Trends in Microbiology

CellPress
REVIEWS

Review

Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research

Zsolt Boldogkői,[1,*] Norbert Moldován,[1] Zsolt Balázs,[1] Michael Snyder,[2] and Dóra Tombácz[1]

Long-read sequencing (LRS) has become increasingly popular due to its strengths in *de novo* assembly and in resolving complex DNA regions as well as in determining full-length RNA molecules. Two important LRS technologies have been developed during the past few years, including single-molecule, real-time sequencing by Pacific Biosciences, and nanopore sequencing by Oxford Nanopore Technologies. Although current LRS methods produce lower coverage, and are more error prone than short-read sequencing, these methods continue to be superior in identifying transcript isoforms including multispliced RNAs and transcript-length variants as well as overlapping transcripts and alternative polycistronic RNA molecules. Viruses have small, compact genomes and therefore these organisms are ideal subjects for transcriptome analysis with the relatively low-throughput LRS techniques. Recent LRS studies have multiplied the number of previously known transcripts and have revealed complex networks of transcriptional overlaps in the examined viruses.

Highlights

Long-read sequencing (LRS) has revolutionized genomics and transcriptomics. These third-generation approaches have a relatively low throughput compared to short-read sequencing, but they can solve problems that used to be a challenge for earlier techniques.

The PacBio and ONT sequencing are able to read full-length transcripts and allow the direct study of base modifications on both DNA and RNA molecules. Nanopore technology is able to sequence RNA directly.

LRS has revealed a much more complex viral transcriptome. Among other capabilities, these techniques allow the discrimination between multispliced transcript variants, RNA length isoforms, embedded RNAs, and polycistronic RNA molecules.

The viral genomes express a highly complex pattern of transcriptional overlaps, the function of which continues to remain unknown.

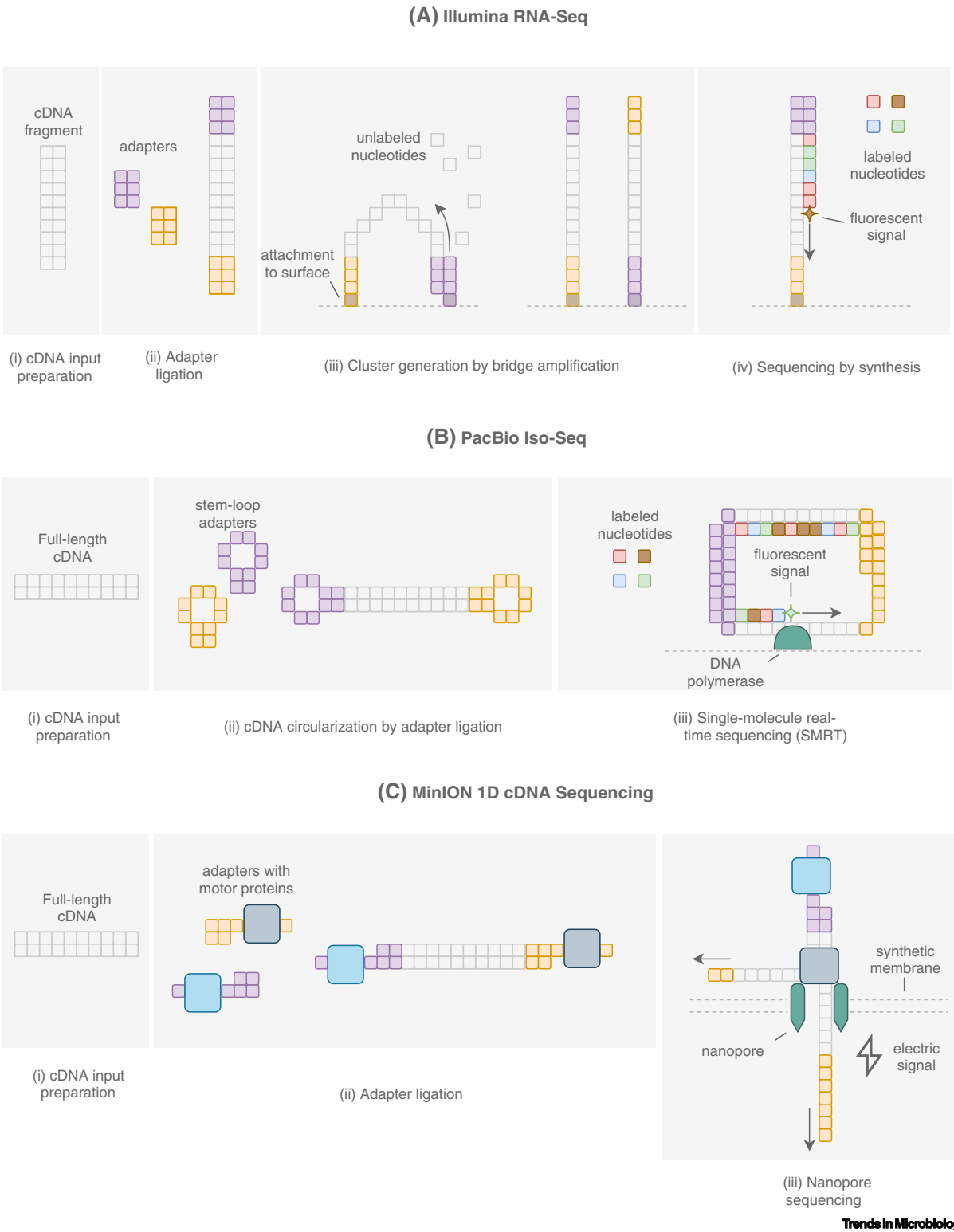
Transcriptome Research on Viruses

Viruses represent a diverse class of microorganisms with polyphyletic origin. Compared to cellular organisms, even the largest DNA viruses have small genomes with closely-spaced genes. This feature makes viruses excellent model systems in molecular biology to explore the general principles of genetic regulation and transcriptome organization.

Alternative splicing increases the coding potential of the genome through the production of multiple RNA and protein molecules from a single gene. Similarly, alternative transcription initiation and termination also contribute to the genomic complexity. Polycistronism is a common phenomenon in bacteria and in their viruses but it is extremely rare in eukaryotes. The reason for this is that, in prokaryotes, the Shine–Dalgarno sequences allow the translation of each gene in the mRNA [1]. Nevertheless, in eukaryotes only the most upstream gene of a polygenic transcript is translated because of the Cap-dependent initiation system. Some small RNA viruses have evolved miscellaneous strategies to solve the problem of translation of multiple (generally two) proteins from a single transcript, which includes the utilization of an internal ribosome entry site (IRES), or mechanisms to bypass the 5′-proximal AUG to enable downstream initiation, such as the leaky ribosomal scanning mechanisms and ribosomal frameshifting [2]. Nonetheless, in the majority of DNA viruses no such mechanisms have been described so far. The canonical termination sequences are not always efficient in stopping the RNA polymerase (RNP); therefore, transcription is continued until the next termination site is reached, which results in transcriptional readthrough (TRT) producing readthrough (rt)RNAs. Transcriptional overlaps (TOs), in most cases produced by TRT, have been shown to represent a common phenomenon in diverse organisms [3]. The latest studies have also shown an intricate meshwork of TRTs and TOs in various viruses.

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary
[2]Department of Genetics, School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA 94305-5120, USA

*Correspondence: boldogkoi.zsolt@med.u-szeged.hu (Z. Boldogkői).



(See figure legend at the bottom of the next page.)

The next-generation short-read sequencing (SRS) technology, released in the mid-2000s, has revolutionized genomic and transcriptomic sciences due to its massively parallel nature, which has enabled sequencing of millions of DNA fragments simultaneously at a relatively low cost. The Illumina platform is by far the most widely applied SRS technique. The enormous number of reads generated by SRS enabled the sequencing of entire genomes of various organisms at an unprecedented speed. The currently running genome programs are mainly based on the SRS approach [4,5]. This technique has also been extensively used to study the transcriptomes of various organisms [6,7]. The third-generation LRS technology emerged in 2011, when Pacific Biosciences (PacBio) commercialized the single-molecule real-time (SMRT®) technology [8]. Currently, two LRS technologies are in use: the PacBio and the Oxford Nanopore Technology (ONT) platforms. MinION, the first prototype of ONT was released in 2014 [9]. Both LRS techniques are based on the development of novel biochemistry, which enables the direct capture of long DNA sequences or cDNAs from full-length transcripts. ONT has also developed a method for sequencing native RNAs [10] and, since Helicos [11] has withdrawn from the market, nanopore sequencing is the only commercially available direct (d)RNA sequencing method. SRS, however, is still outstanding for producing high-quality, deep-coverage datasets. This technique is more cost-effective and has a lower per-base error rate than the LRS approaches [12,13]. Complex genomic regions, including sequences with a high GC content, as well as repetitive sequences, cannot be efficiently resolved by SRS. The short read length also makes computations difficult or impossible for the determination of exon connectivity and for the identification of transcript isoforms, such as multispliced RNAs as well as transcription start site (TSS) and transcription end site (TES) variants [14]. Furthermore, alternative polycistronism and TOs, especially between embedded RNA (eRNA) molecules, pose challenges for the SRS platforms. These challenges can be overcome by the LRS technology since it is able to provide full contig information about transcripts.

In recent years, LRS has been widely utilized in the analysis of the genomes of various organisms including prokaryotic [15] and eukaryotic [16] species as well as viruses [17–22]. Nevertheless, current LRS techniques are only able to characterize small genomes and transcriptomes in high depth due to the comparatively low throughput. Viral transcriptomes used to be investigated by traditional techniques, including Northern blotting [23,24], quantitative PCR [25,26], RACE analysis [27], and microarray studies [28,29]. The introduction of Illumina sequencing [30–32] to virus research has led to a significant progress in the discovery and precise annotation of viral transcripts. Currently, the global transcriptome of several viruses belonging to different families has been analyzed by using various techniques of PacBio and ONT platforms [33–37]. Our review aims to provide an overview of the potentials and limitations of LRS methods, to present the transcriptome diversity that has been detected by long-read sequencing, and to discuss future paths of viral genomics opened up by this technology.

LRS

Similar to the Illumina approach, PacBio also adopts a sequencing-by-synthesis strategy, but while Illumina detects augmented signals from amplified DNA fragments, the PacBio technique captures a single DNA molecule (Figure 1 and Table 1). The PacBio SMRT® sequencing utilizes

Figure 1. Comparison of the Various Sequencing Platforms. (A) Illumina sequencing uses cDNA fragments, each of which is amplified multiple times to form a cluster. Sequencing is based on the fluorescent signal of the incorporated nucleotides. The emitted fluorescence from each cluster is strong enough to generate a specific signal in the detector. (B) During PacBio sequencing, full-length cDNA is circularized by stem-loop adapters and loaded into ZMWs, where the immobilized polymerase incorporates fluorescently labeled nucleotides. ZMWs are capable of detecting the fluorescent signal of a single nucleotide incorporation. The polymerase is able to take multiple passes on the cDNA template; thus, the consensus reads are more accurate than the raw reads. (C) The motor proteins ligated to the cDNA templates bind to nanopores on a synthetic membrane in MinION sequencing. The motor proteins ratchet one strand of the DNA through the nanopore, whereas a detector is measuring the potential changes on the two sides of the membrane. The electric signal is specific to the nucleotides passing through the membrane.

Table 1. Comparison of the Various Sequencing Platforms.

		Illumina		Pacific Biosciences		Oxford Nanopore Technologies	
		HiSeq	MiSeq	RSII	Sequel	MinION 1D	dRNA-Seq
Required amount of input material (ng)		1–50	1–50	–	1000	1000	500–775
Mapped read length (bp)	Mean	92	175	2088	13 800	1503.50	968
	Median	–	–	1720	–	1439	713
	Standard deviation	–	–	1438.14	–	969.18	–
	Maximum	101	250	8006	–	9345	21 866
Average percentage of mapped reads		84.1	84.4	90.29	–	69.27	96.5
Substitutions per base		0.0053	0.0142	0.0212	0.005	0.0754	0.024
INDELs per base		7.2×10^{-6}	–	0.0231	0.001	0.0856	0.0435
Sample type		gDNA	gDNA	cDNA	gDNA	cDNA	RNA
Mapping software used in the study		BWA-MEM	BWA-MEM	GMAP	–	GMAP	Minimap2
Refs		[41]	[41]	[42]	[43]	[42]	[44]

zero-mode waveguides (ZMW) [38] for single-molecule analysis, which allows the detection of fluorescent signals emitted during the incorporation of labeled nucleotides. A single DNA polymerase molecule, fixed at the bottom of a ZMW, reads the circularized template multiple times. When a nucleotide is incorporated in the growing DNA strand, the fluorescent tag is cleaved off, and it gets out of the observation area. The base-call is made by the detection of the fluorescent signal of the nucleotide incorporated within the ZMW [39]. The accuracy of the obtained consensus sequence (reads of inserts, ROI) depends upon the number of polymerase passes around the circular template [12]. Sequel, the newest PacBio platform, launched in 2015, has a capacity sevenfold greater than the former RS II platform [40]. Additionally, the Sequel system has a considerably decreased loading bias compared to RS II; therefore, it does not require size-selection[i]. SMRT® sequencing generates subreads, thereby resulting in multiple base coverage in a given base, which leads to increased precision. The base-calling accuracy depends on the read length and on the movie length. The PacBio Isoform sequencing (Iso-Seq®) allows the generation of full-length cDNA sequences without the need for contig assembly, and thus it is suitable for the confident characterization of the full complement of transcript isoforms across an entire transcriptome or within the targeted genes.

The nanopore technology is based on monitoring the transit of DNA or RNA molecules through a protein pore; it measures variations in electric currents produced by the nucleotides that are threaded through the nanopores aided by a molecular motor protein. Nanopore sequencing is able to determine very long nucleic acid sequences [45]. ONT 1D sequencing has low accuracy (approximately 85%) [46]. The 1D2 technology improves this accuracy by ligating an adapter to the end of the reads, which increases the probability that a strand and its complementary strand pass through a pore consecutively. The base-calling algorithm creates a consensus of the two reads, and it has an average quality of over 95%. ONT can also identify the nucleotide modifications of RNA molecules by native RNA sequencing [10]. The advantages of nanopore sequencing over the PacBio platform are the longer read length, the higher throughput, and the lower costs

[42]. The two LRS techniques are prone to similar errors, such as homopolymer bias and indel errors. High sequencing error rate makes accurate DNA sequencing difficult, including *de novo* sequencing and variant calling. Sequencing errors, however, do not represent a major obstacle in transcriptome research if well-annotated genome sequences to which the transcript reads can be aligned are available. The lower throughput compared to the SRS approach means that LRS can only characterize abundant transcripts and that technical by-products are more difficult to filter out from LRS data. Ligation and template switching are common causes of such artifacts. Template switching is caused by the release of the template strand by the polymerase molecule during synthesis followed by binding to another template that shares homology with the original template and can occur at both the reverse-transcription (RT) [47] and the PCR [48] steps. The advantage of dRNA sequencing of ONT is that it is free from RT and PCR artifacts. The shortcomings of the current dRNA technique are its demand for the starting material (at least 500 ng PolyA$^+$ RNA), very low throughput, and that the produced reads lack short sequences at both the 5′ and 3′ termini [35]; therefore it does not resolve isoforms with base-pair precision. The cDNA sequencing methods require slightly less starting material (at least 250 ng PolyA+ RNA without PCR or 200 ng amplicon after PCR using ONT protocols or 2 ng total RNA using the Iso-Seq® protocol) and have a higher throughput, although not as high as SRS techniques. To date, the majority of LRS protocols have focused on full-length polyA-selected RNAs [49]. Cap-selection of RNA molecules can also be used to enrich full-length RNA molecules [50].

SRS can Complement LRS

It has been demonstrated that SRS coupled with the so-called synthetic long-read sequencing method (SLR-Seq) can represent an alternative approach for full-length characterization of transcripts [51] at the cost of reducing the sequencing yields. SLR-Seq is also afflicted by SRS biases, such as poor characterization of GC-rich regions. LRS is definitely superior to SRS regarding isoform detection and the differential quantitation of isoforms; SRS still offers many advantages and could be used alongside LRS techniques [52]. For example, ChIP-Seq [53] and ribosome profiling [54,55] are methods which provide valuable information and are not well suited for current LRS technologies. SRS can also be used to improve LRS not only through error correction (Box 1)

Box 1. Technology Corner: The Bioinformatic Challenges of LRS

Owing to the higher error rate but greater length of long reads, different tools are needed to analyze LRS and SRS data. The preprocessing of the reads is platform-specific. SMRT Link from PacBio creates accurate consensus reads and also assembles consensus isoforms which can be mapped to the genome or can be analyzed further without the need for a genome sequence [90]. Such consensus isoforms are usually highly accurate, and no further error correction is necessary. The processing of nanopore reads is less standardized. Guppy has recently been declared to be the recommended base caller by ONT. For genome sequencing, the next step would be error correction either by using short reads or based solely on the nanopore sequencing [91–93]. For transcriptome sequencing, however, error correction does not appear to be beneficial prior to isoform identification as it may interfere with both the qualitative and the quantitative analysis [94]. A novel method which uses rolling-circle amplification to produce concatemers of cDNA molecules (R2C2) greatly improves the quality of nanopore reads while preserving the benefits of single-molecule sequencing [95]. Minimap2 is used for the alignment of reads from both sequencing technologies [96]. Recently, a number of software programs have been developed for the task of isoform discovery. Mandalorion was designed for isoform detection in 2D reads [97]; however, ONT no longer supports 2D technology, the new version of the pipeline currently only accepts highly accurate R2C2 reads. FLAIR focuses on the splice isoforms and requires an annotation of splice sites [98]. Pinfish[ii] bases isoform discovery on the clustering of reads to define median exon boundaries, whereas LoRTIA[iii] identifies and filters transcript features (TSS, intron, and TES), then constructs isoforms based on these features. At the moment, all of these tools require an available genome sequence to determine the exact nucleotide sequence of the transcripts, but, except for FLAIR, they do not require an existing annotation of transcript features such as splice sites Therefore they can be applied to the investigation of viral transcriptomes that have not been studied before. Due to the numerous challenges in the analysis of LRS data and the relative novelty of these tools, artifacts may be common and should be filtered out by inspection or, ideally, by quality-control programs, such as SQANTI, a pipeline that characterizes and filters isoforms based on an annotation [99].

but also through the precise characterization of transcript features, which can be helpful in isoform identification. PRO seq [56] identifies TSSs, whereas technologies such as 3′READS+ [57] characterize TESs with higher sensitivity and specificity than is possible with current LRS technologies. Concurrently, LRS can be used for the precise quantitative analysis of the viral transcriptome at the isoform level [58]. Using a non-amplified Iso-Seq® technique, the results of the kinetic categorization of PRV transcripts have been in agreement with earlier observations obtained by real-time RT-PCR analysis [26]. The ability of LRS methods to sequence PCR-free cDNA or RNA provides more accurate quantitation that is devoid of amplification bias. However, due to the low throughput of such LRS methods, SRS is still more efficient in characterizing host transcription. When combined with microfluidic technologies provided by 10x Genomics, LRS can differentiate between transcript isoforms whereas SRS can characterize gene expression at the level of a single cell [59], resulting in a more specific analysis of the viral–host interactions. Using LRS and SRS coupled with other techniques can eschew the deficiencies of each approach and opens the possibility of a wider analysis of the viral transcriptome.

Novel Viral Transcripts Identified by LRS

LRS techniques have already been used in the investigation of the transcriptomes of various viruses, including herpesviruses [33–37], baculoviruses [60], retroviruses [61], circoviruses [62], and poxviruses [22,63]. The application of these techniques has identified a much greater complexity of viral transcriptomes compared to earlier approaches. Most of the newly discovered transcripts belong to categories which are difficult to study with SRS and other techniques, such as embedded messenger RNAs (emRNAs), polygenic transcripts, overlapping transcripts, and RNA isoforms including splice, TSS, and TES variants. LRS studies have also revealed noncoding RNAs (ncRNAs), such as antisense RNAs (asRNAs), intergenic RNAs (iRNAs), and embedded noncoding RNAs (encRNAs). Special classes of transcripts termed near-replication-origin RNAs (nroRNAs) and nro-like transcripts have also been recently described [64,65].

Messenger RNAs

LRS techniques are especially efficient in identifying eRNAs sharing a common TSS or TES with the longer host transcript. An emRNA containing an in-frame open reading frame (ORF) within the coding region of the longer host gene has the potential to specify an N-terminally truncated polypeptide. LRS studies have multiplied the number of these truncated mRNAs in each subfamily of herpesviruses [35–37,65], and in a baculovirus [60].

Noncoding Transcripts

It used to be believed that DNA and protein molecules play a fundamental role in the control of cell functions, whereas RNAs were considered to have only subsidiary roles. Nonetheless, recent transcriptomics studies have revealed a huge variety of ncRNAs with a wide range of functions, including epigenetic, transcriptional, and post-transcriptional regulation of gene expression [66]. These transcripts are classified below according to their locations and orientations.

Intergenic Transcripts

These RNA molecules are located between two coding sequences without or with no significant overlap with the adjacent genes. LRS studies have identified several iRNAs in viruses, although these techniques are not superior to SRS in the detection of these types of transcript.

Embedded Noncoding Transcripts

The encRNAs can be mapped within either mRNAs or ncRNAs. The most typical mRNA-overlapping encRNAs are the 5′- and 3′-truncated transcripts. Generally, encRNAs have a common TSS or TES with other ncRNAs with which they overlap.

Antisense Transcripts

The asRNAs either completely or partially overlap the mRNAs in an antiparallel manner. These transcripts can either be controlled by their own promoters or they can be the result of transcriptional overlaps between neighboring or distal convergent or divergent genes (Figure 2). In this latter case, only the overlapping part of the transcript is antisense.

Replication-Associated RNAs, a Novel Class of Transcripts

The replication-associated RNAs (raRNAs) are mapped in close vicinity to the replication origins (Oris) of viral DNA [33,64,65]. LRS techniques have detected several such RNA molecules, including nroRNAs in herpesviruses [31,36,37], and nro-like transcripts in baculoviruses [60] and circoviruses [62]. Six types of replication-associated transcript can be distinguished in terms of their coding potency and position to the Ori: (i) mRNAs that do not overlap the Ori; (ii) ncRNAs that do not overlap the Ori; (iii) ncRNAs that do overlap the Ori; (iv) mRNA isoforms with very long overlapping alternative TES; (v) mRNA isoforms with very long overlapping alternative TSS; and (vi) mRNA with Ori overlapping ORF (Figure 3).

Readthrough RNAs

The readthrough RNAs (rtRNAs) are produced by occasional TRT of coding or noncoding genes due to the inefficient recognition of transcriptional termination sequences by the RNP molecules. The rtRNAs are TES variants if they have an exact termination site and contain the same ORF as the shorter transcript isoform. Complex transcripts are also the results of TRTs, whereas the polycistronic RNAs can only be considered as rtRNAs if the upstream genes also have their own transcription termination signals. Studies using quantitative RT-PCR [26] and Illumina sequencing [31, 67] have demonstrated a pervasive, genome-wide expression of asRNAs in herpesviruses. It is possible that these transcripts lack poly(A) tails and that is why they are undetected by oligo (dT)-primed sequencing techniques. These molecules may be expressed at a low abundance and/or have a short half-life due to the lack of polyadenylation.

Transcript Isoforms

Transcript isoforms include the length and splice variants of mRNAs and ncRNAs.

Transcription Start-Site Isoforms

Genes can be controlled by alternative promoters which express TSS isoforms [68]. Multiple promoter-controlled genes can be differentially expressed throughout the viral life cycle [69]. Additionally, TSS isoforms can have diverse functions as the various 5′-UTR (untranslated region) structures can control the translation in a differential manner (as reviewed in [70]).

Transcription End-Site Isoforms

In certain cases, the progression of RNPs does not stop at the transcription termination sequences, which leads to longer TES variants [71,72]. On many occasions, these longer molecules overlap the transcripts generated by the downstream genes. SRS has been utilized to identify alternative polyadenylation [73]. LRS provides additional information about the various 5′- and 3′-UTR combinations utilized by the various transcript isoforms. It is important, however, to filter the putative polyadenylation sites for signs of internal priming, as it has been shown that adenine-rich regions may appear as false polyadenylation sites [74].

Splice Isoforms

The PacBio Iso-Seq® technique is especially suitable for the detection of novel splice sites [75–77]. Furthermore, alternatively processed multispliced transcripts can be reliably identified only by the LRS techniques; hence, they are able to sequence full-length transcripts and thus to map exon connectivity. Splicing events are relatively rare in alphaherpesviruses, baculoviruses,

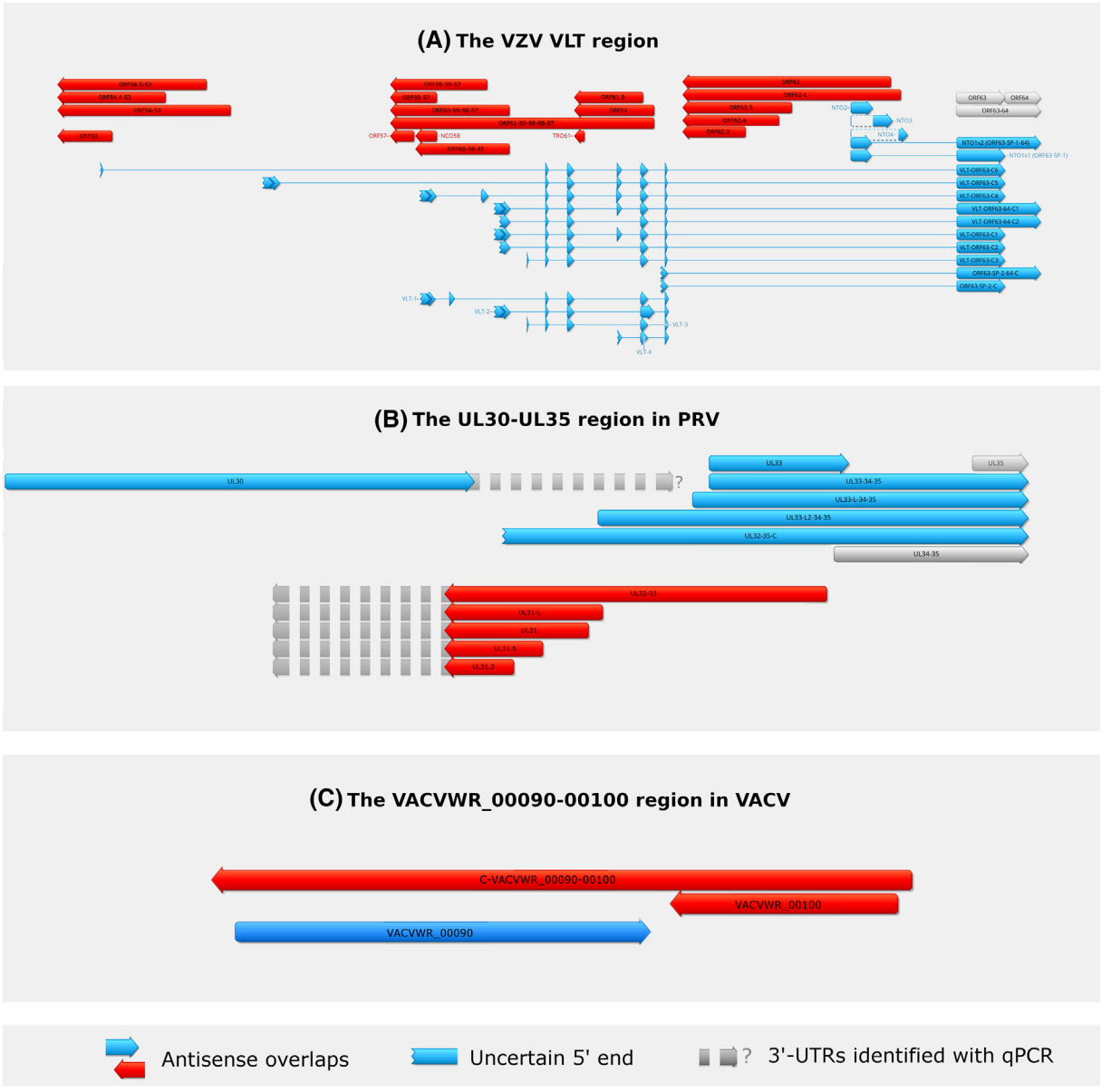


Figure 2. Antisense Transcripts in Two Herpesviruses. (A) The very long splice isoforms of ORF63, ORF63-64, and the VLT_{lyt} RNAs (blue arrows) of varicella-zoster virus overlap several transcripts and transcript isoforms (red arrows) in antisense orientation in an almost 18-kbp long region of the viral genome. The NTO3 and NTO4 antisense transcripts are probably controlled by the same promoter as the NTO1v1 transcript. The unprocessed version of these noncoding transcripts overlaps with several mRNAs in a convergent or divergent manner, thereby generating antisense parts of the RNA molecules. The arrows illustrate exons, while the lines between the arrows are introns. (B) The various RNAs and transcript isoforms produced from the *ul30-35* genomic region of pseudorabies virus (PRV) overlap each other either in convergent (e.g., *ul30* and *ul31*) or in divergent (e.g., *ul32-31* and *ul33*) manners, thereby producing antisense sequences on the RNA molecules. The longer 3′ UTR versions of some transcripts detected by qPCR result in more extended antisense overlaps (gray broken arrows). (C) Convergent overlapping antisense transcripts of the vaccinia virus transcriptome from the VACVWR-00090-00100 genomic region.

and orthomyxoviruses [36,60,78], but they are common in beta- and gammaherpesviruses, retroviruses, and hepadnaviruses [34,61,67,79], whereas there is no splicing in poxviruses at all [80]. As a result of the application of LRS techniques, the number of new splice sites and splice

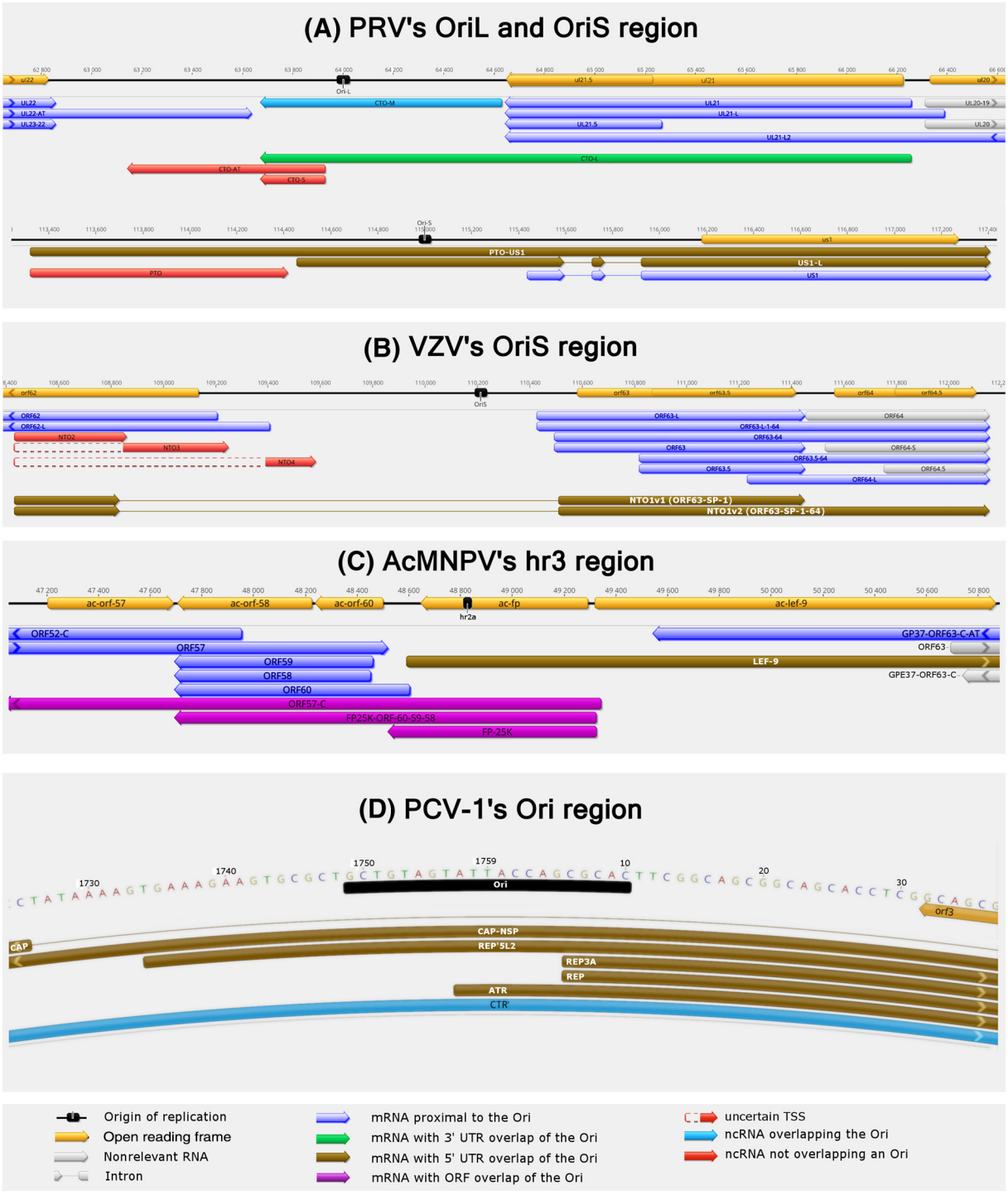


Figure 3. Replication-Associated Transcripts. A large variety of raRNAs have been evolved in various viruses. In herpesviruses, the ncoRNAs can be located near the replication origin, or they can overlap it (panels A and B). These transcripts can be protein-coding (e.g., US1) or noncoding (e.g., cto-s, pto, and NTO2-4), or, alternatively, the noncoding parts of long mRNA isoforms (e.g., PTO-US1, US1-L, and NTO1). The baculovirus hr1 region represents an additional transcript type, which overlaps the Ori with its protein-coding part (ORF; panel C). All transcripts of the porcine circovirus type 1 overlap the viral origin of replication (panel D).

isoforms has been radically increased. Nonetheless, thorough filtering is advised, as template-switching and ligation can both introduce many chimeric cDNA products that may resemble splicing. Common filters require the presence of consensus splice sequences (usually GT/AG, GC/AG, or AT/AC) and/or the absence of short homologous sequences that could facilitate template switching. Even if very strict criteria are used for the identification of splice variants, analysis of individual transcripts by, for example, Northern blot, is needed to confirm their real existence because the RT or the sequencing protocols can produce splice artifacts [47].

Multigenic Transcripts

Multigenic transcripts include polycistronic and complex RNA molecules. Polycistronic transcripts contain two or more genes arranged in tandem array, whereas, in complex transcripts, at least two genes stand in an opposite, convergent or divergent, orientation. Polycistronism is common in viruses; however, the translation of internal genes from these RNA molecules is very rare in large DNA viruses. Two well-described exceptions are the translation of the ORF72-71 and the ORF35-36-37 transcripts of Kaposi sarcoma-associated herpesvirus (KSHV), where the expression of the downstream genes is facilitated by an IRES sequence or an upstream (u)ORF, respectively [81,82]. What could be the function of these RNA molecules, if not translation? LRS characterizes transcriptome diversity with a much higher sensitivity than is possible by SRS. Therefore, LRS studies of viruses with complex transcriptomes, such as DNA viruses, and some RNA viruses, will deepen our understanding of the general aspects of genetic regulation.

Transcriptional Overlaps

Transcripts can overlap with each other in a convergent (tail-to-tail), divergent (head-to-head), or a parallel (tail-to-head) manner (Figure 4). In many cases, TRTs produce overlapping transcripts ('soft'/alternative overlaps), but TOs can also be produced without readthrough ('hard' TOs), for example, in *ul30-31* genes of alphaherpesviruses (Figure 2) [33–37]. This gene pair also produces longer TES variants with uncertain termination. The alternative promoter usage can also produce divergent soft TOs, but in, for example, PRV, most divergent genes overlap each other in a hard manner. This is also the case in the tandem gene clusters of herpesviruses in which the adjacent genes overlap in a tail-to-head manner. LRS studies revealed an intricate meshwork of TOs in every examined virus family [34,36,37,60,62]. The question can be raised whether this TO complexity is functional, or not, and if it is, what could this function be.

Transcriptional Interference Networks

It has been earlier proposed that the role of TOs is to carry out transcription-level interference (TI) between the herpesvirus genes in order to control gene expression [83]. It is supposed that TIs are organized into a system (transcription interference network, TIN) that may coordinate the viral life cycle in a spatiotemporal manner through the physical interaction of the transcriptional apparatuses. transcriptional interference network (TIN) might represent an additional level of gene regulation, which could have coevolved with the transcription factor-dependent regulation of gene expression [84]. TI is considered to be a system-level property because every viral gene produces transcripts that overlap with other transcripts in a variety of ways; therefore, the effect of a change in the expression of a gene can spread throughout the entire genome. It is hypothesized that these interactions form a self-regulatory network and result in a strictly timed alteration of the ON/OFF states of genes along the whole viral DNA. TIN is supposed to coregulate the closely spaced genes through synchronization and/or negative synchronization of the transcriptions, thereby resulting in a well-defined temporal pattern of genome-wide gene expressions. TIN may also be able to reduce the transcriptional noise, that is, it cooperates with the transcription factor-based system for suppressing the expression of genes whose products are not needed at a given stage of the viral life cycle.

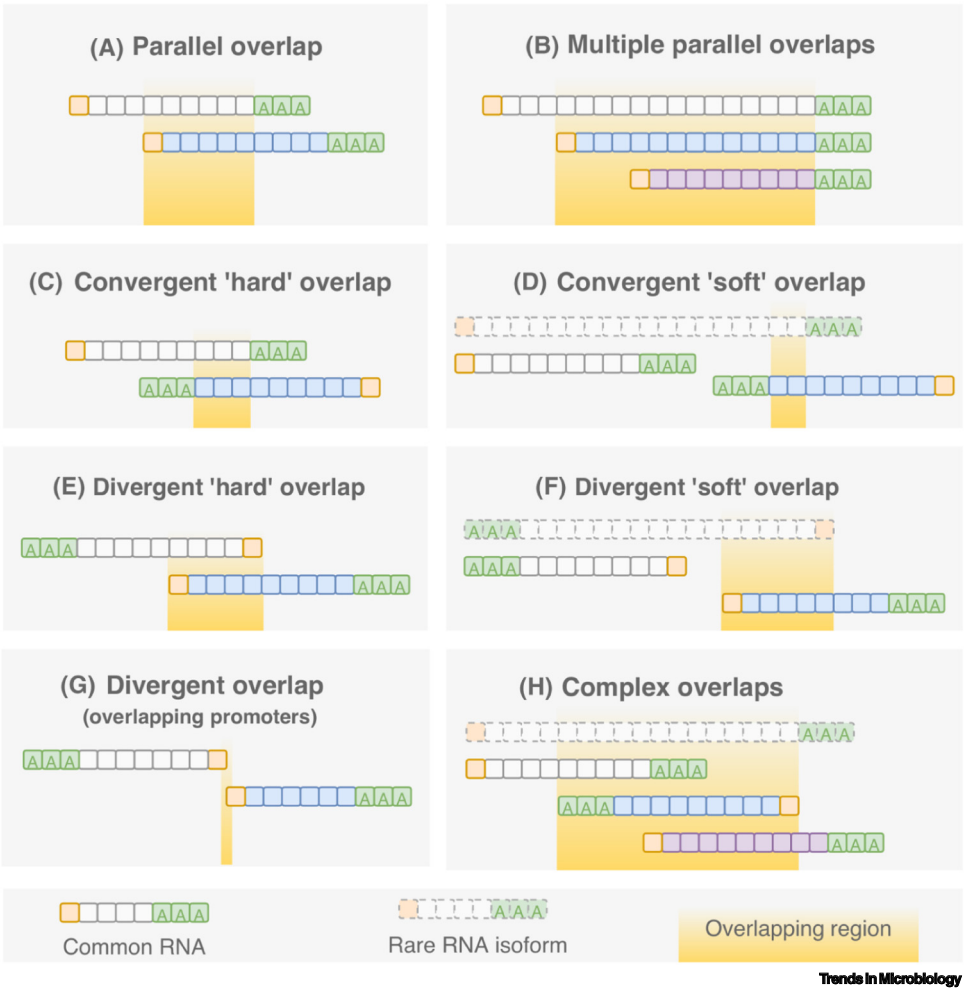


Figure 4. Transcript Overlaps. Viral RNA molecules can form various types of overlap with respect to orientation and length of transcripts. The prototypic organization of the transcriptome of herpesviruses and baculoviruses is that tandem genes express overlapping transcripts with common 3′-termini. The transcripts with 'soft' overlaps have shorter non-overlapping variants. Complex overlaps span at least two complete genes.

Transcriptional and Replication Interference Networks

The expression of raRNAs is supposed to facilitate the regulation of replication initiation and the orientation of replication fork progression through either a collision between the replication and transcription machineries, or by unwinding the DNA strands by the RNP near the Ori. Since the identified nroRNAs are all polyadenylated, they are likely to have a function as RNA molecules beside being the byproducts of a regulatory mechanism. Moreover, an overall decrease in the transcriptional activity in individual herpesvirus genes has been observed following the onset of DNA replication [85], which also suggests an interplay between the two machineries. The interactions between the RNA and DNA synthesis apparatus have been supposed to form a transcription and replication interference network (TRIN) that controls global gene expression and replication in a cooperative manner [64].

Transcriptional overlaps and the overlaps of raRNAs with the replication origins may represent a novel level of genetic regulation. Nevertheless, further investigations are needed to elucidate the role of these RNA molecules.

Upstream ORFs

Ribosome footprint analysis has detected hundreds of translationally active short uORFs in the human cytomegalovirus (HCMV) and KSHV genomes [54,55]. It has also been shown that transcript isoforms differ from each other in the presence or absence of uORFs [35,37]. Some of these short peptides may be functional; it is more likely, however, that their main role is the regulation of translation reinitiation at downstream ORFs. The uORFs permit translation reinitiation of a downstream gene because they are very short, and thus the initiation factors have not yet dissociated. When translation is initiated at an uORF upstream of the HCMV UL4 gene, the mRNA structure stalls the ribosome and it completely inhibits the downstream translation from that RNA molecule [86]. The translation of KSHV gene ORF36 has been shown to be regulated by uORFs upstream of the ORF35 coding sequence through a termination–reinitiation mechanism [81]. This herpesvirus has reversed the host strategy to repress translation by uORFs: it allows the expression of a downstream gene. It has also been shown that RNA isoforms can increase the coding capacity of HCMV genes due to the alternative presence or absence of uORFs [35,37].

Detection of RNA Editing and Modification

G mismatches of the mapped reads can be caused either by natural variation of the viral population or by A-to-I RNA hyperediting. The two cases can be easily distinguished by using LRS techniques when the edited RNA overlaps other transcripts that are unmodified. LRS has been used to discover a hyperediting event on a varicella-zoster virus ncRNA, the NTO3 [65]. Several types of RNA modification have been described in viral transcriptomes with functions including mRNA maturation [87] and evasion of the host's immune response [88,89]. The ability to sequence full-length native RNA molecules simplifies the problem of assorting the detected modified nucleotides among the overlapping transcripts.

Concluding Remarks

In the past few years, LRS techniques have become essential in genome and transcriptome research, and their application is expected to be dramatically increased in the near future. The current LRS techniques produce a lower sequencing coverage than the SRS approaches; therefore, small-genome organisms are ideal subjects for these third-generation sequencing techniques. LRS platforms are able to determine full-length transcripts; thus, they are superior for identifying long, multigenic and multispliced transcripts as well as TSS and TES isoforms. TOs are also easier to study with these techniques. All in all, LRS approaches have multiplied the number of transcripts in every examined viral species. Considering that many viral transcriptomes are poorly annotated, LRS may greatly increase our understanding of the gene expression of most viruses with complex transcriptomes. Direct RNA sequencing is also useful for the examination of viruses with simple transcriptomes [e.g., (+)ssRNA viruses] as it can detect RNA modifications which would not be detectable by other methods. The extremely complex network of TOs suggests a sophisticated interplay between the adjacent and distal genes through physical interaction between the transcriptional apparatuses. Additionally, the discovery of nroRNAs and nro-like transcripts raises the possibility of the interaction between the replication and transcription machineries in order to regulate gene expression and DNA synthesis in a cooperative manner (see Outstanding Questions).

LRS technology is versatile and is rapidly evolving. Its capability of distinguishing between RNA isoforms by reading full-length RNAs and detecting RNA modifications makes it a competent tool for viral transcriptomics.

Outstanding Questions

How could LRS be better exploited for the identification of novel transcripts and transcript isoforms, and how could the potential artefacts of these techniques be eliminated?

While ONT sequencing is superior to the PacBio sequencing in terms of cost-effectiveness, in throughput, and in read length, it possesses a major disadvantage regarding a high error rate. Will ONT be able to solve this problem in the future? Alternatively, will PacBio be able to keep pace with ONT in continuing to improve the above parameters?

10x Genomics has developed an add-on for the short-read sequencing-based Illumina system, which provides long-range genome information. Will this technology outcompete the PacBio and ONT from the market?

The function of most viral noncoding transcripts remains unknown. How can these functions be characterized and harnessed for controlling the virulence of the various viruses?

Can upstream ORFs be utilized for the control of viral gene expression?

What could be the function of RNA editing in some noncoding transcripts?

Do the extensive transcriptional overlaps represent a mere economization of the small viral genomes, or do they represent a novel regulatory layer based on the interference between the transcriptional machineries of adjacent and distal genes?

What could be the function of the 'near replication origin' transcripts? Do they regulate the replication on a collision-based manner? Does the replication regulate the transcription?

Acknowledgments

Z.Bo. was supported by the National Research, Development and Innovation Office (NKFIH) K 128247 grant, M.S. by the National Institutes of Health, Centers of Excellence in Genomic Science Center for Personal Dynamic Regulomes (No. 5P50HG00773502), and D.T. by the National Research, Development and Innovation Office (NKFIH) FK 128252 grant.

Resources

[i]www.pacb.com/wp-content/uploads/Clark-PAG-2017-Full-Length-cDNA-Sequencing-on-the-PacBio-Sequel_Platform.pdf
[ii]https://github.com/nanoporetech/pinfish
[iii]https://github.com/zsolt-balazs/LoRTIA

References

1. Shine J. and Dalgarno L. (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature* 254, 34–38.
2. Firth A.E. and Brierley I. (2012) Non-canonical translation in RNA viruses. *J. Gen. Virol.* 93, 1385–1409.
3. Yuan C. et al. (2017) It is imperative to establish a pellucid definition of chimeric RNA and to clear up a lot of confusion in the relevant research. *Int. J. Mol. Sci.* 18, E714.
4. Weimer B.C. (2017) 100K Pathogen genome project. *Genome Announc.* 5, e00594-17.
5. Turnbull C. et al. (2018) The 100000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* k1687, 361.
6. Wang E.T. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.
7. Goodwin S. et al. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351.
8. McCarthy A. (2010) Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. *Chem. Biol.* 17, 675–676.
9. Ip C.L.C. et al. (2015) MinION Analysis and Reference Consortium: phase 1 data release and analysis. *F1000Research* 4, 1075.
10. Garalde D.R. et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206.
11. Ozsolak F. et al. (2009) Direct RNA sequencing. *Nature* 461, 814–818.
12. Rhoads A. and Au K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289.
13. Quick J. et al. (2014) A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* 3, 22 (2047-217X-3-22).
14. Steijger T. et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184.
15. Chin C.-S. et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569.
16. Pendleton M. et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* 12, 780–786.
17. Szűcs A. et al. (2017) Long-read sequencing reveals a GC pressure during the evolution of porcine endogenous retrovirus. *Genome Announc.* 5, e01040-17.
18. Tombácz D. et al. (2014) Strain Kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. *Genome Announc.* 2, e00628-14.
19. Nakano K. et al. (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* 30, 149–161.
20. Dilernia D.A. et al. (2015) Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Res.* 43, e129.
21. Bull R.A. et al. (2016) A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* 17, 247.
22. Prazsák I. et al. (2018) Full genome sequence of the Western Reserve strain of vaccinia virus determined by third-generation sequencing. *Genome Announc.* 6, e01570-17.
23. Mankertz J. et al. (1998) Transcription analysis of porcine circovirus (PCV). *Virus Genes* 16, 267–276.
24. Farrell M.J. et al. (1991) Herpes simplex virus latency-associated transcript is a stable intron. *Proc. Natl. Acad. Sci. U. S. A.* 88, 790–794.
25. Nagel M.A. et al. (2011) Varicella-zoster virus transcriptome in latently infected human ganglia. *J. Virol.* 85, 2276–2287.
26. Tombácz D. et al. (2009) Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genomics* 10, 491.
27. Sadler R.H. and Raab-Traub N. (1995) Structural analyses of the Epstein–Barr virus BamHI A transcripts. *J. Virol.* 69, 1132–1141.
28. Aguilar J.S. et al. (2006) Quantitative comparison of the HSV-1 and HSV-2 transcriptomes using DNA microarray analysis. *Virology* 348, 233–241.
29. Lacaze P. et al. (2011) Temporal profiling of the coding and noncoding murine cytomegalovirus transcriptomes. *J. Virol.* 85, 6065–6076.
30. Oláh P. et al. (2015) Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol.* 15, 130.
31. O'Grady T. et al. (2014) Global bidirectional transcription of the Epstein–Barr virus genome during reactivation. *J. Virol.* 88, 1604–1616.
32. Chen Y.-R. et al. (2013) The transcriptome of the baculovirus *Autographa californica* multiple nucleopolyhedrovirus in *Trichoplusia ni* cells. *J. Virol.* 87, 6391–6405.
33. Tombácz D. et al. (2016) Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 11, e0162868.
34. O'Grady T. et al. (2016) Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* 44, e145.
35. Moldován N. et al. (2017) Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* 8, 2708.
36. Tombácz D. et al. (2017) Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* 8, 1079.
37. Balázs Z. et al. (2017) Long-read sequencing of human cytomegalovirus transcriptome reveals RNA isoforms carrying distinct coding potentials. *Sci. Rep.* 7, 15989.
38. Levene M.J. et al. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682–686.
39. Eid J. et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.
40. Lin H.-H. and Liao Y.-C. (2015) Evaluation and validation of assembling corrected PacBio long reads for microbial genome completion via hybrid approaches. *PLoS One* 10, e0144305.
41. Schirmer M. et al. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinf.* 17, 125.
42. Weirather J.L. et al. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100.
43. Hebert P.D.N. et al. (2018) A sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19, 219.
44. Workman R.E. et al. (2018) Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv* Published online November 9, 2018. https://doi.org/10.1101/459529.

45. Jain M. et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345.
46. Lu H. et al. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14, 265–279.
47. Cocquet J. et al. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88, 127–131.
48. Kebschull J.M. and Zador A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 43, e143.
49. Zhu Y.Y. et al. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–897.
50. Edery I. et al. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.* 15, 3363–3371.
51. Tilgner H. et al. (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742.
52. Depledge D.P. et al. (2019) Going the distance: optimizing RNA-Seq strategies for transcriptomic analysis of complex viral genomes. *J. Virol.* 93, e01342-18.
53. Park P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680.
54. Stern-Ginossar N. et al. (2012) Decoding human cytomegalovirus. *Science* 338, 1088–1093.
55. Arias C. et al. (2014) KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. *PLoS Pathog.* 10, e1003847.
56. Mahat D.B. et al. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* 11, 1455–1476.
57. Zheng D. et al. (2016) 3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA. *RNA* 22, 1631–1639.
58. Tombácz D. et al. (2017) Characterization of the dynamic transcriptome of a herpesvirus with long-read single molecule real-time sequencing. *Sci. Rep.* 7, 43751.
59. Gupta I. et al. (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* 36, 1197–1202.
60. Moldován N. et al. (2018) Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.* 8, 8604.
61. Moldován N. et al. (2018) Multiplatform next-generation sequencing identifies novel RNA molecules and transcript isoforms of the endogenous retrovirus isolated from cultured cells. *FEMS Microbiol. Lett.* 365 (5), fny013.
62. Moldován N. et al. (2017) Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res.* 237, 37–46.
63. Tombácz D. et al. (2018) Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* 7, giy139.
64. Tombácz D. et al. (2015) Characterization of novel transcripts in pseudorabies virus. *Viruses* 7, 2727–2744.
65. Prazsák I. et al. (2018) Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* 19, 873.
66. Palazzo A.F. and Lee E.S. (2015) Non-coding RNA: what is functional and what is junk? *Front. Genet.* 6, 2.
67. Gatherer D. et al. (2011) High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19755–19760.
68. Pal S. et al. (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* 21, 1260–1272.
69. Isomura H. et al. (2008) Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *J. Virol.* 82, 1638–1646.
70. Leppek K. et al. (2018) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* 19, 158–174.
71. Mainguy G. et al. (2007) Extensive polycistronism and antisense transcription in the mammalian Hox clusters. *PLoS One* 2, e356.
72. Gullerova M. and Proudfoot N.J. (2008) Cohesin complex promotes transcriptional termination between convergent genes in *S. pombe*. *Cell* 132, 983–995.
73. Majerciak V. et al. (2013) A viral genome landscape of RNA polyadenylation from KSHV latent to lytic infection. *PLoS Pathog.* 9, e1003749.
74. Nam D.K. et al. (2002) Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6152–6156.
75. Abdel-Ghany S.E. et al. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7, 11706.
76. Gonzalez-Garay M.L. (2016) *Introduction to Isoform Sequencing Using Pacific Biosciences Technology* (*Iso-Seq*). Springer, pp. 141–160.
77. Wang B. et al. (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708.
78. Shih S.R. et al. (1995) The choice of alternative 5′ splice sites in influenza virus M1 mRNA is regulated by the viral polymerase complex. *Proc. Natl. Acad. Sci. U. S. A.* 92, 6324–6328.
79. Sommer G. and Heise T. (2008) Posttranscriptional control of HBV gene expression. *Front. Biosci.* (13), 5533–5547.
80. Moss B. (1991) Vaccinia virus: a tool for research and vaccine development. *Science* 252, 1662–1667.
81. Kronstad L.M. et al. (2013) Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. *PLoS Pathog.* 9, e1003156.
82. Low W. et al. (2001) Internal ribosome entry site regulates translation of Kaposi's sarcoma-associated herpesvirus FLICE inhibitory protein. *J. Virol.* 75, 2938–2945.
83. Boldogköi Z. (2012) Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* 3, 122.
84. Nasser W. et al. (2002) Transcriptional regulation of fis operon involves a module of multiple coupled promoters. *EMBO J.* 21, 715–724.
85. Takács I.F. et al. (2013) The ICP22 protein selectively modifies the transcription of different kinetic classes of pseudorabies virus genes. *BMC Mol. Biol.* 14, 2.
86. Geballe A.P. and Mocarski E.S. (1988) Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. *J. Virol.* 62, 3334–3340.
87. Sommer S. et al. (1976) The methylation of adenovirus-specific nuclear and cytoplasmic RNA. *Nucleic Acids Res.* 3, 749–765.
88. Anderson B.R. et al. (2011) Nucleoside modifications in RNA limit activation of 2′-5′-oligoadenylate synthetase and increase resistance to cleavage by RNase L. *Nucleic Acids Res.* 39, 9329–9338.
89. Karikó K. et al. (2005) Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* 23, 165–175.
90. Workman R.E. et al. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *Gigascience* 7, giy009.
91. Madoui M.-A. et al. (2015) Genome assembly using nanopore-guided long and error-free DNA reads. *BMC Genomics* 16, 327.
92. Salmela L. and Rivals E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514.
93. Koren S. et al. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736.
94. Lima L.I.S. de et al. (2018) Comparative assessment of long-read error-correction software applied to RNA-sequencing data. *bioRxiv* Published online November 23, 2018. https://doi.org/10.1101/476622.

95. Volden R. et al. (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9726–9731.
96. Li H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100.
97. Byrne A. et al. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027.
98. Tang A.D. et al. (2018) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* Published online September 6, 2018. https://doi.org/10.1101/410183.
99. Tardaguila M. et al. (2017) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–441 Published online August 21, 2017. https://doi.org/10.1101/gr.222976.117.