[278]. The use of formaldehyde crosslinking to produce a reversible covalent bond between an RNA and its interacting proteins[279] improves stability and reduces nonspecific RNA pull-down, but formaldehyde also generates protein–protein crosslinks. The impact of this can be mitigated by mild crosslinking with 0.1% formaldehyde (tenfold lower than that used for ChIP–seq studies), which generates high-quality results across multiple protein targets[280].

The introduction of 254-nm UV crosslinking in CLIP[75,281] was a critical development that increased the specificity and positional resolution of RNA–protein interaction analysis methods. UV crosslinking creates a covalent bond between protein and RNA at their interaction site but, crucially, does not crosslink protein–protein interactions. This stabilizes RNA–protein binding, allowing for stringent enrichment that disrupts native RNA–protein interactions, reducing the background signal. The CLIP protocol has subsequently been the basis of much methodological development. Individual-nucleotide resolution CLIP (iCLIP)[282] incorporates UMIs into the library preparation to remove PCR duplicates. It also takes advantage of the common premature truncation of cDNA synthesis at crosslinked nucleotides to gain quantitative, nucleotide-resolution mapping of the crosslinked sites through amplification of truncated cDNAs. Photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP)[283,284] attains nucleotide resolution by using 4 sU and 356-nm UV crosslinking. The 4 sU is incorporated into endogenous RNAs during cell culture, and 356-nm UV irradiation only generates crosslinks at 4 sU incorporation sites (leading to high specificity). Detection of the reverse-transcription-induced T>C substitutions in the resulting sequence data enables base-pair resolution and allows discrimination of crosslinked versus non-crosslinked fragments, further reducing background signal. More recent improvements to CLIP have increased its efficiency and sensitivity. Infrared CLIP (irCLIP)[285] replaces radioisotopic detection with infrared gel visualization and gel purification with bead-based purification. These changes made the protocols easier to adopt and enabled RNA–protein interaction analysis from as few as 20,000 cells, compared to the 1–2 million cells normally used for iCLIP. Enhanced CLIP (eCLIP)[286] removes quality control and visualization of RNA–protein complexes, incorporates barcodes in the RNA adaptors that allow samples to be pooled earlier in the protocol and replaces gels with beads. These changes were aimed at simplifying the protocols for the user, and eCLIP experiments for almost 200 proteins have been produced as part of the ENCODE project[287]. However, neither irCLIP nor eCLIP is currently widely adopted, partly because some of the increase in sensitivity of eCLIP and irCLIP might result from decreased specificity, as suggested by the decreased enrichment of binding motifs and regulated exons at binding sites of PTBP1 identified by the two methods[288]. As the abundance of publicly available data opens new opportunities for computational analyses, it is important to carefully consider the approaches taken to quality control, filtering, peak calling and normalization of CLIP data, which all affect biological interpretation of the data[288]. For a fuller discussion of CLIP protocols for RNA–protein interactions, we point interested readers to a recent comprehensive review of the topic[289].

The dependence of some RRI, and all RNA–protein binding, methods on IP restricts studies to proteins with well-characterized antibodies, and nonspecific antibody binding remains an issue — although not one confined to this field. RNA structure also affects RNA–protein interactions; some proteins recognize specific RNA secondary structures or compete with those structures for access to RNA, which complicates the transfer of in vitro discoveries to in vivo biology[290,291]. Furthermore, both structural and RNA–protein interaction methods generally report averaged data for a specific transcript or position. Future developments in laboratory methods, in computational approaches and in single-molecule sequencing may help decipher some of this biological variation.

Conclusions

Wang, Gerstein and Snyder were certainly correct in their prediction that RNA-seq would "revolutionise [how] eukaryotic transcriptomes are analysed"[292]. However, even they have most likely been surprised by the scale of the transformation. Today it is possible to analyse the many aspects of RNA biology that are essential to gaining a functional understanding of the genome, to investigating development and to determining the molecular dysregulation that causes cancer and other diseases. While the biological discovery phase is far from over, already RNA-seq tests are being used in the clinic[293,294]. Single-cell sequencing is becoming standard in many laboratories, and spatialomics analysis is likely to follow a similar path, enabling its use outside the laboratories responsible for developing the current methods. It is also possible that long-read sequencing methods will replace Illumina short-read RNA-seq as the default method for a substantial proportion of users. For this scenario to occur, considerable improvements need to be made in long-read sequencing, in terms of increasing throughput and decreasing error rates. However, the advantages of long-read mRNA isoform sequencing are such that, if it becomes as cheap and reliable as short-read sequencing is today, then, for anything other than degraded material, it is likely to be the preferred choice. With this in mind, any predictions of how RNA-seq might develop over the next decade are likely to be too conservative.

Published online: 24 July 2019

1. Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**, 69–73 (2007).
2. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
3. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1350 (2008).
4. Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
5. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
6. Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
7. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
8. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
9. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437 (2014).
10. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* **17**, 207–223 (2016).
11. Illumina. For all you seq. *Illumina* https://emea.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html (2014).
A tour de force that includes a graphical abstract, a brief description and primary references for most sequencing methods.
12. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
The first report of Oxford Nanopore direct sequencing of RNA molecules without reverse transcription or amplification. It reports full-length, strand-specific RNA sequencing and detection of RNA nucleotide analogues.
13. Smith, A. M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/132274 (2017).
14. Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
15. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
16. Cartolano, M., Huettel, B., Hartwig, B., Reinhardt, R. & Schneeberger, K. cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PLOS ONE* **11**, e0157779 (2016).
A paper comparing the performance of reverse transcriptases for long-read RNA-seq, using Pacific Biosciences Iso-Seq, and discussing the challenges of sequencing full-length transcripts, due to RNA degradation, shearing and incomplete cDNA synthesis.
17. Dard-Dascot, C. et al. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19**, 118 (2018).
18. Giraldez, M. D. et al. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat. Biotechnol.* **36**, 746–757 (2018).
19. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.* **23**, 133–140 (2015).
20. Hör, J., Gorski, S. A. & Vogel, J. Bacterial RNA biology on a genome scale. *Mol. Cell* **70**, 785–799 (2018).
21. Saletore, Y. et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* **13**, 175 (2012).
22. Schwartz, S. & Motorin, Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA Biol.* **14**, 1124–1137 (2017).
23. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
24. Su, Z. et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
A thorough comparison of RNA-seq platforms and methods, which assesses multiple performance and quality metrics using cell line and control RNAs across multiple sequencing instruments and multiple laboratories.
25. Li, S. et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).
26. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
27. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. & Vitale, L. GeneBase 1.1: a tool to summarize data from NCBI Gene datasets and its application to an update of human gene statistics. *Database* **2016**, baw153 (2016).
28. Gazzoli, I. et al. Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol.* **13**, 290–305 (2016).
29. Tilgner, H. et al. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**, 231–242 (2018).
30. Wu, I., Ben-yehezkel, T., Genomics, L. & Jose, S. A. Single-molecule long-read survey of human transcriptomes using LoopSeq synthetic long read sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/532135 (2019).
31. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl Acad. Sci. USA* **108**, 9026–9031 (2011).
32. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
33. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
34. Smith, G. R. & Birtwistle, M. R. A mechanistic beta-binomial probability model for mRNA sequencing data. *PLOS ONE* **11**, e0157828 (2016).

35. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).
36. Engström, P. G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).
37. Thomas, S., Underwood, J. G., Tseng, E. & Holloway, A. K. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLOS ONE* **9**, e94650 (2014).
38. Matz, M. et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Proc. Natl Acad. Sci. USA* **27**, 1558–1560 (1999).
39. Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
40. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical. *Nucleic Acids Res.* **46**, 2159–2168 (2018).
41. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16**, 204 (2015).
42. Prazsák, I. et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* **19**, 873 (2018).
43. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
44. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
45. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Preprint at *bioRxiv* https://doi.org/10.1101/459529 (2018).
46. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100 (2017).
A paper providing an assessment of the power of long-read sequencing in transcriptome analysis. It reports hybrid sequencing through the combination of Illumina short reads with Pacific Biosciences or Nanopore long reads.
47. Wongsurawat, T., Jenjaroenpun, P., Wassenaar, T. M. & Taylor, D. Decoding the epitranscriptional landscape from native RNA sequences. Preprint at *bioRxiv* https://doi.org/10.1101/487819 (2018).
48. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl Acad. Sci. USA* **111**, 9869–9874 (2014).
49. Au, K. F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl Acad. Sci. USA* **110**, E4821–E4830 (2013).
50. Sahraeian, S. M. E. et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* **8**, 59 (2017).
A paper that assesses RNA-seq workflows that incorporate RNA variant calling, editing and fusion detection, covering both short- and long-read RNA-seq methods, and that benchmarks 39 analysis tools.
51. Kohli, M. et al. Androgen receptor variant AR-V9 is coexpressed with AR-V7 in prostate cancer metastases and predicts abiraterone resistance. *Clin. Cancer Res.* **23**, 4704–4715 (2017).
52. Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
53. Minoche, A. E. et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* **16**, 184 (2015).
54. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
55. Nottingham, R. M. et al. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA* **22**, 597–613 (2016).
56. Zhao, C., Liu, F. & Pyle, A. M. An ultra-processive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 185–193 (2017).
57. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. HybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
58. Robert, C. & Watson, M. The incredible complexity of RNA splicing. *Genome Biol.* **17**, 265 (2016).
59. Parkhomchuk, D. V. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
60. Levin, J. Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
61. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLOS ONE* **7**, e42882 (2012).
62. Hafner, M. et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3–12 (2008).
63. Chen, Z. & Duan, X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol. Biol.* **733**, 93–103 (2011).
64. Herbert, Z. T. et al. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics* **19**, 199 (2018).
65. Zhao, W. et al. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).
66. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & Von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. *Sci. Rep.* **8**, 4781 (2018).
67. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell. Biol.* **18**, 18–30 (2016).
68. Fullwood, M. J., Wei, C., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19**, 521–532 (2009).
69. Morrissy, A. S. et al. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.* **19**, 1825–1835 (2009).
70. Moll, P., Ante, M., Seitz, A. & Reda, T. Q. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat. Methods* **11**, 972 (2014).
71. Herzog, V. A. et al. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
72. Chen, W. et al. Alternative polyadenylation: methods, findings, and impacts. *Genomics Proteomics Bioinformatics* **15**, 287–300 (2017).
73. Shepard, P. J., Choi, E., Lu, J., Flanagan, L. A. & Hertel, K. J. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).
74. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell* **53**, 1044–1052 (2014).
75. Licatalosi, D. D. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
76. Murata, M. et al. Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).
77. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
78. Islam, S. et al. Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 (2012).
79. The FANTOM Consortium & The RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
80. Adiconis, X. et al. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat. Methods* **15**, 505–511 (2018).
A primary reference for users considering CAGE or similar methods.
81. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
82. Hong, J. & Gresham, D. Incorporation of unique molecular identifiers in TruSeq adapters improves the accuracy of quantitative sequencing. *Biotechniques* **63**, 221–226 (2017).
83. Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531 (2018).
A paper reporting that the majority of RNA-seq duplicates are driven by RNA input rather than sequencing depth and PCR cycles. It also shows that computational removal of duplicates can have unintended consequences on the analysis results.
84. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
A comparison of six scRNA-seq methods that describes the pros and cons of the various approaches and is an excellent introduction to scRNA-seq.
85. Wang, L. et al. Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17**, 58 (2016).
86. Romero, I. G., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42 (2014).
87. Cieslik, M. et al. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).
88. Adiconis, X. et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
A paper covering many of the factors that users with low-quality samples must consider before starting RNA-seq experiments.
89. Schuierer, S. et al. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics* **18**, 442 (2017).
90. Hodges, E. et al. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
91. Sigurgeirsson, B., Emanuelsson, O. & Lundeberg, J. Sequencing degraded RNA addressed by 3′ tag counting. *PLOS ONE* **9**, e91851 (2014).
92. Li, W. et al. Comprehensive evaluation of AmpliSeq transcriptome, a novel targeted whole transcriptome RNA sequencing methodology for global gene expression analysis. *BMC Genomics* **16**, 1069 (2015).
93. Lamarre, S. et al. Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. *Front. Plant Sci.* **9**, 108 (2018).
94. Hansen, K. D., Wu, Z., Irizarry, R. A. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* **29**, 572–573 (2011).
Required reading for anyone considering RNA-seq or other -omics technologies. A well-written reminder of why quantitative RNA experiments will always need replicates, even if RNA assay technologies were perfect. The authors caution users against being over-enthusiastic about new technologies and discarding lessons learned about experimental design.
95. Norton, S. S., Vaquero-Garcia, J., Lahens, N. F., Grant, G. R. & Barash, Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics* **34**, 1488–1497 (2017).
96. Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29**, 656–657 (2013).
97. Wu, Z. & Wu, H. in *Statistical Genomics: Methods and Protocols* (eds Mathé, E. & Davis, S.) 379–390 (Humana Press, 2016).
98. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31**, 233–241 (2015).
99. Gaye, A. Extending the R Library PROPER to enable power calculations for isoform-level analysis with EBSeq. *Front. Genet.* **7**, 225 (2017).
100. Schurch, N. J. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 1641–1641 (2016).
101. Montgomery, S. B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
102. The ENCODE Consortium. Standards, guidelines and best practices for RNA-Seq — V1.0 (June 2011). *UCSC* https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf (2011).
103. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
An overview of computational tools and methods used in RNA-seq analysis.
104. Lei, R., Ye, K., Gu, Z. & Sun, X. Diminishing returns in next-generation sequencing (NGS) transcriptome data. *Gene* **557**, 82–87 (2014).
105. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

106. Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol.* **16**, 131 (2015).
107. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
108. Alamancos, G. P., Agirre, E. & Eyras, E. Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol. Biol.* **1126**, 357–397 (2014).
109. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16**, 59–70 (2013).
110. Williams, C. R., Baccarella, A., Parrish, J. Z. & Kim, C. C. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-seq. *BMC Bioinformatics* **18**, 38 (2017). **A useful overview of several popular computational analysis tools and how they can be used in combination.**
111. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).
112. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
113. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, (15–21 (2013).
114. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
115. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
116. Pertea, M., Kim, D., Pertea, G., Leek, J. T. & Steven, L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2017).
117. Xie, Y. et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
118. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
119. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 4–8 (2016).
120. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
121. Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics* **19**, 510 (2018). **A useful comparison of popular mRNA-seq analysis methods, with particular emphasis on alignment-free tools.**
122. Yang, C., Wu, P.-Y., Tong, L., Phan, J. H. & Wang, M. D. The impact of RNA-seq aligners on gene expression estimation. *ACM BMB* **9**, 462–471 (2015).
123. Robert, C. & Watson, M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* **16**, 177 (2015). **An experimental demonstration of the importance of read mapping and quantification in the computational analysis of mRNA-seq experiments. This paper clearly describes the impact that different alignments and quantification methods can have on biological conclusions.**
124. Zytnicki, M. mmquant: how to count multi-mapping reads? *BMC Bioinformatics* **18**, 411 (2017).
125. McDermaid, A. et al. A new machine learning-based framework for mapping uncertainty analysis in RNA-Seq read alignment and gene expression estimation. *Front. Genet.* **9**, 313 (2018).
126. Fonseca, N. A., Marioni, J. C. & Brazma, A. RNA-Seq gene profiling — a systematic empirical comparison. *PLOS ONE* **9**, e107026 (2014).
127. Teng, M. et al. A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
128. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based versus log-ratio transformation-based methods. *BMC Bioinformatics* **19**, 274 (2018).
129. Vijay, N., Poelstra, J. W., Künstner, A. & Wolf, J. B. W. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.* **22**, 620–634 (2013).
130. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2016).
131. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
132. Turro, E. et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* **12**, R13 (2011).
133. Anders, S., Pyl, P. T. & Huber, W. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
134. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
135. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-seq data. *BMC Bioinformatics* **12**, 480 (2011).
136. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
137. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
138. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* **107**, 9456–9551 (2010).
139. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC* **11**, 94–107 (2010).
140. Dillies, M. A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
141. Li, X. et al. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. *PLOS ONE* **12**, e0176185 (2017).
142. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
143. Robinson, M. D., Mccarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
144. Chen, K. et al. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol. Cell. Biol.* **36**, 662–667 (2016).
145. Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. *Nat. Rev. Genet.* **18**, 473–484 (2017). **A review of the use of spike-in controls and their associated statistical principles. It introduces readers to the concept of commutability: the ability of a spike-in control to perform comparably to experimental RNA samples.**
146. Pine, P. S. et al. Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol.* **16**, 54 (2016).
147. Paul, L. et al. SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. *Preprint at bioRxiv* https://doi.org/10.1101/080747 (2016).
148. Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
149. Lovén, J. et al. Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
150. Risso, D., Ngai, J., Speed, T. & Dudoit, S. in *Statistical Analysis of Next Generation Sequencing Data* (eds Datta, S. & Nettleton, D.) 169–190 (Springer, 2014).
151. Qing, T., Yu, Y., Du, T. T. & Shi, L. M. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Sci. China Life Sci.* **56**, 134–142 (2013).
152. Leshkowitz, D. et al. Using synthetic mouse spike-in transcripts to evaluate RNA-seq analysis tools. *PLOS ONE* **11**, e0153782 (2016).
153. Lun, A. T. L., Calero-nieto, F. J., Haim-vilmovsky, L., Göttgens, B. & Marioni, J. C. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* **27**, 1795–1806 (2017).
154. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
155. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
156. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
157. Frazee, A. et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* **33**, 243–246 (2015).
158. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
159. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
160. Asp, M. et al. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci. Rep.* **7**, 12941 (2017).
161. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
162. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015). **This review provides an overview and in-depth discussion of scRNA-seq transcript quantitation methods. The authors highlight the analytical challenges that are unique to single-cell experiments.**
163. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018). **This review is an excellent introduction to the full range of single-cell sequencing methods.**
164. Leelatian, N. et al. Single cell analysis of human tissues and solid tumors with mass cytometry. *Cytometry B* **92**, 68–78 (2018). **A useful description of the pitfalls of tissue dissociation for users of single-cell sequencing to consider.**
165. Hines, W. C., Su, Y., Kuhn, I., Polyak, K. & Bissell, M. J. Sorting out the FACS: a devil in the details. *Cell Rep.* **6**, 779–781 (2014).
166. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
167. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1098 (2013).
168. Goldstein, L. D. et al. Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18**, 519 (2017).
169. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
170. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
171. Cao, J. et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *Science* **357**, 661–667 (2017).
172. Rosenberg, A. B. et al. Single cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
173. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
174. Sena, J. A. et al. Unique molecular identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci. Rep.* **8**, 13121 (2018).
175. Dal Molin, A. & Di Camillo, B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief. Bioinform.* https://doi.org/10.1093/bib/bby007 (2018).
176. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
177. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
178. 10x Genomics. Application note. Chromium™ — transcriptional profiling of 1.3 million brain cells with the Chromium single cell 3′ solution. *10x Genomics* http://go.10xgenomics.com/l/172142/2017-06-09/bsylz/172142/31729/LIT000015_Chromium_Million_Brain_Cells_Application_Note_Digital_RevA.pdf (2018).

179. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).

180. Insel, T. R., Landis, S. C. & Collins, F. S. The NIH BRAIN initiative. 340, 687–689 (2013).

181. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **599**, 594–599 (2018).

182. Hui Ryu, K., Huang, L., Min Kang, H. & Schiefelbein, J. Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol.* **179**, 1444–1456 (2019).

183. Chen, J. et al. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat. Protoc.* **12**, 566–580 (2017).

184. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

185. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **1467**, 1463–1467 (2019).

186. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.* **16**, 57–66 (2015).

187. Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
This review of spatial RNA-seq methods introduces the main methods in more detail and discusses some of the technical challenges that remain to be resolved.

188. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).

189. Datta, S. et al. Laser capture microdissection: big data from small samples. *Histol. Histopathol.* **30**, 1255–1269 (2015).

190. Lovatt, D., Bell, T. & Eberwine, J. Single-neuron isolation for RNA analysis using pipette capture and laser capture microdissection. *Cold Spring Harb. Protoc.* **2015**, 60–68 (2015).

191. Cubi, R. et al. Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of Plasmodium relapsing species. *Cell. Microbiol.* **19**, e12735 (2017).

192. Giacomello, S. et al. Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* **3**, 17061 (2017).

193. Moncada, R. et al. Integrating single-cell RNA-Seq with spatial transcriptomics in pancreatic ductal adenocarcinoma using multimodal intersection analysis. Preprint at *bioRxiv* https://doi.org/10.1101/254375 (2018).

194. Ke, R. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).

195. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

196. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).

197. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).

198. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).

199. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.* **8**, 4847 (2018).

200. Pichon, X., Lagha, M., Mueller, F. & Bertrand, E. A. Growing toolbox to image gene expression in single cells: sensitive approaches for demanding challenges. *Mol. Cell* **71**, 468–480 (2018).

201. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **93**, 89–93 (2019).

202. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

203. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).

204. Core, L. J., Waterfall, J. & Lis, J. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

205. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**, 1791–1792 (2008).

206. Skalska, L., Beltran-nebot, M., Ule, J. & Jenner, R. G. Regulatory feedback from nascent RNA to chromatin and transcription. *Nat. Rev. Mol. Cell. Biol.* **18**, 331–337 (2017).

207. Tani, H. et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).

208. Paulsen, M. T. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc. Natl Acad. Sci. USA* **110**, 2240–2245 (2013).

209. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).

210. Nojima, T., Gomes, T., Carmo-fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat. Protoc.* **11**, 413–428 (2016).

211. Nagari, A., Murakami, S., Malladi, V. S. & Kraus, W. L. Computational approaches for mining GRO-Seq data to identify and characterize active enhancers. *Methods Mol. Biol.* **1468**, 121–138 (2017).

212. Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T. & Meyer, B. J. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *eLife* **18**, e00808 (2013).

213. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).

214. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).

215. Nojima, T. et al. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).

216. Wallace, E. W. J. & Beggs, J. D. Extremely fast and incredibly close: cotranscriptional splicing in budding yeast. *RNA* **23**, 601–610 (2017).

217. Rabani, M. et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).

218. Schwalb, B. et al. TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).

219. Marzi, M. J. & Nicassio, F. Uncovering the stability of mature miRNAs by 4-thio-uridine metabolic labeling. *Methods Mol. Biol.* **1823**, 141–152 (2018).

220. Riml, C. et al. Osmium-mediated transformation of 4-thiouridine to cytidine as key to study RNA dynamics by sequencing. *Angew. Chem. Int. Ed.* **56**, 13479–13483 (2017).

221. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* **15**, 221–225 (2018).

222. Muhar, M. et al. SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science* **360**, 800–805 (2018).

223. Matsushima, W. et al. SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development* **145**, dev164640 (2018).

224. Jürges, C., Dölken, L. & Erhard, F. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics* **34**, 218–226 (2018).

225. Shah, S. et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* **174**, 363–376 (2018).

226. Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. & Sarnow, P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Natl Acad. Sci. USA* **96**, 13118–13123 (1999).

227. Yamashita, R. et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.* **21**, 775–789 (2011).

228. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).

229. Wang, E. T. et al. Dysregulation of mRNA localization and translation in genetic disease. *J. Neurosci.* **36**, 11418–11426 (2016).

230. Parker, M. W. et al. Fibrotic extracellular matrix activates a profibrotic positive feedback loop. *J. Clin. Invest.* **124**, 1622–1635 (2014).

231. Moreno, J. A. et al. Sustained translational repression by eIF2a–P mediates prion neurodegeneration. *Nature* **485**, 507–511 (2012).

232. Bhat, M. et al. Targeting the translation machinery in cancer. *Nat. Rev. Drug Discov.* **14**, 261–278 (2015).

233. Leibovitch, M. & Topisirovic, I. Dysregulation of mRNA translation and energy metabolism in cancer. *Adv. Biol. Regul.* **67**, 30–39 (2018).

234. Liang, S. et al. Polysome-profiling in small tissue samples. *Nucleic Acids Res.* **46**, e3 (2017).

235. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

236. Floor, S. N., Doudna, J. A., States, U. & Initiative, I. G. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5**, e10921 (2016).

237. Blair, J. et al. Widespread translational remodeling during human neuronal differentiation. *Cell Rep.* **21**, 2005–2016 (2017).

238. Steitz, J. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* **224**, 957–964 (1969).

239. Hsu, P. Y. et al. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **113**, E7126–E7135 (2016).

240. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).

241. Calviello, L. & Ohler, U. Beyond read-counts: ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.* **33**, 728–744 (2017).

242. Erhard, F. et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* **15**, 363–366 (2018).

243. Li, W., Wang, W., Uren, P. J., Penalva, L. O. F. & Smith, A. D. Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics* **33**, 1735–1737 (2017).

244. Zhong, Y. et al. RiboDiff: Detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* **33**, 139–141 (2017).

245. Paulet, D., David, A. & Rivals, E. Ribo-seq enlightens codon usage bias. *DNA Res.* **24**, 303–310 (2017).

246. Gao, X. et al. Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12**, 147–153 (2015).

247. Archer, S. K., Shirokikh, N. E., Beilharz, T. H. & Preiss, T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature* **535**, 570–574 (2016).

248. Iwasaki, S. & Ingolia, N. T. The growing toolbox for protein synthesis studies. *Trends Biochem. Sci.* **42**, 612–624 (2017).

249. Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.* **40**, 221–232 (2015).

250. Holley, R. W. et al. Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).

251. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223–4231 (2005).

252. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nat. Rev. Genet.* **19**, 615–634 (2018).
A good introduction to RNA structural analysis using RNA-seq.

253. Kertesz, M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).

254. Underwood, J. G. et al. FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7**, 995–1001 (2010).

255. Lucks, J. B. et al. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl Acad. Sci. USA* **108**, 11063–11068 (2011).

256. Ding, Y. et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).

257. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).

258. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).

259. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).

260. Incarnato, D. et al. In vivo probing of nascent RNA structures reveals principles of cotranscriptional folding. *Nucleic Acids Res.* **45**, 9716–9725 (2017).

261. Novoa, E. M., Beaudoin, J., Giraldez, A. J., Mattick, J. S. & Kellis, M. Best practices for genome-wide RNA structure analysis: combination of mutational profiles and drop-off information. Preprint at *bioRxiv* https://doi.org/10.1101/176883 (2017).
262. Lee, B. et al. Comparison of SHAPE reagents for mapping RNA structures inside living cells. *RNA* **23**, 169–174 (2017).
263. Tang, Y., Assmann, S. M. & Bevilacqua, P. C. Protein structure is related to RNA structural reactivity in vivo. *J. Mol. Biol.* **428**, 758–766 (2016).
264. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
265. Warner, K. D., Hajdin, C. E. & Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* **17**, 547–558 (2018).
266. Kudla, G., Granneman, S., Hahn, D., Beggs, J. D. & Tollervey, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc. Natl Acad. Sci. USA* **108**, 10010–10015 (2011).
267. Kretz, M. et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**, 231–235 (2013).
268. Engreitz, J. M. et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
269. Lu, Z. et al. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267–1279 (2016).
270. Aw, J. G. et al. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Mol. Cell* **62**, 603–617 (2016).
271. Sharma, E. et al. Global mapping of human RNA-RNA interactions. *Mol. Cell* **62**, 618–626 (2016).
272. Gong, J. et al. RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res.* **46**, 194–201 (2018).
273. Zhang, X. et al. RAID: a comprehensive resource for human RNA-associated (RNA–RNA/RNA–protein) interaction. *RNA* **20**, 989–993 (2014).
274. Schönberger, B., Schaal, C., Schäfer, R. & Voß, B. RNA interactomics: recent advances and remaining challenges. *F1000Res.* **7**, 1824 (2018).
275. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
276. Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl Acad. Sci. USA* **97**, 14085–14090 (2000).
277. Zhao, J. et al. Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010).
278. Mili, S. & Steitz, J. Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses. *RNA* **10**, 1692–1694 (2004).
279. Niranjanakumari, S., Lasda, E. & Brazas, R. Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions in vivo. *Methods* **26**, 182–190 (2002).
280. Hendrickson, G., Kelley, D., Tenen, D., Bernstein, D. & Rinn, J. Widespread RNA binding by chromatin-associated proteins. *Genome Biol.* **17**, 28 (2016).
281. Ule, J. et al. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212–1215 (2003).
282. König, J. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
283. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
284. Garzia, A., Meyer, C., Morozov, P., Sajek, M. & Tuschl, T. Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods* **118**, 24–40 (2017).
285. Zarnegar, B. J. et al. IrCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods* **13**, 489–492 (2016).
286. Van Nostrand, E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
287. Nostrand, E. L. Van et al. A large-scale binding and functional map of human RNA binding proteins. Preprint at *bioRxiv* https://doi.org/10.1101/179648 (2017).
288. Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M. & Ule, J. Data science issues in studying protein–RNA interactions with CLIP technologies. *Annu. Rev.* **1**, 235–261 (2018).
289. Lee, F. C. Y. & Ule, J. Advances in CLIP technologies for studies of protein-RNA interactions. *Mol. Cell* **69**, 354–369 (2018).
A review of RNA–protein interaction methods, with a 5-page table describing the methodological advances of each. Vital reading for anyone considering CLIP–seq analysis.
290. Buenrostro, J. D. et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
291. Cook, K. B., Hughes, T. R. & Morris, Q. D. High-throughput characterization of protein-RNA interactions. *Brief. Funct. Genomics* **14**, 74–89 (2015).
292. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
293. Doebele, R. C. et al. An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor. *Cancer Discov.* **5**, 1049–1057 (2015).
294. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).

Acknowledgements

We thank J. Marioni and J. Ule for their valuable comments on the manuscript.

Author contributions

J.H., R.S. and M.G. researched the literature. J.H. and R.S. discussed the content and wrote and edited the article.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Genetics thanks T. Preiss, J. Ragoussis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s41576-019-0150-2.

RELATED LINKS

nanopolish-polya: https://github.com/jts/nanopolish