frontiers in Genetics

REVIEW
published: 16 August 2019
doi: 10.3389/fgene.2019.00709

Getting the Entire Message: Progress in Isoform Sequencing

*Simon A. Hardwick[1,2], Anoushka Joglekar[1], Paul Flicek[3], Adam Frankish[3] and Hagen U. Tilgner[1]**

[1] Brain and Mind Research Institute, Weill Cornell Medicine, NY, United States, [2] Garvan Institute of Medical Research, Sydney, NSW, Australia, [3] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom

The advent of second-generation sequencing and its application to RNA sequencing have revolutionized the field of genomics by allowing quantification of gene expression, as well as the definition of transcription start/end sites, exons, splice sites and RNA editing sites. However, due to the sequencing of fragments of cDNAs, these methods have not given a reliable picture of complete RNA isoforms. Third-generation sequencing has filled this gap and allows end-to-end sequencing of entire RNA/cDNA molecules. This approach to transcriptomics has been a "niche" technology for a couple of years but now is becoming mainstream with many different applications. Here, we review the background and progress made to date in this rapidly growing field. We start by reviewing the progressive realization that alternative splicing is omnipresent. We then focus on long-noncoding RNA isoforms and the distinct combination patterns of exons in noncoding and coding genes. We consider the implications of the recent technologies of direct RNA sequencing and single-cell isoform RNA sequencing. Finally, we discuss the parameters that define the success of long-read RNA sequencing experiments and strategies commonly used to make the most of such data.

Keywords: RNA, isoforms, long-read, splicing, epitranscriptome

OPEN ACCESS

Edited by:
Andrew J. Mungall, Canada's Michael Smith Genome Sciences Centre, Canada

Reviewed by:
Wei Xu, Texas A&M University Corpus Christi, United States
Jiannis Ragoussis, McGill University, Canada

****Correspondence:***
Hagen U. Tilgner
hut2006@med.cornell.edu

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *31 January 2019*
Accepted: *04 July 2019*
Published: *16 August 2019*

Citation:
Hardwick SA, Joglekar A, Flicek P, Frankish A and Tilgner HU (2019) Getting the Entire Message: Progress in Isoform Sequencing. Front. Genet. 10:709. doi: 10.3389/fgene.2019.00709

AN ABUNDANCE OF ALTERNATIVE RNA PROCESSING EVENTS

The first decade of the new millennium has made it abundantly clear that most genes produce multiple distinct isoforms: Estimates of the fraction of multi-exon genes that are alternatively spliced rose from 42% in 2001 (Modrek et al., 2001) to 74% in 2003 (Johnson et al., 2003), to 86% in 2006 (Harrow et al., 2006). The new technology of RNA-sequencing (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Pan et al., 2008; Sultan et al., 2008; Wang et al., 2008; Wilhelm et al., 2008) and its application to alternative splicing finally pushed this estimation to 95–98% (Pan et al., 2008) and 98–100% (Wang et al., 2008) in 2008. Simultaneously, the RNA community has established the existence of more than 2 million RNA editing sites (Ramaswami and Li, 2016), and that the number of transcription start sites (TSS) outnumbers by an order of magnitude the number of genes (Forrest et al., 2014), implying widespread alternative TSS usage. Also, polyA-site estimates are on the rise, with more than half of all genes now known to have alternative polyA-sites (Tian et al., 2007; Sandberg et al., 2008; Mayr, 2016). Taken together, these observations reveal a vast abundance of alternative processing events that can affect RNA molecules. Beyond the four-letter sequence of RNA molecules, chemical modifications on RNA nucleotides, collectively referred to as the "epitranscriptome" (Meyer et al., 2012) introduce further variables sites on transcripts (Dominissini

et al., 2012; Meyer et al., 2012; Schwartz et al., 2013), which usually are not represented in full-length cDNA sequences. Over 100 different types of RNA modifications have been identified to date, and these have been shown to be involved in nearly every aspect of the mRNA life cycle (Roundtree et al., 2017). For many of these alternative site, functions are known, while for others function remains elusive.

This abundance of alternative sites and events raises a number of key questions, for many (but not all) of which, sequencing of full-length isoforms is giving and is expected to yield significant insights. 1) Which combinations of the previously mentioned variable sites are actually being generated as RNA isoforms? In theory, all these alternative sites can specify an exponential number of distinct RNA molecules by exploiting distinct combinations of the previously discussed sites; however, recent data suggest that for many (but certainly not all) genes, this is not the case. 2) Can we find the precise cell types that generate each isoform? As we will see later, single-cell approaches are beginning to offer a window into this field. 3) What is the relative timing of multiple alternative processing events within a gene? 4) With multiple long-read sequencing approaches now available, we must ask to which extent these may give different answers. This review will focus on the contributions made by long-read RNA sequencing to date and will also include a discussion of the technical challenges that have been overcome and that will need to be overcome in the future.

FUNCTIONALITY OF ALTERNATIVE RNA PROCESSING EVENTS

We will only briefly touch on other important questions, such as "Which isoforms harbor function?"—a question whose negative is not easily assessed, given the variety of possible settings in which an event may be relevant. Here, we will limit ourselves to saying that there are many clear examples for the functionality of alternative RNA processing events. This is exemplified by the FAS receptor (Cheng et al., 1994) and the finding that for the majority of tested genes, distinct alternative isoforms differ in their protein interaction partners, once translated (Yang et al., 2016). Detailing all the functional consequences of alternative splicing is beyond the scope of this review, and this topic has been recently reviewed by others (Cieply and Carstens, 2015; Daguenet et al., 2015; Raj and Blencowe, 2015; Vuong et al., 2016; Fiszbein and Kornblihtt, 2017; Gallego-Paez et al., 2017; Mauger and Scheiffele, 2017).

LIMITATIONS OF SHORT-READ RNA SEQUENCING

High-throughput transcriptional profiling ("RNA-seq") was pioneered in 2008, which enabled a transcriptome-wide survey of gene expression and alternative splicing in a quantitative fashion (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Pan et al., 2008; Sultan et al., 2008; Wang et al., 2008; Wilhelm et al., 2008). Despite the success of RNA-seq in greatly expanding our knowledge of the mammalian transcriptome, it relies on short sequencing reads (~100–150 bp), which must be computationally assembled into longer transcript models. This can be a notoriously difficult and error-prone task, particularly when alternative splicing generates multiple partially redundant isoforms at a given locus (Steijger et al., 2013; Tilgner et al., 2013). With saturating coverage, short-read RNA-seq can accurately measure percent spliced-in (PSI) scores for individual exons but cannot unambiguously resolve the connectivity between distant exons because they are never represented on the same sequenced fragment (Tilgner et al., 2015; Tilgner et al., 2018). With the emergence of third-generation sequencing, it is now possible to sequence full-length transcripts "in one go," thereby obviating the challenges posed by computational assembly and delivering reliable isoform structures.

A BRIEF HISTORY OF THIRD-GENERATION ISOFORM SEQUENCING

In the second decade of the millennium, third-generation sequencing experienced and is experiencing rapid growth (**Figure 1**). This occurred with an abundance of alternative RNA processing events described (see the previous section), which set the stage for questions that could be addressed with long-read isoform sequencing. Note, this review largely ignores long-read

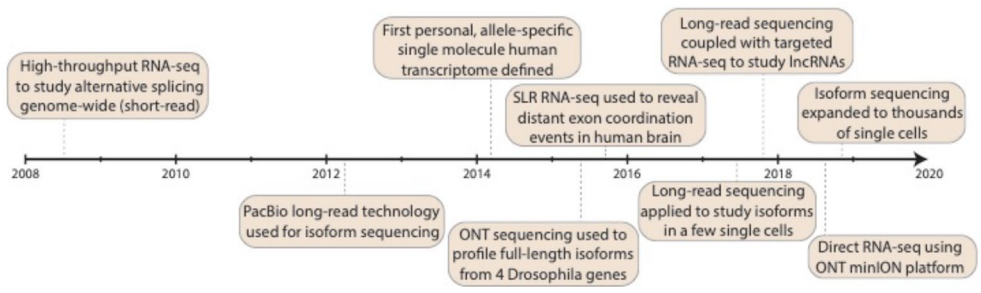

FIGURE 1 | Progress in isoform sequencing. Timeline highlights some of the key milestones in the history of isoform sequencing, dating back to the advent of short-read RNA-seq back in 2008. Note that this is presented as a summary only and is not intended to be exhaustive of all work done in the field. RNA-seq: RNA sequencing; PacBio: Pacific Biosciences; SLR: synthetic long-read; lncRNA: long noncoding RNA; ONT: Oxford Nanopore Technologies.

and linked-read applications to non-transcriptome work, which have been reviewed (focusing on PacBio) in 2015 (Rhoads and Au, 2015) and considering PacBio, nanopore, and linked-read technologies in 2018 (Sedlazeck et al., 2018). The first long-read platform that truly allowed sequencing of full-length isoforms in a single read was the Pacific Biosciences (PacBio) platform (Eid et al., 2009), which started to be used for isoform descriptions in 2012. PacBio sequencing works by utilizing a DNA polymerase that is affixed at the bottom of a zero-mode waveguide (ZMW) with a single molecule of DNA as a template. As with earlier sequencing technologies, each of the four DNA nucleotides is attached to one of four different fluorescent dyes, and nucleotide incorporation is observed in real time. Many ZMWs are incorporated on a single chip, enabling massive parallelization. Koren et al. (2012) investigated the corn transcriptome using methods of error correction (see later). In 2013, Sharon et al. (2013) exploited a panel of human organs, which theoretically harbors large amounts of splice variants, to describe full-length molecules in a PCR-free fashion based on circular consensus sequences (Eid et al., 2009; Travers et al., 2010) (CCS, see later), and Au et al. (2013) described the transcriptome of human embryonic stem cells, again using error correction. In 2014, we (Tilgner et al., 2014) produced an enhanced GENCODE annotation, adding full-length isoforms from lymphoblastoid cells and from a panel of human organs to the GENCODE annotation (Harrow et al., 2006; Harrow et al., 2012). This same year (2014) also saw important work, investigating the connectivity of neurexin exons in a targeted manner (Schreiner et al., 2014; Treutlein et al., 2014). The following year (2015) saw the emergence of the first non-PacBio long-read strategies. Bolisetty et al. (2015) and Roy et al. (2015) pioneered the use of Oxford Nanopore Technologies (ONT) sequencing to study exon connectivity for a set of target genes (Bolisetty et al., 2015; Roy et al., 2015). Nanopore sequencing works by detecting changes in current that occur when a biological molecule (e.g., DNA) passes through a nanoscale pore. These changes in current ("squiggles") are measured and then computationally converted into DNA nucleotides. We (Tilgner et al., 2015) exploited the dilution-based Moleculo approach (McCoy et al., 2014; Voskoboynik et al., 2013) for RNA sequencing to reveal nonrandomly paired alternative exon pairs genome-wide. Likewise, we developed another PacBio competitor—sparse isoform sequencing (SpISO-Seq)—which is based on linked-read sequencing (Zheng et al., 2016) and allows the description of many more millions of RNA molecules (Tilgner et al., 2018). Despite the availability of competitors, PacBio continues to be heavily used and developed for isoform sequencing (Gordon et al., 2015; Weirather et al., 2015; Shi et al., 2016; Tevz et al., 2016; Tombácz et al., 2016; Lagarde et al., 2017; Sahraeian et al., 2017; Tseng et al., 2017; Anvar et al., 2018; Balázs et al., 2018; Deveson et al., 2018; Dougherty et al., 2018; Gupta et al., 2018; Tardaguila et al., 2018; Wyman and Mortazavi, 2019). Both ONT (through its now available PromethION instrument) and PacBio (through an announced 8 million ZMW SMRT cell[1]) are poised for large throughput increases, which could dramatically alter our view of isoform biology. The year 2017 saw the first long-read strategies for a few (~10^1) single cells (Byrne et al., 2017; Karlsson and Linnarsson, 2017), and in 2018, we and others introduced the first applications of long-read technologies to ~10^2 (Volden et al., 2018) and 10^3–10^4 (Gupta et al., 2018) individual cells (see later).

[1] https://www.pacb.com/press_releases/pacific-biosciences-launches-new-sequel-ii-system-featuring-8-times-the-dna-sequencing-data-output/.

CHARACTERIZATION OF lncRNAs AND THEIR BIOLOGY THROUGH ISOFORM SEQUENCING

Due to the relatively shallow sequencing depth provided by third-generation sequencing platforms, the majority of studies to date have focused on protein-coding genes due to their higher overall expression. Early isoform studies with third-generation sequencing studies (or "454" 400–700-bp reads) (Au et al., 2013; Sharon et al., 2013; Tilgner et al., 2013; Tilgner et al., 2014; Tilgner et al., 2015) revealed consistently novel aspects of long noncoding RNA (lncRNA) expression. Using "454" (Tilgner et al., 2013), PacBio (Sharon et al., 2013; Tilgner et al., 2014), and Moleculo (Tilgner et al., 2015) sequencing, we found that 30–40% of all long reads aligned to known GENCODE lncRNA loci were inconsistent with all annotated isoforms for the loci in question. This was dramatically higher than for long reads aligned to protein-coding loci. A simple explanation for these observations appeared to be that lncRNAs had been less comprehensively investigated than protein-coding genes. Therefore, increased novelty rates would be simply a reflection of our more limited (in comparison with protein-coding genes) knowledge of lncRNA biology. Recent research has shed new light on this observation: the Mercer and Mattick laboratories found universal alternative splicing of noncoding exons (Deveson et al., 2018), including those in lncRNAs. That is, unlike protein-coding exons, almost all noncoding exons were found to be alternatively spliced (i.e., had a PSI score < 95%). This suggests that splicing patterns in lncRNAs may not fall under the same level of constraint as those in protein-coding genes; as the requirement to maintain an ORF is not imposed on noncoding RNA, this allows the spliceosome to explore the full range of noncoding exon combinations available. This would in turn explain the much larger fraction of lncRNA long reads that are inconsistent with all annotated isoforms. In summary, lncRNA isoforms appear to exploit all exons as alternative when interrogated in bulk tissue. A key question now is whether the previous observation could change profoundly when considering highly specific cell types or cell populations. In other words, there are two scenarios that could lead to the observation of universal alternative splicing in noncoding exons: 1) Within specific cell populations, splicing of these noncoding exons may be constitutive but different between populations; 2) splicing of these noncoding exons could also be alternative within all cell populations.

Simultaneous with the previous results, the Wong laboratory's hybrid sequencing approach (Au et al., 2013) (i.e., combining both short- and long-read sequencing technologies) revealed 216

novel gene loci, which were unknown to GENCODE (Harrow et al., 2006; Harrow et al., 2012), RefSeq (O'Leary et al., 2016), and the UCSC (Karolchik et al., 2014) annotation. A subset of these were lncRNAs preferentially expressed in pluripotent cell lines (Au et al., 2013), and a further subset of three of such lncRNAs was later shown to play a role in preimplantation embryo development (Durruthy-Durruthy et al., 2015). Earlier work had established a number of characteristic features of intergenic lncRNAs, including (but not limited to) a lower number of exons and a shorter transcript length compared with those of protein-coding transcripts (Derrien et al., 2012). Targeted RNA capture (Clark et al., 2015) in conjunction with PacBio long-read sequencing, however, increased the estimates of average lncRNA transcript length and exon number substantially, although the estimates for protein-coding genes were not entirely matched (Lagarde et al., 2017). This work by the GENCODE consortium vastly improved lncRNA annotations, with the number of lncRNA genes and transcripts now easily outnumbering their protein-coding counterparts. Another way of posing the previous key question for the future is whether specific lncRNA isoforms are characteristic of precise cell populations.

COMBINATION PATTERNS AND TIMING OF MULTIPLE RNA PROCESSING EVENTS

As noted previously, a single gene can harbor multiple distinct alternative exons and other alternative processing events. Let us consider a hypothetical gene with n alternative sites with only two options each, which (for simplicity of the argument) are all used in 50% of the molecules (**Figure 2**). At one extreme, of the spectrum of possible combinations, these n exons (or more general sites) could produce 2^n combinations at relative abundances of $1/(2^n)$ through exhaustive random pairing. Under random pairing, short-read sequencing would be the method of choice because it would yield a usage probability for each variable site at the lowest cost. The frequency of each complete isoform could then be determined by multiplication over the associated probabilities. At the other end of the spectrum, perfectly coordinated exon pairing could result in just two isoforms (e.g., the isoform including all n exons and the one skipping all n exons). In this setting, short-read probabilities of variable sites would be uninformative for complete isoforms. The interest of combinations has been noticed and investigated for a long time (MacLeod et al., 1985; Helfman et al., 1986; Cramer et al., 1997; Fededa et al., 2005), revealing the combination patterns of such alternative processing RNA sites. Thus, Helfman et al. observed nonrandom pairing of an internal exon and a 3' exon (Helfman et al., 1986), Cramer et al. (1997) showed the dependence of inclusion of an internal exon on promoter structure, and Fededa et al. (2005) showed the dependent splicing outcome of two alternative exons in the *fibronectin* gene. Most interestingly, these authors showed that inclusion levels of the upstream alternative exon conditioned inclusion levels of the downstream one but not in the opposite way, which revealed a gene polarity mechanism probably due to changes in RNA polymerase II elongation

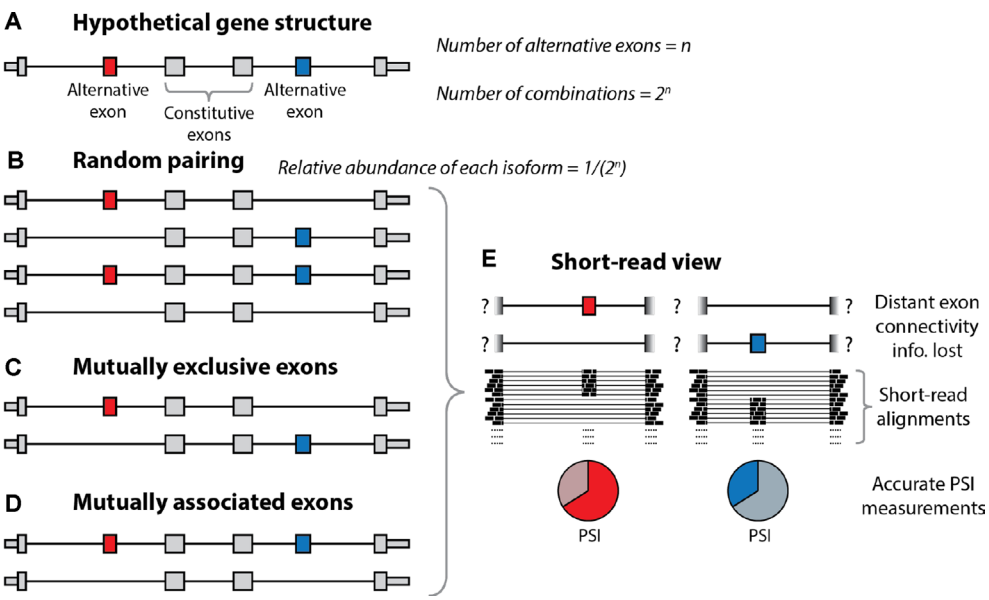

FIGURE 2 | Resolution of alternative splicing events with long-read sequencing. **(A)** Schematic illustration of the structure of a hypothetical gene undergoing alternative splicing. The gene contains two alternatively spliced exons (red and blue) separated by constitutive exons (gray). In theory, if we let *'n'* equal the number of alternative exons, then there are 2^n different combinations of these exons. **(B)** Under random pairing, we would expect to see all of these 2^n combinations, each at a relative abundance of $1/2^n$. In this case, short-read RNA-seq would be sufficient, as it can accurately quantify percent spliced-in (PSI) scores for individual exons. **(C, D)** However, coordinated exon pairing can result in a situation whereby the alternative exons are mutually exclusive **(C)** or mutually associated **(D)**. **(E)** With short-read RNA-seq, these three scenarios are indistinguishable, as the information regarding the connectivity of the alternative exons is lost. Conversely, with long-read sequencing, it is trivial to determine which scenario is present.

rates. At a genome-wide level, Fagnani et al. (2007) revealed correlated exon inclusion using short-reads, but this approach did not allow them to distinguish between two models: A) One type of molecule would include exon 1 and another would include exon 2, with upregulation of both types (at the expense of molecules including both or none of the exons) leading to the observation of correlation. B) Preferential expression of molecules employing either both or none of the exons. In 2014, the neurexin genes were investigated using PacBio by the Sudhof and Scheiffele laboratories (Schreiner et al., 2014; Treutlein et al., 2014), with nonadjacent exons being mostly randomly paired. In 2015, we employed deep long-read sequencing of ~5 million ~2-kb reads (Tilgner et al., 2015) to reveal >100 human genes with coordinated pairs of alternative exons. Consistent with the previously discussed work, neurexins could not be shown to harbor any nonrandom pairing of nonadjacent exons. That same year, the Graveley and Moore labs employed ONT sequencing to establish the connectivity of alternative exons in a mouse gene and the *Drosophila Dscam* gene (Roy et al., 2015) as well as four *Drosophila* genes (Bolisetty et al., 2015). Finally, we recently estimated that 40% of genes with multiple distant alternative splicing events show coordination of these events (Tilgner et al., 2018), and Anvar et al. (2018) revealed thousands of coordination events between exons (including adjacent exon pairs), TSS, and polyA-sites. Interestingly, coordination of distant splicing events in bulk tissue occurs in the presence of different isoform expression between cell types more frequently than coordination of adjacent exons (Gupta et al., 2018). In summary, these combination patterns generally warrant the use of long-read strategies for isoform descriptions.

The combination patterns of n alternative binary processing events can result in 2^n combinations. However, the relative order of n events could in principle be carried out in $n!$ distinct orders, which defines an even greater search space, than the previously mentioned 2^n combinations. As a general background, it is widely appreciated that splicing occurs very frequently co-transcriptionally, that is, while the RNA molecule is still in proximity to the chromatin template (Beyer and Osheim, 1988; Ameur et al., 2011; Khodor et al., 2011; Tilgner et al., 2012; Schor et al., 2013). An important finding in this realm, involving third-generation sequencing, was recently revealed by Carillo Oesterreich and colleagues, who employed PacBio sequencing to track the splicing status of introns in yeast. These authors revealed that once RNA polymerase has transcribed 45 nt of downstream DNA, half of the preceding introns have undergone splicing in yeast (Carrillo Oesterreich et al., 2016), implying very fast intron removal. The same laboratory more recently aimed at investigating the order of intron removal in the fission yeast *Schizosaccharomyces pombe*, revealing most multi-intron transcripts to be fully spliced or fully unspliced (Herzel et al., 2018).

DIRECT RNA SEQUENCING

Until recently, high-throughput RNA-seq assays have relied on an initial step in which the RNA is first converted to cDNA before sequencing. Thus, these methods detect the products of a synthesis reaction rather than directly reading the RNA molecule itself. Crucially, any RNA modifications are lost in the process of cDNA conversion. While the first direct RNA sequencing method was published almost a decade ago—the Helicos platform (Ozsolak et al., 2009)—this method relied on short sequence reads. Long-read direct RNA sequencing provides a framework in which TSSs, splice sites, polyA-sites, RNA-editing, as well as a number of RNA modifications, whose positions are lost during reverse transcription, can theoretically be interrogated simultaneously on single molecules. This can advance the identification of single sites but, above all, can also reveal the combination patterns of all these different alterations. Recently, ONT has provided proof of concept of direct RNA sequencing in yeast (Garalde et al., 2018), showing that the MinION platform can detect all the previously mentioned variables that define the sequence of an RNA molecule. The Nanopore WGS Consortium (Workman et al., 2018) has recently extended this technique to directly sequence a human polyA transcriptome, impressively generating ~10 million aligned sequence reads that were filtered into ~78,000 high-confidence isoforms (the majority of which contained novel splice junctions missing from GENCODE). However, for the moment, it appears difficult to define all variable sites accurately, based on a single read only.

ISOFORM SEQUENCING FROM SINGLE CELLS

Short-read single-cell splicing studies had revealed the existence of bimodality for percent spliced-in (PSI) distributions across individuals cells (Shalek et al., 2013). That is, individual cells of similar type could differ drastically in their inclusion of a specific exon. More recent work (Song et al., 2017) showed that 20% of alternative exons show this phenotype of bimodality. The advent of long-read third-generation sequencing made it only natural to wonder if full-length isoforms could be profiled from individual cells. Thus, Karlsson and Linnarsson (2017) employed PacBio sequencing to monitor isoforms in six individual mouse brain cells (one vascular, one leptomeningeal, and four oligodendrocyte type cells in different maturation stages) and revealed strong isoform diversity within single cells. The Vollmers lab (Byrne et al., 2017) used and benchmarked the ONT system on seven individual B-cells and found widespread usage of novel TSSs and transcription end sites (TESs), as well as 100–1,000 alternative splicing events.

In 2018, the same lab extended the single-cell long-read view to 96 cells (Volden et al., 2018), also describing a CCS-like method for nanopore sequencing (see previous discussion). Still in 2018, our laboratory described singe-cell isoform RNA sequencing for 5,000–10,000 cells (Gupta et al., 2018), which produces complete cDNAs tagged for their cell of origin (here by using 10x Genomics) (Zheng et al., 2017), and PacBio or ONT to produce full-length isoforms. By identifying barcodes in each long read, one can assign each read to its cell of origin. The advantage of this last approach is that the number of cells (>5,000) allows clustering of cells into cell types and, therefore,

an isoform description of all (sufficiently abundant) cell types in a bulk sample. This technology enables a wealth of applications: First, it allows the tracing of the effect of single nucleotide polymorphisms (SNPs) (or germline mutations) into distinct cell types. Interestingly, such sequence alterations are, in principle, present in every single cell and cell type and may affect genes that are expressed across multiple cell types. By sequencing isoforms of thousands of single cells, we may be able to understand to which extent the action of such SNPs differs across cell types or single cells. Second, in case–control settings of diseases, we may be able to trace the consequences of disease-causing genome alterations or environmental factors into specific cell types—which may pave the way for devising strategies that "correct" isoform regulation in a cell-type specific way.

PARAMETERS OF ISOFORM SEQUENCING

The advantages and disadvantages of different long-read sequencing strategies can be summarized using several criteria (summarized in **Supplementary Table 1**). In this review, we will not attempt to mathematically define these but rather to explain the intuition behind them. These measurements include "completeness of reads," "correctness of sequence," "bias of representation," "sequencing depth," and the "minimal input amount."

1. *Completeness of reads:* Completeness of reads describes the extent to which a long read represents the entire underlying RNA molecule. An interesting twist to this question is that a long read may represent a complete RNA molecule (which was turned into cDNA) but not a complete transcript, as the RNA molecule may have suffered damage in the cell or during the experiment. Whether a read is complete at its 3' end is, in theory, relatively easily assessed by considering its polyA-tail. Of note, a cDNA molecule that is generated through reverse transcription with a polydT primer must contain a polyT (or polyA depending on its orientation) region at its end. This is the case even if the polydT primer annealed to a non-perfect genomic A-rich region because the sequence in the cDNA is determined by the primer, not by the transcript's region that the primer bound to. Given these observations, it was a surprise that we initially, using a hidden Markov model, only found 67% of PacBio CCSs to contain a polyA-tail (Sharon et al., 2013). Broadly consistently, we recently found 61.4% of single-cell long reads to contain a polyA-tail, and Lagarde et al. (2017) report 73% (human) and 64% (mouse) of all reads of insert to yield an identifiable polyadenylation site. Given the previously discussed considerations, it is likely that the missing polyA-tails are lost during CCS generation or possibly earlier in the experiment. A measure of completeness that applies to both 5' and 3' end of reads can be obtained through the comparison with annotated transcripts. This measure is, however, intrinsically subject to the completeness and correctness of the employed annotation (Sharon et al., 2013; Tilgner et al., 2014; Tilgner et al., 2015; Uszczynska-Ratajczak et al., 2018).

2. *Correctness of sequence:* PacBio and ONT raw reads have much higher per base error rates than Illumina sequencing. Linked-read-based methods exploit the repetitive sequencing of individual cDNA molecules to reach quality comparable (and superior) with Illumina short reads (Tilgner et al., 2015; Tilgner et al., 2018). However, there is far less support software available. For PacBio and ONT, the error rates in raw reads have ranged from 10 to 20% but are subject to change in the future. There are currently three approaches to limit the consequences of these error rates. A) The first relies on building CCSs of lower error rates from multiple low-quality read outs of the same molecule. This has been pioneered by PacBio (Eid et al., 2009; Travers et al., 2010) and has been widely used ever since. The advantage of this approach is that all the information in an individual CCS originates from one original RNA molecule, with the disadvantage being that reads shorter than the molecule of interest cannot generate such CCS. While ignoring such reads may introduce a bias against long molecules, the ever-increasing read length of PacBio is likely to increase CCS numbers. For PacBio, the CCS approach has gone through rounds of optimization. For ONT, a recent report (Li et al., 2016) engineered a CCS-like strategy: circularization of molecules and rolling-circle amplification generated long molecules, which repeatedly contain the molecule of interest. Sequencing of this repeat allowed the construction of consensus reads, similar to PacBio CCS. Volden et al. (2018) applied this approach to mRNA, obtaining an accuracy of 94%. This is considerably higher than raw ONT accuracy but still lags behind PacBio CCS accuracy. Possibly, algorithmic improvements to the method could raise the 94% accuracy, although the nonrandom nature of ONT errors could limit such improvements. B) The second approach employs higher-quality short Illumina reads to correct errors in higher-quality long reads. This method was first used in 2012 (Au et al., 2012; Koren et al., 2012) and is also employed in recent software, including LoRDEC (Salmela and Rivals, 2014) and *proovread* (Hackl et al., 2014). This approach has the advantage of rescuing many long reads that cannot form consensus and that would otherwise be lost, with the disadvantage being that the resulting corrected long read is a hybrid of multiple distinct molecules that may not harbor identical sequence. A relatively recent publication has compared the effects of error correction on PacBio and ONT reads (Weirather et al., 2017), although the employed data predate both the PacBio Sequel and the ONT PromethION. C) Last but not least, consensus and error correction can also be achieved from multiple long reads after grouping similar long reads. This simplifies experimental procedures, as only one sequencing experiment has to be performed. However, if systematic biases are present in the original reads, these could persist in the final consensus. Relevant tools include Tofu (Gordon et al., 2015), TAPIS (Abdel-Ghany et al., 2016), and CARNAC-LR (Marchet et al., 2019).

3. *Bias of representation:* Different cDNA molecules can differ in a variety of characteristics, including length, sequence (often summarized as GC) content, structure, to name only a few. Looking at the bias of representation fundamentally asks whether the molecules that are presented to the machine differ significantly in any of the previously discussed characteristics from those that are reported as long reads. On the PacBio machine, there is little to no bias of coverage in GC-rich region (Ferrarini et al., 2013); however, there is a bias for shorter molecules. This length bias has

been counteracted by sequencing distinct size selections (Chin et al., 2013), which ensures that larger molecules are not lost due to preferential sequencing of shorter molecules. The ONT system was recently tested (Oikonomopoulos et al., 2016) on the External RNA Controls Consortium (ERCC) synthetic spike-ins (Baker et al., 2005), observing no length or GC bias (see later discussion) and then applied to human HEK-293 cells. However, it is worth noting that the longest ERCC spike-in transcript is only ~2 kb in length. ONT sequencing was recently benchmarked using "sequin" spike-ins (Hardwick et al., 2016; Hardwick et al., 2019), which include 15 multi-exonic transcripts in the 2.5–7-kb range.

4. *Sequencing depth:* Sequencing depth has been the Achilles' heel of isoform sequencing for a long time. Our initial PacBio isoform sequencing paper yielded ~500,000 (Sharon et al., 2013) CCS reads, and we then increased this to ~2 million a year later (Tilgner et al., 2014). These limitations (along with required input amount) were our primary motivation to explore dilution-based methods (Tilgner et al., 2015; Tilgner et al., 2018), which yielded 5 and 25 million long reads, respectively. It now seems that a breakthrough has been achieved with the PromethION from ONT, which at the time of writing appears to yield 20–50 million long reads, although currently at lower quality. Likewise, PacBio has announced an 8-million ZMW SMRT cell.

5. *Minimal input amount:* Both PacBio and now ONT require large amounts of input material. This requires either starting with large amounts of material or extensive PCR, the latter of which of course can decrease library complexity and introduce quantitative bias. Rolling-circle PCR, however, such as used by Volden et al. (2018), is an attractive work-around, as it amplifies molecules, while ensuring that all copies of the original cDNA molecule are sequenced in one single read. Therefore, no PCR duplicates are created unless standard PCR is used before or after. Dilution-based isoform sequencing methods (Tilgner et al., 2015; Tilgner et al., 2018) start with 100 pg to 1 ng. While there is PCR involved, this PCR occurs in barcoded wells or droplets, and all reads originating from one molecule can be collapsed back onto the original molecule.

As for deciding which sequencing platform to use, it largely depends on the specific goals and priorities of the study. For example, if high splice site accuracy is needed and money is no object, then PacBio arguably remains the best option in most cases. While ONT sequencing has a relatively high error rate, this may be tolerable in cases where perfect splice site accuracy is not required. If the goal is to detect RNA modifications or perform direct RNA sequencing, then ONT is the method of choice. Likewise, if portability is essential—e.g., for use in the field—then, ONT's MinION device is recommended.

MAPPING OF LONG READS

The long, noisy sequencing reads produced by third-generation technologies have posed new bioinformatic challenges for accurate spliced read alignment (Križanović et al., 2018). This has necessitated the development of specialized long-read alignment tools, the most popular of which currently include GMAP (Wu and Watanabe, 2005), STAR (Dobin et al., 2013), BLASR (Chaisson and Tesler, 2012), Minimap2 (Li, 2018), and Magic-BLAST (Boratyn et al., 2018). All of these aligners are splice-aware with the exception of BLASR, which was designed for alignment of genome sequencing data and, thus, is not recommended for RNA sequencing. The performance of GMAP and STAR was comprehensively benchmarked for PacBio long reads in the Association of Biomolecular Research Facilities (ABRF) next-generation sequencing study (Li et al., 2014). This study, however, predates the advent of nanopore sequencing, the introduction of the PacBio Sequel instrument, as well as the introduction of the Minimap2 and Magic-BLAST software. A detailed study of the performance of these mappers, especially with respect to accuracy on PacBio and ONT reads, would therefore be of high interest. The higher error rates of long-read sequencing platforms can also confound the precise determination of splice junctions. This has led to the emergence of several tools designed to cluster, polish, and collapse long reads into high-confidence isoforms, including Mandalorion (Byrne et al., 2017), Carnac-LR (Marchet et al., 2019), Pinfish[2], and FLAIR (Tang et al., 2018). Some of these tools can optionally be run in conjunction with short-read RNA-seq data to help increase accuracy of splice junction detection and quantification.

CONCLUDING REMARKS

RNA molecules can be multiple kilobases long and even up to 100 kilobases if premature molecules are considered. Yet, for the first decade of the new millennium, close to all transcriptome-wide approaches worked on RNA or cDNA fragments. From 2010 on, a revolution started that allowed the consideration of full-length RNA molecules, and at the time of writing, the resulting technologies are on the verge of going mainstream. The ultimate goal is to unambiguously decipher the sequence, structure, and abundance of each RNA molecule produced by a given cell, including its TSS, splicing structure, RNA modifications, and poly-A tail. Moving forward, the main technical challenges that will need to be overcome are the relatively high error rates, low throughput, and large input material requirements (compared with short-read RNA-seq). The continued development of novel bioinformatic approaches designed specifically for long, noisy reads can be expected to lead to further increases in performance. It seems that in the near future, a lot of biological reasoning could be performed with the isoform as a unit, rather than with single exons, splice sites, RNA edits, and modifications. Eventually, this development will hopefully further the large body of knowledge on the interactions between the respective machineries and allow us to appreciate all variables on individual RNA molecules at once.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

[2] https://github.com/nanoporetech/pinfish.

FUNDING

HT is a Leon Levy Research Fellow in Neuroscience and is furthermore grateful for a generous gift by Anita Garoppolo. SH acknowledges an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship (APP1156531). PF and AF acknowledge support from the National Human Genome Research Institute (U41HG007234), the Wellcome Trust (WT108749/Z/15/Z), and the European Molecular Biology Laboratory. PF is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00709/full#supplementary-material

REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7, 11706. doi: 10.1038/ncomms11706

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., et al. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* 18, 1435–1440. doi: 10.1038/nsmb.2143

Anvar, S. Y., Allard, G., Tseng, E., Sheynkman, G. M., de Klerk, E., Vermaat, M., et al. (2018). Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* 19, 46. doi: 10.1186/s13059-018-1418-0

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., et al. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4821–E4830. doi: 10.1073/pnas.1320101110.

Au, K. F., Underwood, J. G., Lee, L., and Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS One* 7, e46679. doi: 10.1371/journal.pone.0046679

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., et al. (2005). The External RNA Controls Consortium: a progress report. *Nat. Methods* 2, 731–734. doi: 10.1038/nmeth1005-731

Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M., and Boldogkői, Z. (2018). Dual platform long-Read RNA-sequencing dataset of the human cytomegalovirus lytic transcriptome. *Front. Genet.* 9, 432. doi: 10.3389/fgene.2018.00432

Beyer, A. L., and Osheim, Y. N. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.* 2, 754–765. doi: 10.1101/gad.2.6.754

Bolisetty, M. T., Rajadinakaran, G., and Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* 16, 204. doi: 10.1186/s13059-015-0777-z

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T. L. (2018). Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *bioRxiv*. doi: 10.1101/390013

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi: 10.1038/ncomms16027

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K. M. (2016). Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell* 165, 372–381. doi: 10.1016/j.cell.2016.02.045

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform.* 13, 238. doi: 10.1186/1471-2105-13-238

Cheng, J., Zhou, T., Liu, C., Shapiro, J. P., Brauer, M. J., Kiefer, M. C., et al. (1994). Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science* 263, 1759–1762. doi: 10.1126/science.7510905

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Cieply, B., and Carstens, R. P. (2015). Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip. Rev. RNA* 6, 311–326. doi: 10.1002/wrna.1276

Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., et al. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods* 12, 339–342. doi: 10.1038/nmeth.3321

Cramer, P., Pesce, C. G., Baralle, F. E., and Kornblihtt, A. R. (1997). Functional association between promoter structure and transcript alternative splicing. *Natl. Acad. Sci. U.S.A.* 94, 11456–11460. doi: 10.1073/pnas.94.21.11456

Daguenet, E., Dujardin, G., and Valcarcel, J. (2015). The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep.* 16, 1640–1655. doi: 10.15252/embr.201541116

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Deveson, I. W., Brunck, M. E., Blackburn, J., Tseng, E., Hon, T., Clark, T. A., et al. (2018). Universal alternative splicing of noncoding exons. *Cell Syst.* 6, 245–255. e5. doi: 10.1016/j.cels.2017.12.005

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Dougherty, M. L., Underwood, J. G., Nelson, B. J., Tseng, E., Munson, K. M., Penn, O., et al. (2018). Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* 28, 1566–1576. doi: 10.1101/gr.237610.118

Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E. J., et al. (2015). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat. Genet.* 48, 44–52. doi: 10.1038/ng.3449

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

Fagnani, M., Barash, Y., Ip, J. Y., Misquitta, C., Pan, Q., Saltzman, A. L., et al. (2007). Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol.* 8, R108. doi: 10.1186/gb-2007-8-6-r108

Fededa, J. P., Petrillo, E., Gelfand, M. S., Neverov, A. D., Kadener, S., Nogués, G., et al. (2005). A polar mechanism coordinates different regions of alternative splicing within a single gene. *Mol. Cell* 19, 393–404. doi: 10.1016/j.molcel.2005.06.035

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., et al. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genom.* 14, 670. doi: 10.1186/1471-2164-14-670

Fiszbein, A., and Kornblihtt, A. R. (2017). Alternative splicing switches: Important players in cell differentiation. *BioEssays* 39, 1600157. doi: 10.1002/bies.201600157

Forrest, A. R. R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M. J. L., Haberle, V., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Gallego-Paez, L. M., Bordone, M. C., Leote, A. C., Saraiva-Agostinho, N., Ascensão-Ferreira, M., and Barbosa-Morais, N. L. (2017). Alternative splicing: the pledge, the turn, and the prestige: the key role of alternative splicing in human biological systems. *Hum. Genet.* 136, 1015–1042. doi: 10.1007/s00439-017-1790-y

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi: 10.1038/nmeth.4577

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10, e0132628. doi: 10.1371/journal.pone.0132628

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* 36, 1197–1202. doi: 10.1038/nbt.4259

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30, 3004–3011. doi: 10.1093/bioinformatics/btu392

Hardwick, S. A., Bassett, S. D., Kaczorowski, D., Blackburn, J., Barton, K., Bartonicek, N., et al. (2019). Targeted, high-resolution RNA sequencing of non-coding genomic regions associated with neuropsychiatric functions. *Front. Genet.* 10, 309. doi: 10.3389/fgene.2019.00309

Hardwick, S. A., Chen, W. Y., Wong, T., Deveson, I. W., Blackburn, J., Andersen, S. B., et al. (2016). Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* 13, 792–798. doi: 10.1038/nmeth.3958

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7(Suppl 1), S4. doi: 10.1186/gb-2006-7-s1-s4

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Helfman, D. M., Cheley, S., Kuismanen, E., Finn, L. A., and Yamawaki-Kataoka, Y. (1986). Nonmuscle and muscle tropomyosin isoforms are expressed from a single gene by alternative RNA splicing and polyadenylation. *Mol. Cell. Biol.* 6, 3582–3595. doi: 10.1128/MCB.6.11.3582

Herzel, L., Straube, K., and Neugebauer, K. M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* 28, 1008–1019. doi: 10.1101/gr.232025.117

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., et al. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141–2144. doi: 10.1126/science.1090100

Karlsson, K., and Linnarsson, S. (2017). Single-cell mRNA isoform diversity in the mouse brain. *BMC Genom.* 18, 126. doi: 10.1186/s12864-017-3528-6

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., et al. (2014). The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* 42, D764–D770. doi: 10.1093/nar/gkt1168

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C.-H. A., Marr, M. T., and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev.* 25, 2502–2512. doi: 10.1101/gad.178962.111

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700. doi: 10.1038/nbt.2280

Križanović, K., Echchiki, A., Roux, J., and Šikić, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* 34, 748–754. doi: 10.1093/bioinformatics/btx668

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., et al. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* 49, 1731–1740. doi: 10.1038/ng.3988

Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A., and Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* 5, 34. doi: 10.1186/s13742-016-0140-7

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., et al. (2014). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* 32, 915–925. doi: 10.1038/nbt.2972

MacLeod, A. R., Houlker, C., Reinach, F. C., Smillie, L. B., Talbot, K., Modi, G., et al. (1985). A muscle-type tropomyosin in human fibroblasts: evidence for expression by an alternative RNA splicing mechanism. *Natl. Acad. Sci. U.S.A.* 82, 7835–7839. doi: 10.1073/pnas.82.23.7835

Marchet, C., Lecompte, L., Silva, C. D., Cruaud, C., Aury, J.-M., Nicolas, J., et al. (2019). De novo clustering of long reads by gene from transcriptomics data. *Nucleic Acids Res.* 47, e2. doi: 10.1093/nar/gky834

Mauger, O., and Scheiffele, P. (2017). Beyond proteome diversity: alternative splicing as a regulator of neuronal transcript dynamics. *Curr. Opin. Neurobiol.* 45, 162–168. doi: 10.1016/j.conb.2017.05.012

Mayr, C. (2016). Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol.* 26, 227–237. doi: 10.1016/j.tcb.2015.10.012

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., et al. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 9, e106689. doi: 10.1371/journal.pone.0106689

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859. doi: 10.1093/nar/29.13.2850

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., and Ragoussis, J. (2016). Benchmarking of the oxford nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* 6, 31602. doi: 10.1038/srep31602

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., et al. (2009). Direct RNA sequencing. *Nature* 461, 814–818. doi: 10.1038/nature08390

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Raj, B., and Blencowe, B. J. (2015). Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. *Neuron* 87, 14–27. doi: 10.1016/j.neuron.2015.05.004

Ramaswami, G., and Li, J. B. (2016). Identification of human RNA editing sites: a historical perspective. *Methods* 107, 42–47. doi: 10.1016/j.ymeth.2016.05.011

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Roy, C. K., Olson, S., Graveley, B. R., Zamore, P. D., and Moore, M. J. (2015). Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation. *Elife* 4, e03700. doi: 10.7554/eLife.03700

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., et al. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* 8, 59. doi: 10.1038/s41467-017-00050-4

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647. doi: 10.1126/science.1155390

Schor, I. E., Gómez Acuña, L. I., and Kornblihtt, A. R., (2013). "Coupling Between Transcription and Alternative Splicing," in *RNA and Cancer*. Ed. J. Y. Wu (Berlin, Heidelberg: Springer Berlin Heidelberg), 1–24. doi: 10.1007/978-3-642-31659-3_1

Schreiner, D., Nguyen, T.-M., Russo, G., Heber, S., Patrignani, A., Ahrné, E., et al. (2014). Targeted combinatorial alternative splicing generates brain

region-specific repertoires of neurexins. *Neuron* 84, 386–398. doi: 10.1016/j.neuron.2014.09.011

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., et al. (2013). High-Resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* 155, 1409–1421. doi: 10.1016/j.cell.2013.10.047

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346. doi: 10.1038/s41576-018-0003-4

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240. doi: 10.1038/nature12172

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi: 10.1038/nbt.2705

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* 7, 12065. doi: 10.1038/ncomms12065

Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., et al. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67, 148–161.e5. doi: 10.1016/j.molcel.2017.06.003

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi: 10.1038/nmeth.2714

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960. doi: 10.1126/science.1160342

Tang, A. D., Soulette, C. M., Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., et al. (2018). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv*. doi: 10.1101/410183

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–411. doi: 10.1101/118083

Tevz, G., McGrath, S., Demeter, R., Magrini, V., Jeet, V., Rockstroh, A., et al. (2016). Identification of a novel fusion transcript between human relaxin-1 (RLN1) and human relaxin-2 (RLN2) in prostate cancer. *Mol. Cell. Endocrinol.* 420, 159–168. doi: 10.1016/j.mce.2015.10.011

Tian, B., Pan, Z., and Ju, Y. L. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 17, 156–165. doi: 10.1101/gr.5532707

Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Natl. Acad. Sci. U.S.A.* 111, 9869–9874. doi: 10.1073/pnas.1400447111

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., et al. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742. doi: 10.1038/nbt.3242

Tilgner, H., Jahanbani, F., Gupta, I., Collier, P., Wei, E., Rasmussen, M., et al. (2018). Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* 28, 231–242. doi: 10.1101/gr.230516.117

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A, Chakrabortty, S., Djebali, S., et al. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625. doi: 10.1101/gr.134445.111

Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M., and Snyder, M. (2013). Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3 (Bethesda).* 3, 387–397. doi: 10.1534/g3.112.004812

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., et al. (2016). Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 11, e0162868. doi: 10.1371/journal.pone.0162868

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38, e159. doi: 10.1093/nar/gkq543

Treutlein, B., Gokce, O., Quake, S. R., and Südhof, T. C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci.* 111, E1291–E1299. doi: 10.1073/pnas.1403244111

Tseng, E., Tang, H. T., AlOlaby, R. R., Hickey, L., and Tassone, F. (2017). Altered expression of the FMR1 splicing variants landscape in premutation carriers. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1860, 1117–1126. doi: 10.1016/j.bbagrm.2017.08.007

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19, 535–548. doi: 10.1038/s41576-018-0017-y

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., 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.* 115, 9726–9731. doi: 10.1073/pnas.1806447115

Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., et al. (2013). The genome sequence of the colonial chordate, Botryllus schlosseri. *Elife* 2, e00569. doi: 10.7554/eLife.00569

Vuong, C. K., Black, D. L., and Zheng, S. (2016). The neurogenetics of alternative splicing. *Nat. Rev. Neurosci.* 17, 265–281. doi: 10.1038/nrn.2016.27

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Weirather, J. L., Afshar, P. T., Clark, T. A., Tseng, E., Powers, L. S., Underwood, J. G., et al. (2015). Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res.* 43, e116. doi: 10.1093/nar/gkv562

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100. doi: 10.12688/f1000research.10571.2

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., et al. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243. doi: 10.1038/nature07002

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., et al. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*. doi: 10.1101/459529

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Wyman, D., and Mortazavi, A. (2019). TranscriptClean: variant-aware correction of indels, mismatches, and splice junctions in long-read transcripts. *Bioinformatics* 35, 340–342. doi: 10.1093/bioinformatics/bty483

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164, 805–817. doi: 10.1016/j.cell.2016.01.029

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311. doi: 10.1038/nbt.3432

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. doi: 10.1038/ncomms14049

Conflict of Interest Statement: PF is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.