Making heads or tails of the untranslated parts of genes: computational methods applied to high-throughput transcriptomic data

Krzysztof J. Szkop[1] and Irene Nobeli[1,$]

1. Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, Malet Street, London WC1E 7HX, UK.
$ Corresponding author email: i.nobeli@bbk.ac.uk

Subtitle

Computational methods to discover and quantify isoforms with alternative untranslated regions

Keywords

Alternative poly-adenylation; alternative transcription start site; untranslated region; RNA-seq

Abbreviations

APA, alternative poly-adenylation; **PCS**, poly-adenylation and cleavage site; **RNA-seq**, RNA sequencing; **TSS**, transcription start site; **UTR**, untranslated region.

Summary

The fate of eukaryotic transcripts is closely linked to their untranslated regions, which are determined by where transcription starts and ends on a genomic locus. The extent of alternative transcription start and alternative poly-adenylation has been revealed by sequencing methods focused on the ends of transcripts, but the application of these methods is not yet widely adopted by the community. In this review we highlight the importance of defining the untranslated parts of transcripts

and suggest that computational methods applied to standard high-throughput technologies are a useful alternative to the expertise-demanding 5' and 3' sequencing. We present a number of computational approaches for the discovery and quantification of alternative transcription start and poly-adenylation events, focusing on technical challenges and arguing for the need to include better normalization of the data and more appropriate statistical models of the expected variation in the signal.

Introduction

The development of high-throughput technologies for transcriptome profiling has allowed us unprecedented access to the ensemble of RNA transcripts in and out of the cell. Evidence from a growing number of studies suggests that both the extent of the RNA repertoire and its complexity have been vastly underestimated[1]. Nearly a decade after the introduction of next-generation sequencing technologies for surveying the transcriptome, it is still possible to find novel splicing junctions that were originally overlooked because of low expression of the corresponding transcripts[2]. Although the term "gene" [3] has served geneticists well as the unit of heredity, a complete understanding of the workings of a cell at the molecular level dictates a shift from a gene-level to a transcript (or "isoform")-level analysis.

A transcript-centric approach to quantifying RNA is, however, challenging. Alternative splicing of the primary RNA is the most commonly studied mechanism of producing gene variants in eukaryotes, but the possibility of starting and ending a transcript at different points in the genome through the use of alternative transcription start sites (TSS) and alternative poly-adenylation and cleavage sites (PCS) adds another layer of complexity to the study of the transcriptome. Interestingly, the choice of TSS and PCS are not disconnected[4], adding to the intricacies of regulation mediated by the transcript's untranslated regions (UTRs). Although alternative TSS/PCS can result in altered protein-coding sequences, they often produce isoforms that vary only in their non-coding

parts (Fig. 1). Such differences in the untranslated regions have been shown to affect the stability, localization and translation efficiency of the parent transcripts and, more recently, the localization of proteins post-translationally, as documented by a number of studies (a selection of examples is presented in Box 1 and reviews [5-8] provide a more comprehensive coverage of the subject).

Considering the importance of alternative TSS and alternative PCS, it is instructive to understand why the vast majority of transcriptomic studies employing microarrays or RNA-seq technologies do not attempt to quantify the relative expression of isoforms that differ only in their untranslated regions. We believe a major issue is lack of comprehensive annotation. Transcription start and end points differ between different tissues [9, 10], developmental stages [11, 12], and states of health[13, 14](with many more examples presented in the the review by Curinha et al.[15]). An accurate analysis of transcriptomic data would require a comprehensively annotated version of the reference genome, something that is arguably lacking even for the best-studied organisms. Existing databases are often focused on humans, but even then agreement between them is limited. For example, our comparison of entries in APADB[16] and PolyAsite[17] databases shows that of the 392912 PCS clusters in PolyAsite, 60186 overlap with at least one of the 71829 clusters in APADB. This means that 16% of the PCS in the smaller database have no match in the larger one (these numbers have been obtained using coordinates rather than gene names which would result in a lot fewer hits; overlap is defined very generously as at least one nucleotide in common between clusters, with a tolerance of 5 nucleotides on either side). As with many bioinformatics resources, there is an additional issue with keeping such databases up to date when new data are published or when new reference genome sequences are released. For example, none of the major alternative poly-adenylation databases uses the current human genome release (hg38), yet this is already the standard for most sequence analysis pipelines. Finally, there is an argument that aberrant poly-adenylation might not affect only the relative quantities of transcribed isoforms but also the actual sites used. In other words, even a complete reference set of poly-adenylation or transcription start sites may not be appropriate for samples originating from

conditions that allow cryptic sites to become sufficiently utilized resulting in novel transcription products.

The lack of comprehensive annotation is a barrier to analyzing transcriptomic data but could be overcome by methods that deduce the transcript structure directly from the data. Although tempting, this approach presents several challenges. Reconstructing the transcriptome from a pool of sequenced transcript fragments and quantifying the relative expression of isoforms is only straightforward in the simplest scenario of one gene producing a single transcript (see relevant reviews by [18-21] for a more thorough coverage of the challenges associated with transcriptome reconstruction and isoform quantification). In addition to the problems shared with methods that attempt to quantify alternative splicing events, methods for probing the ends of transcripts are plagued by technical issues linked to biases in next-generation sequencing technologies, as discussed below. Sequencing of whole transcripts, as promised by the more recent third-generation sequencing technologies (nanopores and the single molecule real-time technology (SMRT) platforms), would potentially eliminate these issues. Early successes in this direction [22-25] are very promising, but there is disagreement in the literature about the error rate associated with nanopore sequencing[26-28]. Clearly, these technologies are still under active development and many of the technical challenges, such as variability of the time intervals used to identify each base, are unresolved [29]. The challenges associated with sequencing cDNA with nanopores are also hampering direct RNA sequencing with this technology, and in addition, there is less evidence in this case that coverage of the transcriptome is to a satisfactory degree. One recent non-peer reviewed study[30] suggests promising results but also makes it clear that the timeframe required for nanopores to deliver on their promises remains uncertain.

Until sequencing of full transcripts becomes routine, it appears that there are two routes to a comprehensive study of the transcripts' untranslated parts: one is to carry out additional experiments that are specifically aimed at probing the ends of the transcript; the other is to use computational methods to obtain information about alternative TSS and PCS from standard transcriptomic data produced by microarrays and RNA-seq. The present review focuses on the

second option, with a brief description of experimental methods aimed at capturing the transcript ends of a gene given in Box 2. The latter methods are obviously necessary for benchmarking computational approaches. In summary, there are a variety of methods that have been developed for probing both TSS and PCS at nucleotide resolution. Many have been used already to provide us with great insights on the differential use of transcript ends in different tissues, stages of development or disease. However, we believe that the availability of these specialized methods to the wider research community remains a challenge. The majority of these methods are still only accessible to research groups that have the necessary expertise, funds and laboratory set ups to carry out the often complex and laborious protocols involved. Until these targeted methods become more widely accessible, bioinformatics approaches will constitute a much-needed alternative solution. In this spirit, we review below computational approaches for discovering and quantifying the use of alternative TSS and PCS from standard transcriptomic data.

Computational prediction of alternative TSS and PCS from standard transcriptomic data

The availability of genome-wide surveys of transcripts has opened the doors to deducing TSS and PCS data for all genes expressed in a given sample using computational methods. Although it may be possible to computationally infer potential starting and ending points of transcription using genomic data alone (see for example studies by [31-45] among many others), methods to do so have to rely on sequence signals/motifs or their associated structure and thermodynamic properties, that can, at best, suggest the possibility of a TSS or PCS in a genome but cannot ascertain their use at any given time. The topic of finding such signals is outside the scope of this review but interested readers are directed to the review by Tian and Graber[46]; instead we are focusing here on the prediction of TSS and PCS from transcriptomic data.

Some of the earliest estimates of the use of alternative TSS/PCS have their origin in studies that were focused on alternative splicing and, in the absence of microarray and next-generation sequencing data, relied on Expressed Sequence Tag (EST) contigs and pioneering spliced aligners to identify splicing events. Although these studies did not explicitly set out to identify alternative TSS and PCS, they did discover large numbers of alternative splicing events in the untranslated regions ([47-49]) and often gave information on the predicted starts and ends of transcripts based on ESTs. It was the advent of high-throughput transcriptomic technologies though that fuelled an interest in exploring transcripts in detail, including the diversity of untranslated regions. Below, we concentrate on bioinformatics methods developed for and applied to the more recent transcriptomic data from microarray and RNA-seq experiments.

Quantifying the use of alternative TSS and PCS from microarray data is limited by probe design and signal variability

Microarray chips for surveying the transcriptome predate RNA-seq and in theory, if designed appropriately, could be used for accurate prediction of the relative use of known TSS and PCS sites in any gene. In microarrays, RNA expression is measured through the amount of cDNA that hybridizes to pre-designed short DNA fragments (probes) immobilized on a chip. Analysis of expression at the gene level requires only that some probes cover parts of each transcript, but the analysis of untranslated regions and relative quantification of alternative 5' and 3' ends dictates in addition the requirement for a satisfactory coverage of the untranslated parts. It is primarily this requirement that renders most microarray chips unusable for discovering alternative transcript ends, or at best, limits their use to a small subset of genes. Illumina and Agilent platforms, for example, used relatively longer (50- to 60-mer) reads but with their numbers of probes per chip ranging between 40,000-50,000, it is obvious that only a small subset of genes (if any at all) could be covered adequately for UTR probing. Thus, we focus here on chips produced by Affymetrix, which are arguably the most widely used in quantification of gene expression. The spread of probes was very limited in the first two generations of Affymetrix platforms. GeneChip arrays lacked probes in the 5' UTR, and both they and Exon ST arrays had insufficient

probes across the 3'UTR to satisfactorily cover alternative-length transcript tails (Fig. 2A). As a consequence, there have been only few attempts to quantify alternative poly-adenylation site selection using these arrays and, to our knowledge, the only study that proposed a method for mapping both ends of a transcript from microarray data utilized a high-density tiling array covering the whole genomic sequence of *Saccharomyces cerevisiae*[50].

All methods developed to find alternative PCS using microarray data rely on the fact that microarray chips use "probe sets" rather than individual probes to quantify the expression of a gene. Most microarray analysis protocols group probes together, if they correspond to sequences derived from the same gene/transcript/exon, thus summarizing probe values to the appropriate feature level. However, if the individual probes cover both the part of the transcript before and after a poly-adenylation site, then it is possible to extract information about the relative use of the "short" and "long" tailed transcripts in different conditions by calculating the ratio of expression for the probes up- and down-stream of that site. When applied to 3'UTR events, these methods benefit from the fact that UTRs often lack introns (less than ~10% of all annotated introns are located outside the protein-coding region[51]) and hence alternative splicing is less likely to interfere with the APA signal. Similarly, limiting the search to the last exon or employing microarrays with probes heavily biased towards the 3'UTR further focuses the method on events due to alternative poly-adenylation. Although theoretically straightforward and computationally easy to implement, comparing levels of individual probes is hindered by the variability in probe signal whose origins are technical rather than biological. Our recent experience working with several published datasets agrees with earlier studies [52, 53] suggesting that the expression levels of individual probes vary widely due to biases such as the probe location within the transcript or the similarity of the sequence probe to sequences originating from other non-target genes.

Despite the non-trivial challenges involved, several methods have been published attempting to quantify APA events from microarray data. In a pioneering study, Sandberg *et al.*[54] introduced the probe-level alternative transcript analysis (PLATA), a method that uses the individual microarray probes to assess differences in expression within a gene, after correcting for

probe-specific variations and normalizing for gene-level intensities. This approach relies on prior knowledge of the PCS (Sandberg *et al* used EST-supported poly(A) sites), as does the more recent APAdetect method[55] which relies on poly(A) sites from the PolyA_DB database[56]. In contrast, the *Rmodel* method[57], implemented as a package in R, allows identification of novel events by comparing individual probe expression ratios in two conditions along the gene body and identifying the optimum segmentation point using a modified *t*-test. More recently, Li et al[58] replaced the modified *t*-test in *Rmodel* with a Bayesian analysis following the method of Erdman & Emerson[59] implemented in the R package *bcp*[60]. In their approach, a list of tandem 3'UTRs is first constructed from the coordinates of known transcripts with identical 3' UTR start sites but different poly-adenylation sites. Then, the change point is identified as the probe with the highest posterior change probability. Finally, the fold change between the expression levels of the common and extended regions is calculated, filtering out insignificant or unreliably measured changes.

Although the methods described above are promising, their performance is highly dependent on the microarray design, as well as the quality of the signal at the individual probe level. To improve performance, methods usually employ some form of filtering to exclude "outlier" probes whose intensities lie on the extremes of the distribution defined by the intensities of probes belonging to a given set. However, in the case of a small number of probes (<10), a distribution would be difficult to define with any confidence, making the identification of outliers a practice of debatable value. Additionally, "size" filters can be applied to exclude transcripts that do not have enough probes covering the area of interest. In practice, however, these filters may be set at unrealistically small cut-offs (APAdetect[55], for example, eliminates transcripts with a single distal or proximal probe) meaning that as few as two probes may be considered acceptable, although it is clear that the statistical value of a comparison of intensities between control and condition samples in this case would be questionable.

Most existing microarrays were not designed with complete coverage of potential untranslated regions in mind and would not be suitable for discovering new TSS/PCS or even quantifying accurately differential use of known sites in

samples originating from distinct conditions. The recent introduction of the Affymetrix Human Transcriptome Array 2.0 is a promising step in the right direction with increased coverage of both the 5' and 3' UTR of a substantial part of protein-coding isoforms. Although, the platform remains largely unexplored in this aspect, it appears that it has the potential to be used for this type of analysis. However, it should be kept in mind that even probes in this array cover on average ~70% of the 5' human UTRs but less than 40% of the 3' and (Fig. 2B), limiting its applicability to a subset of annotated genes.

RNA-seq can reveal alternative TSS and PCS events but accounting for technical biases is not trivial

RNA-seq technology transformed transcriptome profiling providing information on the full length of both known and novel transcripts, including the much under-studied untranslated regions. In RNA-seq data, microarray-style predefined probes are replaced by "reads", fragmented pieces of RNA ranging in length from 25 to few hundreds of nucleotides depending on the protocol and platform. These reads potentially cover the entire length of each RNA in a sequenced sample, making it possible, in theory at least, for any region of an RNA transcript that is sufficiently expressed to be detected by the method. In reality, most RNA-seq protocols suffer from relatively poor coverage of the 5′ ends of RNAs, and poor representation or biased over-representation of the 3' end, depending on the technology used[61]. For example, fragmentation bias can lead to either strong depletion of reads at the 5'end (when cDNA is fragmented) or milder depletion of reads at both ends (when RNA is fragmented)[61]. Given the difficulty of obtaining good quality data for the 5' end, it is not surprising that hardly any methods have been developed specifically for identification of TSS from RNA-seq data. RNA degradation, a common problem during the storage and preparation of samples, plagues RNA-seq libraries and as a result, disappointing levels of RNA integrity are not uncommon among transcripts in publicly available RNA-seq datasets. In RNA-seq libraries prepared using the polyA+ selection method, the 3' ends of transcripts are protected from degradation, leading to a higher number of reads at the end of the transcript, and a corresponding density peak close to the end of degraded transcripts. In libraries

derived from ribosome depletion, the 3' end is not protected and read coverage is generally lower at both 5' and 3' ends[62]. Despite these challenges, the reasonable coverage of 3' ends by RNA-seq reads in combination with the promise of potential discovery of novel sites, has prompted the development of a variety of methods for the analysis of alternative PCS using next-generation sequencing data. It should be obvious here that for all computational methods that rely on changes in the expression signal, rather than sequence or structure motifs, the two ends of the transcript are indistinguishable; a method defined for one end would generally be applicable to both ends, with only minor modifications.

Current computational tools fall into either of two major categories: Those that rely on existing annotations of TSS/PCS and aim only to quantify the different-length isoforms and those able to predict the position of the transcript ends from the distribution of RNA-seq reads. Tools from the first category are clearly more limited in their applicability as databases are currently focused on a limited number of tissues and organisms, and as highlighted earlier, it is debatable that any single one used in isolation comprehensively covers all potential alternative events (Table 1 summarises web-accessible databases with information on alternative PCS and TSS).

Despite the challenges of using existing annotation to support the prediction of alternative PCS or quantify their use, a number of methods follow this approach. The mixture-of-isoforms model, or MISO[63], was developed to estimate the expression of alternatively spliced exons or isoforms but can be used also to estimate expression of isoforms resulting from APA events. For MISO to infer the abundance of isoforms using Bayes' rule (the 'percent spliced in' or PSI value[64]), it requires knowledge of which isoforms a read is compatible with and hence knowledge of the splice junctions (or in the case of poly-adenylation, the PCS sites) in order to build the isoform compatibility matrix. The Bioconductor package *roar* (=ratio of a ratio) (Grassi 2013) requires similarly an annotation file with the coordinates of the canonical and alternative poly-adenylation site for each gene. Although *roar* allows only two sites per gene, the package can handle multiple pairwise comparisons between the canonical end and different alternative PCS. This is used to assign each read to either the

portion of the gene that belongs to the long isoform only ('POST'), or the portion that is common to the short and long isoforms ('PRE'). The ratio of short to long isoform expression is then estimated, taking into account the length of each isoform. The statistical significance of the difference between PRE and POST counts in two samples (e.g. treatment versus control) is assessed using a Fisher test. This approach, like others that attempt to estimate the actual isoform expression, assumes that reads are uniformly distributed along the length of the transcript, an assumption that is unlikely to be true in any real RNA-seq dataset.

Methods that rely on transcript reconstruction are computationally expensive and must solve the problem of assigning reads to overlapping isoforms

The second category of methods do not rely on an existing annotation of PCS but instead try to deduce the use of alternative poly-adenylation sites directly from the data. Some of these tools rely on transcript reconstruction methods, such as Cufflinks[65], to provide them with a list of transcripts that are consistent with the available reads in a sample. Thus, they delegate the difficult problem of putting transcripts together to external software, whereas they deal with the relatively easier question of identifying 3'UTR lengths that do not match the genome annotation, and then comparing quantitatively the use of alternative PCS across samples. The 3'UTR Sequence Seeker (3USS) server[66] is an example of a method that relies on the output from genome-guided transcript reconstruction carried out by Cufflinks. The 3'UTR lengths of transcripts sharing a common chain of introns are compared to the corresponding annotated transcripts to discover alternative PCS. This way, both previously known and novel alternative PCS can be identified in the data. Although the 3USS server can highlight cases where an alternative PCS is exclusively used in one sample, it cannot quantify differential use of alternative PCS in different samples. Given that in the majority of cases the selection of a PCS is not a binary event (i.e. not simply on or off), methods attempting to quantify the relative use of two or more PCS across samples are potentially more useful. An additional issue with tools that employ the output of transcript reconstruction methods is that they cannot distinguish between short and long 3'UTR isoforms, if these isoforms contain the

same exons. In these cases, the short isoform is embedded in the long and reads originating from the part of the transcript common to both isoforms cannot be assigned to short or long in a straightforward manner. Hence, methods that rely on reconstruction of transcripts based on their exon composition will almost certainly fail to identify alternative PCS or correctly deduce their relative use across samples.

In order to overcome the shortcomings of incomplete genome annotations, some tools employ *de novo* transcriptome reconstruction prior to searching for alternative 3'UTRs. Both KLEAT[67] and PASA[68] implement pipelines that rely on such *de novo* reconstruction, a process notorious for its demands on computational power (it should be noted that PASA can be used in the context of APA analysis but was developed with the aim of automatically modeling gene structures using spliced alignments and hence, its scope extends far beyond APA discovery and quantification). However, even assuming that *de novo* reconstructions will become more routinely accessible with time, their inability to reconstruct lowly expressed transcripts fully and correctly is likely to remain an issue. Perhaps more problematic than either accuracy or complexity in this case is the fact that when it comes to deciphering poly-adenylation sites, these tools seem to rely heavily, if not entirely, on the presence of poly(A)-capped reads in the data, i.e. reads that are produced from the end of the transcript and have thus stretches of untemplated As that cannot be mapped to the genome. The idea of using such reads has been around since the early days of RNA-seq[69] and, in combination with the use of a library enriched in poly(A)-spanning reads, it allowed the discovery of a great number of unannotated poly-adenylation events in *Drosophila melanogaster*[70]. More recently, it has been suggested that mapping of poly(A)-spanning reads should become standard practice which led to the incorporation of this method into the Context Map 2 RNA-seq mapping pipeline[71]. Although these studies are evidence of the increasing popularity of this approach, the issue with relying on poly(A)-containing reads is that they are actually rather scarce in standard RNA-seq datasets (where libraries are not specifically enriched in poly(A)-carrying reads), their number being largely dictated by read coverage of the 3' end. Our analysis of the relatively recent dataset by Bayerlová et al.[72] showed that only ~0.1% of

reads have at least 6 As at the end, a percentage consistent with that obtained for other datasets we have examined in the past. A more thorough analysis by Kim et al.[73] found just over 10% of 130 million unmapped reads had at least two untemplated As in their 3'end, and following removal of these As only ~0.1% of the original reads could be uniquely mapped to the genome. The reads that mapped successfully provided poly-adenylation cleavage site information for just ~2000 protein-coding genes, suggesting that reads with untemplated As are of limited use in the analysis of APA events. In our opinion, the fact that fairly complex *de novo* reconstruction protocols are actually limited by the relatively trivial step of identifying poly(A) stretches at the end of reads makes them less attractive for analyzing the 3' end of transcripts. In addition, this approach is clearly not suitable for the 5' end, where no equivalent sequence signals the beginning of the transcript.

Detection of read density fluctuations in RNA-seq data allows the discovery of novel events but is prone to high false positive rates

A final group of algorithms bases PCS recognition on read density fluctuations. The number of mapped RNA-seq reads at each position along the UTR is considered and the algorithms generally search for a sudden fluctuation in the number of reads, indicative of a transcript start or termination event that is embedded in the genomic sequence within a longer transcript (Fig. 3). A method that delineates both the 5' and 3' ends of transcripts using read density fluctuation was included in the very first paper describing RNA-seq[74] although its description was limited to the detection of a "sharp" reduction in the level of transcription with no further details. Since then, several studies have adopted this approach employing a variety of computational algorithms. It is important to highlight that all methods relying on read density fluctuations are prone to inaccurate predictions caused by variations whose origin is not biological (see, for example, [75, 76] and, at least in principle, need to normalize or smooth the data to avoid such technical biases. Kim et al's GETUTR method[73] applies a choice of three algorithms to smooth the RNA-seq signal prior to finding the local maximum-gradients in the new monotonically decreasing signal landscape that are thought to coincide with the poly-adenylation cleavage sites. Due to the low

computational complexity of the algorithm, the method is applicable to processing very large datasets but the software seems to be limited to finding the cleavage sites, without attempting a quantification of relative expression of the corresponding isoforms. Lu and Bushell's PHMM method[77] employs Hidden Markov Models (HMMs) to treat alternative poly-adenylation events as transitions between two distinct (and hidden) states, a "high expression" state corresponding to the part of the UTR that is contained in both short and long isoforms and a "low expression" state corresponding to part covered only by the longer isoform. Transcripts for which a two-state model has a better fit than a one-state model are selected and the method looks for transitions in the Markov chain from high to low expression states in the sequences. A sliding window of 100 to 800 bases (depending on the length of the 3' UTR) is used to smooth the data and reads are counted against these overlapping windows. Read counts at any given window are dependent on the state generating them and are modeled using a Poisson distribution. The Poisson assumption is problematic with data that is very likely to be over-dispersed (as is the case with RNA-seq data), and this may be one of the reasons why the estimated specificity of the method is a lot lower than its sensitivity. Moreover, issues of heterogeneity of the read density in RNA-seq may be contributing to the high number of false positives, and there are additional issues that have not been addressed, such as the problem of taking into account multiple samples per condition. ChangePoint[78] is another method that attempts to discover the points that mark the transition in read density expected at an internal poly-adenylation site. It uses a generalized likelihood ratio statistic to evaluate the significance of these transitions and, additionally, it applies a directional multiple testing procedure for controlling the mixed directional False Discovery Rate (FDR). In essence, the latter ensures that only the most significant events are reported, because in a multiple testing scenario such as the one here (where thousands of genes are routinely considered), it is these large magnitude events that are more likely to be relevant to the conditions examined. A drawback of ChangePoint is that it cannot analyse more than one sample per condition. The Isoform Structural Change Model (IsoSCM)[79] approach is built on the same principle of looking for change-points in the read density, but it is based on a Bayesian model and

allows for multiple change-points to be discovered. Importantly, the discovery of the sites by IsoSCM is entirely done *ab* initio and does not rely on the presence of a reference annotation, which limits the applicability of a method to model, well-studied organisms. However, the current implementation of IsoSCM can also only be applied to pairwise comparisons of samples, making it difficult to analyse data with biological replicates, unless samples from the same condition are pooled together. DaPars [80], in contrast allows the analysis of an arbitrarily large number of samples. It works by first predicting the proximal PCS by minimizing the difference between observed and estimated read density, based on a two-site model. Although the original publication by Xia et al. suggests both a way to infer the distal PCS from RNA-seq data and a generalization to solving the problem when multiple sites are present, the publicly available DaPars software relies on annotated distal PCS as input and can only work with a two-site model. Following identification of a proximal site, DaPars quantifies the relative use of the two sites between two conditions using the Percentage of Distal PolyA site Usage Index, or PDUI, value. PDUI is defined for each sample as the ratio of the estimated expression of the long transcript over the sum of the estimated expression values for the long and short transcripts. Hence, the greater the PDUI value, the more the distal poly-adenylation site is used. Mean PDUI values can be calculated across any number of samples and the difference of these values between two conditions is indicative of a differential APA event. DaPars assigns statistical significance to these differences by carrying out a Fisher exact test on estimated expression values for long and short transcripts between two conditions and corrects for multiple testing using the Benjamini-Hochberg approach. Finally, DaPars suggests filtering further the results for size of effect by highlighting events with higher ΔPDUI values (>=0.2) and fold change in PDUI of at least 1.5 (both are user-defined parameters and can be set to different values). Although DaPars is possibly the most widely applicable of all methods detecting APA events, given that it can be used with multiple samples and does not require prior knowledge of the proximal PCS, it has its own limitations. From a practical point of view, we have found that DaPars requires very good coverage at the 3' end to discover APA events, and will often only consider fewer than 10% of all genes expressed in a sample, rejecting the rest

due to lack of adequate coverage. Another limitation (not inherent in the DaPars algorithm but stemming from the software implementation of it), common to all methods that search the read data for cleavage events prior to annotated ends of transcripts, is that events representing lengthening of the 3'UTR (compared with the reference) are missed. For a method to discover these events, it needs to look beyond the annotated ends of transcripts. Moreover, in order to estimate the long and short transcript expression, DaPars implements a basic normalisation method for RNA-seq data, adjusting the counts for a region in a given sample by a weight factor corresponding to the sequencing depth difference between that sample and the mean of all other samples. This effective normalization by library size is known to cause biases in the outcome, if there are large differences in the expressed gene populations between samples. Additional biases resulting from covariates (e.g. gender, age, sequencing centre etc) affecting gene expression in each sample are also not taken into account, but these can influence the results where heterogeneous samples per condition are analysed. The statistical test used by DaPars may also be problematic, although the extent to which this affects the predictions of significant events is difficult to estimate. DaPars essentially compares normalized transcript expression values using a Fisher exact test. In other words, the values in the contingency table are no longer raw counts and hence the validity of how the test is applied may be questioned. Independence tests are common in many similar scenaria and have been used since the early days of analyses of differential PCS usage events [64]. However, applications using single test methods like chi-square or Fisher tests do not take into account replicates of a biological condition. Instead, they find ways of averaging values from the replicates in order to construct a contingency table with just one value per condition. The recently proposed non-parametric RAX2 method (ranking analysis of chi-squares)[81] attempts to overcome this problem by extending the chi-square test to the case of replicated count data from high-throughput transcriptomic experiments. An additional problem with the traditional chi-square or Fisher exact tests emerges in the case of multiple PCS/TSS; these tests ignore the order of columns in the contingency table and hence, there is an argument for the use of alternatives that would be more sensitive when information from multiple sites is available. Regression models of

the expected read counts may be a more suitable alternative, especially if they allow the error distribution to account for overdispersion. In addition, such models can naturally accommodate the presence of covariates, as implemented in the Generalised Linear Models included in the differential expression software solutions offered by DESeq2[82] and EdgeR[83].

More sophisticated methods for detecting and quantifying alternative poly-adenylation (as well as alternative TSS selection) are likely to be inspired by methods concentrating on the identification of alternative splicing events from RNA-seq data. The recently proposed change-point model that discovers splicing events in the 5' and 3' end of transcripts[84] is a promising step towards a more statistically robust method at finding read-density fluctuations in transcriptomic data. It uses negative binomial distributions with different parameters to describe read coverage in the common and extended regions of two isoforms separated by a splicing event. Additionally, it utilises annotation information or splice junction reads to assign different weights to each position, thus affecting the prior probability of each being a change-point. It is easy to see how a poly-adenylation sequence signal could be used in a similar manner when searching for isoforms that share the same splicing events but differ in their UTR lengths. Finally, the method employs an empirical Bayes estimator that allows pooling information across genes and results in powerful and accurate predictions of the change-points in the data.

As annotation of TSS and PCS becomes more complete and computational methods for the detection of alternative UTRs from high-throughput transcriptomic data mature, it is likely that initiatives will be set up to facilitate the study of the 5' and 3' UTRs using a combination of databases and software. In this spirit, the recently established expressRNA site (http://www.expressrna.org/) offers an integration of computational tools for the analysis of APA events. This and other future initiatives promise a more detailed, and ultimately more accurate, picture of the transcriptome.

Conclusions and prospects

The undisputed functional role of untranslated regions in eukaryotic regulation of gene expression dictates an emphasis on the development of isoform-centric approaches to data analysis. Until sequencing methods directed at both untranslated regions become widely accessible, it is important to focus on the easier to implement and more cost-effective computational approaches that can be applied to data from standard transcriptomic experiments. Methods aimed at identifying the TSS and PCS from the expression signals are of particular interest, as it remains debatable whether current annotations can be used reliably in the case of biological samples originating from conditions that have not been studied before. Importantly, although many existing methods were developed to study only one end of the transcripts, many have the potential to be applied to both transcript ends, allowing for a complete analysis of the transcriptome, given enough coverage.

The accurate identification of alternative TSS and PCS and the quantification of their differential use across samples by computational methods are undoubtedly challenging. One of the major hurdles remains the bias in read densities that is attributable to technical rather than biological origin. In this case, methods can take advantage of existing normalization procedures borrowed from recently developed robust differential expression software. The ability to process together multiple samples in order to contrast the use of alternative ends in different conditions is also a significant challenge. The use of multiple samples from heterogeneous origins adds the complication of dealing with confounding variables such as the site or type of sequencing, or characteristics of the sample's provenance (e.g. age, gender or ethnicity that often confound studies of human biological samples). Statistical methods that allow the inclusion of such covariates would be required for accurate modeling of the data. Finally, future developments should aim to include robust statistical methods that are appropriate for dealing with over-dispersed count data. The conspicuous lack of such methods may be at the root of low specificity in many cases where significance of variations in the use of sites across conditions is evaluated. We believe the field is ripe for further exploration and method development.

Conflict of interest

The authors declare they have no conflict of interest in the production of this manuscript.

References

1. **Mercer TR**, **Gerhardt DJ**, **Dinger ME**, **Crawford J**, et al. 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**: 99–104.

2. **Nellore A**, **Jaffe AE**, **Fortin J-P**, **Alquicira-Hernández J**, et al. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**: 266.

3. **Johannsen W**. 1913. *Elemente der exakten Erblichkeitslehre*. Jena: Fischer.

4. **Winter J**, **Kunath M**, **Roepcke S**, **Krause S**, et al. 2007. Alternative polyadenylation signals and promoters act in concert to control tissue-specific expression of the Opitz Syndrome gene MID1. *BMC Mol. Biol.* **8**: 105.

5. **Mignone F**, **Gissi C**, **Liuni S**, **Pesole G**. 2002. Untranslated regions of mRNAs. *Genome Biol.* **3**: REVIEWS0004.

6. **Barrett LW**, **Fletcher S**, **Wilton SD**. 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **69**: 3613–34.

7. **Hsin-Sung Yeh JY**. 2016. Alternative Polyadenylation of mRNAs: 3′-Untranslated Region Matters in Gene Expression. *Molecules and Cells* **39**: 281–5.

8. **Miura P**, **Sanfilippo P**, **Shenker S**, **Lai EC**. 2014. Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us? *Bioessays* **36**: 766–77.

9. **Lianoglou S**, **Garg V**, **Yang JL**, **Leslie CS**, et al. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**: 2380–96.

10. **MacDonald CC**, **McMahon KW**. 2010. Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *Wiley Interdiscip Rev RNA* **1**: 494–501.

11. **Ji Z**, **Lee JY**, **Pan Z**, **Jiang B**, et al. 2009. Progressive lengthening of 3'

untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.* **106**: 7028–33.

12. **Shepard PJ**, **Choi E-A**, **Lu J**, **Flanagan LA**, et al. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761–72.

13. **Park JY**, **Li W**, **Zheng D**, **Zhai P**, et al. 2011. Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules. *PLoS ONE* **6**: e22391.

14. **Fu Y**, **Sun Y**, **Li Y**, **Li J**, et al. 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21**: 741–7.

15. **Curinha A**, **Oliveira Braz S**, **Pereira-Castro I**, **Cruz A**, et al. 2014. Implications of polyadenylation in health and disease. *Nucleus* **5**: 508–19.

16. **Müller S**, **Rycak L**, **Afonso-Grunz F**, **Winter P**, et al. 2014. APADB: a database for alternative polyadenylation and microRNA regulation events. *Database (Oxford)* **2014**: bau076–6.

17. **Gruber AJ**, **Schmidt R**, **Gruber AR**, **Martin G**, et al. 2016. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**: 1145–59.

18. **Steijger T**, **Abril JF**, **Engström PG**, **Kokocinski F**, et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**: 1177–84.

19. **Angelini C**, **De Canditiis D**, **De Feis I**. 2014. Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics* **15**: 135.

20. **Kanitz A**, **Gypas F**, **Gruber AJ**, **Gruber AR**, et al. 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**: 150.

21. **Hayer KE**, **Pizarro A**, **Lahens NF**, **Hogenesch JB**, et al. 2015. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* **31**: 3938–45.

22. **Sharon D**, **Tilgner H**, **Grubert F**, **Snyder M**. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**: 1009–14.

23. **Bolisetty MT**, **Rajadinakaran G**, **Graveley BR**. 2015. Determining exon

connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16**: 204.

24. **Hoenen T**, **Groseth A**, **Rosenke K**, **Fischer RJ**, et al. 2016. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerging Infect. Dis.* **22**: 331–4.

25. **Quick J**, **Loman NJ**, **Duraffour S**, **Simpson JT**, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**: 228–32.

26. **Laver T**, **Harrison J**, **O'Neill PA**, **Moore K**, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3**: 1–8.

27. **Jain M**, **Fiddes IT**, **Miga KH**, **Olsen HE**, et al. 2015. Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**: 351–6.

28. **Loman NJ**, **Quick J**, **Simpson JT**. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**: 733–5.

29. **Deamer D**, **Akeson M**, **Branton D**. 2016. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**: 518–24.

30. **Garalde DR**, **Snell EA**, **Jachimowicz D**, **Heron AJ**, et al. 2016. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv* : 068809.

31. **Salamov AA**, **Solovyev VV**. 1997. Recognition of 3'-processing sites of human mRNA precursors. *Comput. Appl. Biosci.* **13**: 23–8.

32. **Tabaska JE**, **Zhang MQ**. 1999. Detection of polyadenylation signals in human DNA sequences. *Gene* **231**: 77–86.

33. **Graber JH**, **McAllister GD**, **Smith TF**. 2002. Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites. *Nucleic Acids Res.* **30**: 1851–8.

34. **Legendre M**, **Gautheret D**. 2003. Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**: 7.

35. **Cheng Y**, **Miura RM**, **Tian B**. 2006. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics* **22**: 2320–5.

36. **Ohler U**, **Niemann H**, **Liao Gc**, **Rubin GM**. 2001. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* **17 Suppl 1**: S199–206.

37. **Down TA**, **Hubbard TJP**. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12**: 458–61.

38. **Havukkala I**, **Vanderlooy S**. 2007. On the reliable identification of plant sequences containing a polyadenylation site. *J. Comput. Biol.* **14**: 1229–45.

39. **Ahmed F**, **Kumar M**, **Raghava GPS**. 2009. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. *In Silico Biol. (Gedrukt)* **9**: 135–48.

40. **Akhtar MN**, **Bukhari SA**, **Fazal Z**, **Qamar R**, et al. 2010. POLYAR, a new computer program for prediction of poly(A) sites in human sequences. *BMC Genomics* **11**: 646.

41. **Kalkatawi M**, **Rangkuti F**, **Schramm M**, **Jankovic BR**, et al. 2012. Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences. *Bioinformatics* **28**: 127–9.

42. **Xie B**, **Jankovic BR**, **Bajic VB**, **Song L**, et al. 2013. Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics* **29**: i316–25.

43. **Cui H**, **Wang J**. 2013. Machine Learning-Based Approaches Identify a Key Physicochemical Property for Accurately Predicting Polyadenlylation Signals in Genomic Sequences. In *Intelligent Computing Theories and Technology*. Springer Berlin Heidelberg. p 277–85.

44. **Lee T-Y**, **Chang W-C**, **Hsu JB-K**, **Chang T-H**, et al. 2012. GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. *BMC Genomics* **13 Suppl 1**: S3.

45. **Tzanis G**, **Kavakiotis I**, **Vlahavas I**. 2011. PolyA-iEP: A data mining method for the effective prediction of polyadenylation sites. *Expert Systems with Applications* **38**: 12398–408.

46. **Tian B**, **Graber JH**. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**: 385–96.

47. **Mironov AA**, **Fickett JW**, **Gelfand MS**. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9**: 1288–93.

48. **Brett D**, **Hanke J**, **Lehmann G**, **Haase S**, et al. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**: 83–6.

49. **Modrek B**, **Resch A**, **Grasso C**, **Lee C**. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29**: 2850–9.

50. **Huber W**, **Toedling J**, **Steinmetz LM**. 2006. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**: 1963–70.

51. **Bicknell AA**, **Cenik C**, **Chua HN**, **Roth FP**, et al. 2012. Introns in UTRs:

why we should stop ignoring them. *Bioessays* **34**: 1025–34.

52. **Li C**, **Wong WH**. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS* **98**: 31–6.

53. **Cambon AC**, **Khalyfa A**, **Cooper NGF**, **Thompson CM**. 2007. Analysis of probe level patterns in Affymetrix microarray data. *BMC Bioinformatics* **8**: 146.

54. **Sandberg R**, **Neilson JR**, **Sarma A**, **Sharp PA**, et al. 2008. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science* **320**: 1643–7.

55. **Akman HB**, **Oyken M**, **Tuncer T**, **Can T**, et al. 2015. 3' UTR Shortening and EGF signaling: Implications for breast cancer. *Hum. Mol. Genet.* **24**: ddv391–6920.

56. **Zhang H**, **Hu J**, **Recce M**, **Tian B**. 2005. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33**: D116–20.

57. **Salisbury J**, **Hutchison KW**, **Wigglesworth K**, **Eppig JJ**, et al. 2009. Probe-level analysis of expression microarrays characterizes isoform-specific degradation during mouse oocyte maturation. *PLoS ONE* **4**: e7479.

58. **Li L**, **Wang D**, **Xue M**, **Mi X**, et al. 2014. 3'UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network. *Scientific Reports* **4**: 5406.

59. **Erdman C**, **Emerson JW**. 2008. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics* **24**: 2143–8.

60. **Erdman C**, **Emerson JW**. 2007. bcp: an R package for performing a Bayesian analysis of change point problems. *J. Stat. Soft.*

61. **Wang Z**, **Gerstein M**, **Snyder M**. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**: 57–63.

62. **Sigurgeirsson B**, **Emanuelsson O**, **Lundeberg J**. 2014. Sequencing degraded RNA addressed by 3' tag counting. *PLoS ONE* **9**: e91851.

63. **Katz Y**, **Wang ET**, **Airoldi EM**, **Burge CB**. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**: 1009–15.

64. **Wang ET**, **Sandberg R**, **Luo S**, **Khrebtukova I**, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.

65. **Trapnell C**, **Roberts A**, **Goff L**, **Pertea G**, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–78.

66. **Le Pera L**, **Mazzapioda M**, **Tramontano A**. 2015. 3USS: a web server for detecting alternative 3'UTRs from RNA-seq experiments. *Bioinformatics* **31**: 1845–7.

67. **Birol I**, **Raymond A**, **Chiu R**, **Nip KM**, et al. 2015. Kleat: cleavage site analysis of transcriptomes. *Pac Symp Biocomput* : 347–58.

68. **Haas BJ**, **Delcher AL**, **Mount SM**, **Wortman JR**, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**: 5654–66.

69. **Pickrell JK**, **Marioni JC**, **Pai AA**, **Degner JF**, et al. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–72.

70. **Smibert P**, **Miura P**, **Westholm JO**, **Shenker S**, et al. 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* **1**: 277–89.

71. **Bonfert T**, **Friedel CC**. 2017. Prediction of Poly(A) Sites by Poly(A) Read Mapping. *PLoS ONE* **12**: e0170914.

72. **Bayerlová M**, **Klemm F**, **Kramer F**, **Pukrop T**, et al. 2015. Newly Constructed Network Models of Different WNT Signaling Cascades Applied to Breast Cancer Expression Data. *PLoS ONE* **10**: e0144014.

73. **Kim M**, **You B-H**, **Nam J-W**. 2015. Global estimation of the 3' untranslated region landscape using RNA sequencing. *Methods* **83**: 111–7.

74. **Nagalakshmi U**, **Wang Z**, **Waern K**, **Shou C**, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–9.

75. **Wu Z**, **Wang X**, **Zhang X**. 2011. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* **27**: 502–8.

76. **Roberts A**, **Trapnell C**, **Donaghey J**, **Rinn JL**, et al. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**: R22.

77. **Lu J**, **Bushel PR**. 2013. Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. *Gene* **527**: 616–23.

78. **Wang W**, **Wei Z**, **Li H**. 2014. A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics* **30**: 2162–70.

79. **Shenker S**, **Miura P**, **Sanfilippo P**, **Lai EC**. 2015. IsoSCM: improved and

alternative 3' UTR annotation using multiple change-point inference. *RNA* **21**: 14–27.

80. **Xia Z**, **Donehower LA**, **Cooper TA**, **Neilson JR**, et al. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274.

81. **Tan Y-D**, **Deng J**, **Neilson JR**. 2015. RAX2: a genome-wide detection method of condition-associated transcription variation. *Nucleic Acids Res.* **43**: e96–6.

82. **Love MI**, **Huber W**, **Anders S**. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**: 550.

83. **Robinson MD**, **McCarthy DJ**, **Smyth GK**. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–40.

84. **Zhang J**, **Wei Z**. 2016. An empirical Bayes change-point model for identifying 3“ and 5” alternative splicing by next-generation RNA sequencing. *Bioinformatics* **32**: 1823–31.

85. **Mayr C**, **Bartel DP**. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673–84.

86. **Rojas-Duran MF**, **Gilbert WV**. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**: 2299–305.

87. **Berkovits BD**, **Mayr C**. 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**: 363–7.

88. **Dieudonné F-X**, **O'Connor PBF**, **Gubler-Jaquier P**, **Yasrebi H**, et al. 2015. The effect of heterogeneous Transcription Start Sites (TSS) on the translatome: implications for the mammalian cellular phenotype. *BMC Genomics* **16**: 986.

89. **Hollerer I**, **Curk T**, **Haase B**, **Benes V**, et al. 2016. The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion. *RNA* **22**: 1441–53.

90. **Valen E**, **Pascarella G**, **Chalk A**, **Maeda N**, et al. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**: 255–65.

91. **Kanamori-Katayama M**, **Itoh M**, **Kawaji H**, **Lassmann T**, et al. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**: 1150–9.

92. **Plessy C**, **Bertin N**, **Takahashi H**, **Simone R**, et al. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* **7**: 528–34.

93. **Tang DTP**, **Plessy C**, **Salimullah M**, **Suzuki AM**, et al. 2013. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Res.* **41**: e44–4.

94. **Hestand MS**, **Klingenhoff A**, **Scherf M**, **Ariyurek Y**, et al. 2010. Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Res.* **38**: e165–5.

95. **Velculescu VE**, **Zhang L**, **Vogelstein B**, **Kinzler KW**. 1995. Serial analysis of gene expression. *Science* **270**: 484–7.

96. **Derti A**, **Garrett-Engele P**, **Macisaac KD**, **Stevens RC**, et al. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**: 1173–83.

97. **Wang L**, **Dowell RD**, **Yi R**. 2013. Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages. *RNA* **19**: 413–25.

98. **Zawada AM**, **Rogacev KS**, **Müller S**, **Rotter B**, et al. 2014. Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. *Epigenetics* **9**: 161–72.

99. **Jan CH**, **Friedman RC**, **Ruby JG**, **Bartel DP**. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**: 97–101.

100. **Martin G**, **Gruber AR**, **Keller W**, **Zavolan M**. 2012. Genome-wide Analysis of Pre-mRNA 3′ End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3′ UTR Length. *Cell Rep* **1**: 753–63.

101. **Wilkening S**, **Pelechano V**, **Järvelin AI**, **Tekkedil MM**, et al. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* **41**: e65–5.

102. **Harrison PF**, **Powell DR**, **Clancy JL**, **Preiss T**, et al. 2015. PAT-seq: a method to study the integration of 3'-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA* **21**: 1502–10.

103. **Hafner M**, **Renwick N**, **Brown M**, **Mihailović A**, et al. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**: 1697–712.

104. **Routh A**, **Ji P**, **Jaworski E**, **Xia Z**, et al. 2017. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA

enrichment or fragmentation. *Nucleic Acids Res.*

105. **Ozsolak F**, **Milos PM**. 2011. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA* **2**: 565–70.

106. **Pelechano V**, **Wei W**, **Steinmetz LM**. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–31.

107. **Ruan X**, **Ruan Y**. 2012. Genome wide full-length transcript analysis using 5“ and 3” paired-end-tag next generation sequencing (RNA-PET). *Methods Mol. Biol.* **809**: 535–62.

108. **Lee JY**, **Yeh I**, **Park JY**, **Tian B**. 2007. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.* **35**: D165–8.

109. **Derti A**, **Garrett-Engele P**, **Macisaac KD**, **Stevens RC**, et al. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22**: 1173–83.

110. **Dassi E**, **Re A**, **Leo S**, **Tebaldi T**, et al. 2014. AURA 2. *Translation* **2**: e27738.

111. **Wu X**, **Zhang Y**, **Li QQ**. 2016. PlantAPA: A Portal for Visualization and Analysis of Alternative Polyadenylation in Plants. *Front Plant Sci* **7**: 889.

112. **Brockman JM**, **Singh P**, **Liu D**, **Quinlan S**, et al. 2005. PACdb: PolyA Cleavage Site and 3'-UTR Database. *Bioinformatics* **21**: 3691–3.

113. **You L**, **Wu J**, **Feng Y**, **Fu Y**, et al. 2015. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* **43**: D59–67.

114. **Suzuki A**, **Wakaguri H**, **Yamashita R**, **Kawano S**, et al. 2015. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res.* **43**: D87–91.

Box 1.

Selected studies highlighting the role of alternative transcription start and alternative poly-adenylation sites in gene expression

Alternative promoters regulate the choice of poly-adenylation sites in the X-linked MID1 gene [4]
The MID1 3' UTR comprises four polyadenylation sites and the selection of these sites is linked to promoter usage, as confirmed by RT-PCR experiments utilizing primers covering different exons in the 5' UTR. This suggests that factors interacting with both ends of the gene may be simultaneously regulating the two UTR lengths.

Cancer cells evade miRNA repression by shortening the 3' UTR of transcripts [85]
In some cancer cell lines shortening of the 3' UTR leads to isoforms that are more highly expressed, primarily through the evasion of miRNA-mediated repression.

The efficiency of translation depends on the length of the 5' UTR [86]
Short and long isoforms of 9 yeast genes show different translation efficiencies depending on the 5' UTR length, with two of the genes tested showing > 100-fold difference in translation activity.

3' UTRs control the fate of a gene product post-translationally [87]
In human cell lines, the 3' UTR of the CD47 transcript recruits a protein complex and transports it to the site of translation, where it subsequently interacts with the newly synthesized protein to allow its translocation to the plasma membrane.

Selective recruitment of transcript variants to polysomes based on their 5' UTRs allows rapid cell response with no changes to the transcriptional programme [88]
A study of the transcriptome and translatome (transcripts associated with polysomes) in both tumour and normal breast cell lines points to cell-specific heterogeneity in transcript leaders (5' UTRs) as well as differential recruitment of transcript variants by ribosomes in disease states.

Stress induces alternative poly-adenylation and regulates gene expression in mammalian cells [89]
Poly(A) sites were mapped using the 3'T-fill method in control and stress-provoked HEKT293 cells. 401 genes were shown to exhibit stress-induced alternative poly-adenylation and the majority of these events did not affect directly mRNA abundance but instead were linked to changes in the mRNA configuration.

Box 2.

An overview of experimental methods for probing alternative TSS and PCS

A relatively limited number of methods have been developed to capture the 5' end of transcripts at nucleotide resolution. Initial efforts based on a 5' version of Serial Analysis of Gene Expression (SAGE)[Hashimoto et al. 2004)] or the capturing of the 7-methylguanosine cap (Cap Analysis of Gene Expression or CAGE [Shiraki et al. 2003; Carninci et al. 2005; Kodzius et al. 2006) have been superseded by versions that rely on next-generation sequencing (DeepCAGE[90] and HeliScopeCAGE[91]. HeliScopeCAGE is of particular interest as it is built on single molecule sequencing technology, avoiding the clonal amplification step that potentially leads to quantification bias, as well as several other error-prone steps such as second strand synthesis, ligation and digestion. Nano-cap analysis of gene expression (nanoCAGE)[92] uses the template switching method (Zhu et al. 2001) for reverse transcription, reducing the amount of starting RNA material needed but, importantly, not at the cost of coverage. However, template switching can introduce artifacts (cDNAs that are artificially shorter than the corresponding RNAs) owing to a process known as strand invasion[93]. Moreover, as with all CAGE-based methods, nanoCAGE may produce tags covering the whole transcript (a phenomenon known as "exon painting"[94]), rendering the calling of TSS challenging. Hence, caution is advised in the interpretation of intra-genic clusters of CAGE reads as alternative TSS.

Like its 5' equivalent, the 3' specific SAGE-based method[95] has been superseded by a variety of RNA-seq derivatives focused on the 3' end of transcripts. Several protocols have been developed but there is considerable overlap between many of these approaches. For example, the PolyA-seq [96] and PAS-Seq [12] methods vary only in minor technical details such as the type of primer used. The majority of protocols capture the ends of transcripts by relying on oligo(dT)-primed reverse transcription–PCR (RT–PCR) for library construction. This approach has the drawback that primers hybridize to intergenic stretches of adenosines, resulting in internal priming and interfering with the prediction of the PCS. A variety of methods have been created to deal

with this problem. The earlier Sequencing APA Sites (SAPAS) method filters reads resulting from internal priming during the bioinformatics analysis pipeline[14]. Similarly, Wang et al.[97] have suggested a bioinformatics solution to filtering out internal priming events; the method takes advantage of the distinct nucleotide composition patterns found around poly-adenylation sites to computationally identify authentic 3' end cleavage events. The 3'READS method (Hoque et al. 2012) minimises the internal priming problem by elongating oligo(T) primers to 45 bp, whereas "hot priming" has been used by Muller et al.[16] during the hybridization step of Massive Analysis of cDNA Ends (MACE)[98] to avoid the same problem. In 3P-Seq[99] a biotinylated double-stranded oligo (with overhanging stretch of Ts) is ligated to the end of the poly(A) tail, avoiding the need for oligo(dT) priming. Another issue that must be addressed by protocols that sequence in the sense direction, i.e. that start the sequencing within the transcript and proceed towards the poly(A) tail (e.g. A-seq[100]), is fragment selection: fragments must be long enough to allow the production of mappable reads to the genome, whilst at the same time allowing the polya-denylation site to be reached. When the protocol supports reading through the poly(A) site, the resulting reads suffer from low quality[101]. The 3' T-fill method[101] avoids reading through the poly(A) tail by filling in the stretch of As with base-pairing unlabeled dTTPs, forcing sequencing to start directly after the tail and into the 3'UTR region. The PAT-Seq method[102] opts for the opposite approach, deliberately including the poly(A) tail in the sequencing, but avoiding the problems associated with sequencing homopolymers by starting the sequencing on the 5' end of fragments. The inclusion of the poly(A) stretch allows studies that correlate the adenylation state to gene expression, measuring 3'UTR dynamics on a genome-wide scale. The above protocols are powerful in providing the precise locations of PCS, but complex library preparation steps make them error-prone and introduce technical biases; adaptor ligation, for example, can yield poor quantification in deep sequencing because of significant bias introduced by RNA ligase[103]. The recently proposed Poly(A)-Clickseq (PAC-seq) method[104] avoids sample purification and poly(A) enrichment, as well as the RNA fragmentation and adaptor ligation steps by using azido-nucleotides to terminate cDNA synthesis just upstream of the 3'UTR-poly(A)

junction, followed by click-chemistry to allow the ligation of special Illumina adaptors prior to PCR amplification. Another promising alternative approach is Direct RNA Sequencing (DRS) [105]. Similarly to HeliScope-CAGE, DRS has the major advantage of avoiding the reverse transcription–PCR step during library construction by sequencing directly captured poly-adenylated RNAs on oligo(dT)-coated slides. The DRS method requires very small amounts of RNA and has been shown to quantitatively reflect the amount of transcript isoform present in the RNA sample [105]. This promising technology is yet to be widely implemented, possibly due to higher error rates compared with other methods.

A limitation of methods that focus on either the 5' or the 3' end is that they essentially require three different experiments to probe the two UTRs and coding region of the mRNA. Consequently, experimental methods that combine 3' and 5' sequencing are obviously advantageous and potentially time and cost-efficient. Tag-based transcript-isoform sequencing (TIF-Seq)[106] and RNA-PET[107] offer such solutions using paired-end sequencing. Both methods combine approaches developed for directed 3' and 5' sequencing. RNA-PET combines protocols from 3'-SAGE and 5'-CAGE to capture both ends of transcripts, whereas TIF-Seq employs nanoCage for 5' and oligo(T)- primer for 3' sequencing. Both methods are limited in their coverage. More specifically, RNA-PET identifies a much smaller number of isoforms compared with the directed sequencing approaches, whereas TIF-Seq results are affected by strong bias towards short RNA molecules[106].

Table 1.

List of databases with information on alternative transcription start sites (TSS) and alternative poly-adenylation and cleavage sites (PCS)

Database name	Alternative PCS or TSS	Organism	Technology	Comments	Web address	Reference
PolyA_DB (2)	PCS	Human; Mouse; Rat; Chicken; Zebrafish	cDNA/EST and TRACE		http://exon.umdnj.edu/polya_db/	Lee et al. 2007[108]
AURA 2	PCS	Human; Rhesus; Dog Mouse; Rat	PolyA-seq	This is a "meta" database. Alternative UTR information is from [109]	http://aura.science.unitn.it/	Dassi et al. 2014[110]
PlantAPA	PCS	Plants	EST, PAT-seq, Poly(A)-tag	Only plant species are included in this database.	http://bmi.xmu.edu.cn/plantapa/	Wu et al. 2016[111]
PolyASite	PCS	Human; Mouse	2P-seq, 3'Seq, 3'READS, 3P-Seq, A-seq, A-seq2, DRS, PAS-Seq, PolyA-seq, SAPAS		http://polyasite.unibas.ch	Gruber et al. 2016[17]
PACdb			EST	Uses an older version of the human genome (hg18)	http://harlequin.jax.org/pacdb/	Brockman et al. 2005[112]
APADB	PCS	Human; Mouse; Chicken	Massive Analysis of cDNA Ends (MACE)		http://tools.genxpro.net/apadb/	Müller et al. 2014[16]
APASdb		Human; Mouse; Zebrafish	SAPAS		http://genome.bucm.edu.cn/utr/	You et al. 2015[113]
DBTSS	TSS	Human	TSS-seq	Uses an older version of the human genome (hg18)	http://dbtss.hgc.jp	Suzuki et al. 2015[114]

Figure Captions

Figure 1.

Schematic representation of alternative transcription start site and alternative poly-adenylation events of interest to this review.

The presence of two alternative TSS and two alternative PCS creates four possible transcripts from the same genomic locus (top). TSS1 creates transcripts with longer 5' UTRs compared with TSS2, whereas PCS1 creates transcripts with shorter 3'UTRs compared with PCS2. The use of alternative TSS/PCS regulates the inclusion or exclusion of functional elements such as upstream open reading frames (uORFs) in the 5' UTR or miRNA and protein-binding sites in the 3' UTR. The coding regions (light grey boxes and grey lines) may or may not be different between the transcripts, depending on the action of alternative splicing. In this review we are only concerned with differences in the untranslated regions (blue and yellow).

Figure 2.

The distribution of microarray probes along the length of the transcript limits their usability for genome-wide studies of alternative TSS and PCS.

A & B:

The mapped position of microarray probes for three generations of Affymetrix chips are shown as coloured rectangles along the 5' (A) and 3' UTR (B) of two transcripts (brown, pink and blue for Human Genome 133 Plus, Human Exon 1.0 ST and Human Transcriptome Array 2.0 respectively). Annotated TSS and PCS sites are shown as brown vertical lines along the body of the UTR. Even in the simplest case of a single TSS site (A), it is obvious that only the recently developed Human Transcriptome Array contains enough probes to give satisfactory coverage of both alternative 5' UTRs. In the more complex case of multiple known sites shown in (B), even the extended coverage provided by HTA 2.0 is not sufficient for unambiguous discovery and quantification of alternative poly-adenylation events for this transcript.

C & D:

The Human Transcriptome Array 2.0 platform displays good coverage of the 5' UTR (C; left) with a median number of 20 probes per transcript. Splitting the 5' UTR into quartiles along its length and calculating the coverage of each from the HTA 2.0 probes indicates a reasonable distribution of probes along the length of the UTR (C; right). The coverage of the 3'UTR by the same microarray chip is worse (D; left), despite a higher median number of probes (23). This discrepancy is due to the generally longer lengths of the 3' as compared with the 5' UTRs. Encouragingly, the distribution of probes along the length of the 3'UTR is also good with only a small bias towards the first quartile (D; right).

Figure 3.

Discovering alternative poly-adenylation events from read density fluctuations in RNA-seq data.

Reads from an RNA-seq experiment[72] are mapped to the reference genome and the density of the reads is displayed as a Sashimi plot, depicting coverage along the gene body (only the 3' end of the transcripts are shown for genes *EIF2S3* and *XRN2*). Black vertical lines mark the annotated PCS sites from the PolyASite database[17]. The two vertical red lines represent the sites predicted by the program DaPars as alternative poly-adenylation sites, based on a drop in read densities along the 3' UTR. (DaPars is using a number of samples to identify the sites, not just the one shown here). The DaPars prediction agrees well with one of the annotated PCS sites in the *EIF2S3* gene (A) but lies far from the annotated proximal PCS site in the *XRN2* gene (B). However, the drop in read density around the DaPars predicted site suggests the possibility that the prediction is correct. Clearly, the presence of multiple peaks in density in (A) would be challenging for any software that predicts APA events using density fluctuations in the RNA-seq signal.