Notable challenges posed by long-read sequencing for the study of transcriptional diversity and genome annotation

Carolina Monzó, Adam Frankish and Ana Conesa

Genome Res. published online March 3, 2025
Access the most recent version at doi:10.1101/gr.279865.124

Notable challenges posed by long-read sequencing for the study of transcriptional diversity and genome annotation.

Carolina Monzó[1], Adam Frankish[2], Ana Conesa[1]

1 Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council (CSIC), Paterna, 46980, Spain

2 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus Hinxton, Cambridge, CB10 1SA, United Kingdom

Corresponding author: ana.conesa@csic.es,

Running title:

Long-read sequencing challenges: diversity and annotation

Abstract

Long-read sequencing (LRS) technologies have revolutionized transcriptomic research by enabling the comprehensive sequencing of full-length transcripts. Using these technologies, researchers have reported tens of thousands of novel transcripts, even in well-annotated genomes, while developing new algorithms and experimental approaches to handle the noisy data. The LRGASP community effort benchmarked LRS methods in transcriptomics and validated many novel, lowly-expressed, often times sample-specific transcripts identified by long reads. These molecules represent deviations of the major transcriptional program that were overlooked by short-read sequencing methods but are now captured by the full-length, single-molecule approach. This Perspective discusses the challenges and opportunities associated with LRS' capacity to unravel this fraction of the transcriptome, both in terms of transcriptome biology and genome annotation. For transcriptome biology, we need to develop novel experimental and computational methods to effectively differentiate technology errors from rare but real molecules. For genome annotation, we must agree on the strategy to capture molecular variability while still defining reference annotations that are useful for the genomics community.

Main text

Long-read sequencing (LRS) technologies, such as those developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have revolutionized genomic and transcriptomic research. Their ability to generate very long reads has enabled significant advancements, including complete sequencing of human chromosomes (Nurk et al. 2022) and full-length sequencing of single-molecule transcripts spanning kilobases (Sharon et al. 2013; Weirather et al. 2017; Soneson et al. 2019). This unprecedented capability earned LRS recognition by Nature Methods as the Method of the Year in 2022, highlighting its transformative impact on both fields (Marx 2023). One of the most significant contributions of long-read methods to the study of transcription is their capacity to uncover alternative isoforms with a confidence not present in short-read methods, which has led to the discovery of tens of thousands of novel transcripts even in well-annotated organisms (Glinos et al. 2022; Veiga et al. 2022; Zhang et al. 2022; Roach et al. 2020), and represents a data source of great value for the *de novo* annotation of the Earth Biogenome (Lawniczak et al. 2022). Despite its strengths, LRS presents several shortcomings. The quality of long-read RNA sequencing (lrRNA-seq) can be compromised by factors such as RNA degradation, biases introduced during library preparation, sequencing errors, and inaccurate bioinformatic processing during mapping, transcript assembly and quantification, which may lead to the incorrect identification of transcript models i.e. computational representations of transcripts depicting their transcription start and termination sites (TSS and TTS) and intron composition (Marx 2023; Amarasinghe et al. 2020). Most lrRNA-seq experiments rely on cDNA libraries, as they provide high sequencing throughput and accuracy. However, reverse transcription (RT) may introduce errors driven by specific sequences present in the RNA's primary sequence. These sequences can promote single-nucleotide errors and mis-priming, resulting in faulty cDNA molecules (technical artifacts) that inaccurately represent structural variations (Verwilt et al. 2023). The ONT direct RNA sequencing method can potentially overcome these issues while also identifying RNA modifications in the native molecule. However,

sequencing throughput from current direct RNA protocols is still relatively low compared to cDNA-based protocols (~20 M reads in direct RNA protocols (Oxford Nanopore Technologies 2019) versus ~130 M reads in cDNA-based protocols (Aguzzoli Heberle et al. 2024)), which compromises transcript identification. Despite these shortcomings, direct RNA holds great potential for improving transcript identification in the future.

A significant challenge in the analysis of lrRNA-seq data is the accurate identification of novel transcripts while effectively distinguishing them from artifacts introduced by the technology. To address this, various software tools have been created for reconstructing transcript models from LRS, and recently the technology has been subjected to rigorous benchmarking (Soneson et al. 2019; Kuo et al. 2020; Dong et al. 2023; Križanovic et al. 2018; Pardo-Palacios et al. 2024b; Su et al. 2024). The most comprehensive study to evaluate lrRNA-seq methods to date is the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP), a community effort aimed at systematically evaluating library preparation, sequencing platforms, and analysis tools for the identification and quantification of transcripts using LRS technologies (Pardo-Palacios et al. 2024b).

LRGASP included sequencing from PacBio, ONT, and Illumina short-reads, four library preparation methods, and the use of SQANTI3, a common tool for LRS quality control, to evaluate approximately 50 analysis pipelines (Pardo-Palacios et al. 2024a). In this scheme, full-splice-match (FSM) transcripts align with reference transcripts at all splice junctions, incomplete-splice-match (ISM) transcripts lack one or more junctions at the 5′ or 3′ ends, which could indicate RNA degradation or alternative initiation/termination sites, novel-in-catalog (NIC) transcripts exhibit new combinations of splice sites, and novel-not-in-catalog (NNC) transcripts have at least one novel donor or acceptor site (Figure 1A) (Tardaguila et al. 2018).

The LRGASP uncovered significant discrepancies among lrRNA-seq methods, particularly in the number and identity of transcripts in the novel SQANTI3 structural categories (ISM, NIC and NNC), as well as the degree of support from orthogonal data (Pardo-Palacios et al. 2024b). The authors concluded that these discrepancies were in part motivated by the differences in analysis goals pursued by each method. For instance, while Bambu (Chen et al. 2023b), IsoQuant (Prjibelski et al. 2023), and FLAIR (Tang et al. 2020) rely on the reference annotation to identify transcript models and consequently call few novel transcripts, other tools such as Lyric (Kaur et al. 2024) are designed to detect highly supported novel transcripts. Additionally, LRGASP conducted validation of long-read transcript models both experimentally and by manual curation. Surprisingly, a large number of novel transcripts were confirmed. Many transcripts identified by only one or a few tools were experimentally validated by PCR (Figure 1B), while several novel transcripts that had not been detected by any of the benchmarked tools were confidently annotated by GENCODE (Figure 1C). Most of these novel transcripts belonged to known genes and included novel combinations of annotated exons, displaced splice donor or acceptor sites, or previously undescribed intron retention events. Moreover, many validated novel transcripts were lowly expressed and found in *single samples* of the same experimental condition.

The LRGASP project highlighted two known but previously underappreciated insights: i) LRS technologies are also single-molecule sequencing methods that, unlike short-read sequencing, reveal the actual RNA molecules in a biological sample, and ii) RNA transcription is inherently noisy, and during the transcription and splicing processes, RNA molecules that deviate from the major gene expression products may be synthesized. This results in a pool of rare but real transcripts populating biological samples. Noisiness in RNA synthesis is long known and has been described in association with several stresses (Castells-Roca et al. 2011), diseases (Larsen et al. 2017; Wu et al. 2021), and also as a source of evolutionary adaptation (Singh and Ahi 2022; Wright et al. 2022). These rare transcripts were usually overlooked in conventional short-read sequencing because standard

analysis methods interpreted them as sporadic misalignment events and only library preparation techniques designed to target particular transcripts had the resolution to confidently detect lowly expressed transcript variants (Mercer et al. 2014). In contrast, the ability to capture full-length transcripts at ever greater sequencing depths by LRS means these deviations cannot be ignored any longer, presenting both a challenge and an opportunity in transcriptome research.

In this perspective, we discuss two significant consequences of LRS capturing unique, often times sample-specific, real RNA molecules: (i) the need to acknowledge this fraction of the transcriptome as different from the condition-specific transcriptional program and develop bioinformatics methods that distinguish these authentic molecules from technological artifacts to study their biological relevance, and (ii) the considerations of incorporating this transcriptomic complexity into genome annotation efforts. For both issues, we discuss the technology's potential to advance the field, the conceptual shifts it imposes on the research community, and the challenges associated with data processing and analysis.

Figure 1: LRGASP identified many transcripts only expressed in one or few samples. A) Main SQANTI3 structural categories for transcript models of known genes. B) Fraction of experimentally validated transcripts as a function of the number of WTC11 samples in which supportive reads were observed. C) Structural category classification and intersection of 50 loci from the WTC11 samples that were not identified by any transcript identification tool but were manually annotated by GENCODE. These loci were selected for having mapped reads across all six library preparation and sequencing platform combinations. B-C) Adapted from (Pardo-Palacios et al. 2024b).

Understanding the rare but real transcriptome

We refer to the previously described pool of diverse and rare RNA molecules, expressed alongside their major expression products, as Transcript Divergency (TD). TD encompasses the stochastic variability in gene expression resulting from deviations in transcription and splicing processes; in other words, it is a population of RNA molecules that diverge from the condition-defining transcriptional state. TD is distinct from transcriptional noise, defined as across-cell variations in expression levels (Tosti et al. 2021; Angelidis et al. 2019; Bartz et al.

2023), and alternative or aberrant splicing, defined as RNA processing characteristic of a cell-type or biological condition (Castells-Roca et al. 2011; Larsen et al. 2017; Kahles et al. 2018; Jose et al. 2019). TD molecules are, on the contrary, often times sample-specific and heterogeneous (Figure 2A).

Why should we care about TD in the first place? An increased transcriptome stochasticity has been described in relation to ageing (Angelidis et al. 2019; Enge et al. 2017; Kimmel et al. 2019; Martinez-Jimenez et al. 2017), disease (Larsen et al. 2017; Wu et al. 2021; Merlotti et al. 2023; Levy et al. 2023; Patowary et al. 2024) and cellular stress in a variety of organisms (Wu et al. 2021; Nicolas et al. 2018; Castells-Roca et al. 2011; Ramnarine et al. 2022). For example, in cancer, misregulated splicing patterns have been proposed as a source of neoantigens that contribute to tumour progression (Merlotti et al. 2023; Levy et al. 2023). In Alzheimer's disease, misregulated alternative splicing has been observed in genes associated with synaptic plasticity and neuronal function (Larsen et al. 2017; Patowary et al. 2024). As cells age, the accumulation of transcriptional noise increases errors in gene expression and disrupts pathway coordination, making cells more susceptible to functional decline and age-related conditions (Martinez-Jimenez et al. 2017). TD could contribute to or be a consequence of cellular stress by generating erroneous transcripts or proteins that may interfere with normal cellular functions, potentially leading to an altered phenotype. Moreover, RNA processing deviations from the main transcriptional program may represent a source of transcriptional innovation part of the evolutionary process, and low abundance structural variations may have biological functions, as has been shown for long non-coding RNAs (Mattick et al. 2023). Faithfully identifying these molecules within the LRS signal, will be invaluable to study and characterize these processes and understand how transcriptional stochasticity contributes to transcriptome biology.

Having stated the relevance of TD in biology, the next question is how to analyze and describe this fraction of the transcriptome, and how to differentiate it from technical noise and condition-specific transcript isoforms.

Recent efforts have moved beyond annotating alternative isoforms to conducting in-depth analysis of transcriptional diversity per gene. For instance, Reese et al. (2023) evaluate non-FSM reads based on three diversity mechanisms: TSS, exon junction chain and TTS, analyzing each gene to determine which of these three diversity mechanisms it is predominantly influenced by (Reese et al. 2023). However, structural elements of the RNA molecules are treated as independent variables, despite substantial research on exon exclusion and co-inclusion patterns that highlight the interdependent effects of these elements (Dong and Chen 2020; Merlotti et al. 2023; Singh and Ahi 2022; Oehler et al. 2022; Tilgner et al. 2015). Other tools assess transcriptional diversity through quantification of structural differences and evaluation of distance metrics between transcript models (Nanni et al. 2024). Due to the heterogeneous and stochastic nature of the TD, it is essential to study these molecules comprehensively, considering the full tandem combination of TSS, exon junction chains and TTS. Therefore, these findings underscore the need for developing new analytical methods that consider transcripts as integrated entities rather than the sum of their parts.

Numerous algorithms that aim at identifying and quantifying transcript models from long-read sequencing data have been published (Chen et al. 2023b; Kovaka et al. 2019; Gao et al. 2023; Wyman et al. 2019; Holmqvist et al. 2021; Orabi et al. 2023; Wang et al. 2021; Nip et al. 2023; de la Rubia et al. 2022; Petri and Sahlin 2023; Prjibelski et al. 2023; Tang et al. 2020; Volden et al. 2023; Bushmanova et al. 2019; Lienhard et al. 2023; Sahlin and Medvedev 2020). Generally, they do not make a specific distinction between TD and condition-associated isoforms, but rather between novel and known transcripts or between tissue-specific and ubiquitously expressed transcripts. We advocate for replication as one of the most effective strategies to differentiate TD from regular isoforms. For example, analyzing a high depth mouse brain lrRNA-seq PacBio Iso-Seq dataset (Supplementary Methods), we have observed a bimodal distribution of transcript models across replicates, with most transcripts either present in one sample or in all (5) analyzed samples, with

considerably fewer transcripts detected in 2 to 4 samples (Figure 2B). This is relevant as lack of biological replication is still frequent in many LRS studies that favor increased sequencing depth over replication, risking the soundness of the transcriptome composition derived from the LRS data.

A greater challenge, however, is distinguishing between technical and biological noise. A critical, and sometimes overlooked, first step is to run exhaustive quality control analyses on the data. Tools such as LongQC (Fukasawa et al. 2020), SQANTI3 (Pardo-Palacios et al. 2024a), GffCompare (Pertea and Pertea 2020) and SQANTI-reads (Keil et al. 2024) can be used to identify and discard reads and transcript models with common artifacts, such as intrapriming and reverse transcriptase switching. These tools can also be used to identify possible TD molecules, for example, those having non-canonical splicing – splicing events that do not follow the standard *GT-AG* or *GU-AG* splice site patterns–. Another important question is the approximate relative magnitude of these two sources of noise, which depends on the chosen long-read sequencing method. One way to begin addressing this question is by comparing the amount of deviating reads associated with an invariable set of RNAs, such as those provided by synthetic spike-in RNA variants (SIRVs) or Sequins spike-in controls (Hardwick et al. 2016), to those in a real sample. For example, we used the SQANTI3 framework to analyze the reads associated with Lexogen's E0 mix SIRV transcripts spiked into the aforementioned mouse brain samples, sequenced by different LRS methods, and the reads associated with mouse genes (Figure 2C-D). Since SIRVs are synthetic known transcripts, all reads should be classified by SQANTI3 as FSM in a technology-error-free scenario. Therefore, the amount of non-FSM reads provides a baseline for the errors associated with the long-read methodology, and any excess of non-FSM reads in the real mouse samples are candidates for TD. We found this difference to be between 10 and 15% of the sequencing output (Figure 2C).

However, this estimation has several caveats and only represents a possible upper limit. For example, real samples may be subject to higher RNA degradation than the carefully

controlled SIRV reagents, so the increased ISM fraction in real samples is likely to represent, at least partly, this additional RNA degradation. Conversely, some native RNAs could be more stable via protection by RNA-binding proteins, tertiary structures etc. Additionally, SIRV transcript structures are limited in complexity and fall short of faithfully capturing the distribution of transcript length and exon number present in a mammalian transcriptome, being limited for fully recapitulating library preparation and mapping errors. Consequently, for a more precise identification of TD molecules, three efforts should proceed in parallel: (i) continuing to improve long-read sequencing and library preparation methods by technology providers, (ii) developing novel, realistic ground-truth standards to differentiate them from any residual technological error, and (iii) developing new and improved isoform and quantification tools capable of identifying and quantifying TD.

To identify reads that represent TD, employing multiple LRS technologies on the same source RNA or utilizing orthogonal data such as short-reads, CAGE-seq (Takahashi et al. 2012), and Quant-seq (Moll et al. 2014) can be effective for validating junction sites, TSS and TTS, respectively. Tools like SQANTI3 are capable of integrating such information to label and filter long-read transcripts, potentially flagging reads that might represent rare RNA species or artifacts. However, this approach is both expensive and may lack generalizability. Moreover, as LRS technologies continue to advance and increase in throughput, their sensitivity for detecting rare junctions, TSS, and TTS will likely surpass that of orthogonal methods, diminishing their utility as supportive evidence. An alternative approach involves statistically modelling RNA processing deviations or learning these patterns from large datasets to train machine learning models capable of classifying them as TD. However, using very strict rules to filter sequencing datasets or training machine learning models on specific sets of data have the potential of discarding real biological TD from less abundant transcripts or those present in sample-specific conditions. Therefore, developing accurate algorithms to identify TD represents a new and important challenge for the computational biology and RNA communities.

Figure 2. Transcript Divergency detection. A) Differences between transcriptional noise (TN), alternative splicing (AS) and transcript divergency (TD). TN indicates stochastic variation in transcript expression levels. AS is programmed alternative processing of RNA that can be consistently detected. TD represents rare but real RNA molecules. B) Structural category classification and intersection of loci among biological replicates in mouse brain samples. Replication strongly reduces the biological and technical noise in the samples. Structural category classification of reads among PacBio Iso-Seq and Kinnex, and ONT (R9 flowcell) sequencing methods in C) spike-in RNA variants (SIRV) and D) mouse brains. Non FSM reads in spikes represent the technical noise level of the technology, while the excess of these reads in the real sample is indicative of TD.

Genome annotation in a context of increasing transcriptional diversity

The discovery of thousands of transcripts across different species through LRS experiments, once technical and bioinformatic artifacts have been eliminated from the equation, poses a significant challenge for genome annotation. The task is to balance the comprehensive description of this variety with the need to maintain practical and useful reference transcriptomes (Figure 3A). Effective strategies must be developed to incorporate this complexity without overwhelming the annotation process.

The two reference human gene and transcript annotation resources with the longest standing are NCBI's RefSeq (Maglott et al. 2000; O'Leary et al. 2016) and EMBL-EBI's Ensembl/GENCODE (Harrow et al. 2006; Frankish et al. 2023), both of which started more than 20 years ago. The historic sets of human transcriptomic data that supported the manually annotated transcript models made by both teams consisted of what we would now consider very small numbers of Sanger sequenced expressed sequence tags (ESTs) and cDNAs, with fewer than 0.5M cDNAs and approximately 8.5M ESTs ever captured. In an era of relative data sparsity, annotating every transcript detected was a plausible goal notwithstanding the immaturity of much of the software and computational tooling supporting the annotation effort. Transcript models based on short-read RNA-seq data have been

available for many years, and there is a huge volume of public RNA-sequencing data on which to base models. While detailed discussion of RNA-seq methods is out of scope for this perspective, we note that none of the issues described are unique to long transcriptomic methods. For example, with a few years of completion of the human genome sequence, analysis of millions of EST and cDNA sequences identified splicing complexity (Hayashizaki and Carninci 2006) as did RNA-seq assembly efforts from Fantom (Hon et al. 2017) and MiTranscriptome (Iyer et al. 2015) a decade later. However, well understood problems of read length and uncertainty over reliability of assignment of exons and introns to transcript models, particularly in genes with significant evidence of alternative splicing have meant that they have not been adopted by Ensembl/GENCODE as part of the reference annotation, and sets produced by RefSeq (tagged as XM, XR, XP), are released as an adjunct to manually annotated models (NM, NR, NP) and not subject to further manual curation (O'Leary et al. 2016). Today, decades-worth of our original endeavours to sequence full-length transcripts can be overtaken by a single PacBio Kinnex or ONT cDNA/directRNA sequencing experiment. The depth of data now available challenges both our technical ability to identify and describe every transcript and our philosophical approach to producing reference annotation. Put simply, sampling depth will increase and many novel transcripts will be captured by reads at a quality that will allow them to be accurately mapped and this has the potential to massively increase the number of transcripts that can be added to reference transcript sets like GENCODE, perhaps by many millions. Increasing depth of long read sequencing will also improve detection of other RNA species such as circular RNAs (circRNAs), single stranded covalently closed loops of RNA formed by an alternative splicing pathway. CircRNA detection requires the use of specific library preparation protocols, but these have been developed for both ONT and PacBio sequencing platforms (Rahimi et al. 2021; You et al. 2015). CircRNAs also present challenges to representation in standard linear genome browsers to identify their distinct biology compared to linear RNA species. An analogous challenge applies to intragenic trans-spliced RNAs, where a transcript contains non-collinear splicing, e.g. exons may be spliced out of order. Trans-spliced RNAs will be

sequenced in long transcriptomic experiments but are likely to be classified incorrectly without specific mapping strategies (Chen et al. 2023a). How should reference annotation resources approach this challenge?

Reference annotation resources must provide their users with information about genes and transcripts that supports the downstream analysis they are undertaking. Different users will want to perform different downstream analyses that may benefit from using different transcript annotations which may appear incompatible. For example, when using RNA-seq data to perform gene-level transcriptomic analysis, a maximal representation of transcriptomic complexity is beneficial to ensure that as many RNA-seq reads as possible are correctly assigned to their gene of origin. This would require that the annotation captures the maximal extent of the gene even where there is variability in the transcript start and end through the use of alternative TSS and TTS sites, also capturing all alternative splicing events and potentially even intron retention. Similarly, LRGASP demonstrated the benefit of a comprehensive reference annotation, as reference guided transcript annotation tools were able to accurately identify many more transcripts than those methods that did not use reference annotations (Pardo-Palacios et al. 2024b). It is reasonable to assume that a complete reference annotation that captures every observed transcript would support reference-guided annotation methods to maximise read mapping.

A further benefit of comprehensive reference annotation is highlighting features that may potentially be overlooked in the absence of annotation. For example, ‘deep intronic’ variants – genetic variants that are more than 100bp away from the closest exon-intron boundary – may not be identified by standard variant annotation pipelines, or if identified may not be considered significant. Comprehensive annotation that captures infrequently included exons and splice sites may highlight possible functional significance of these previously un- or under-annotated variants. Such variation in exonic and intronic splice enhancing or splice silencing signals may affect the inclusion rate of the exon, and by upregulating the inclusion

of the exon in more transcripts from a haploinsufficient gene, could be implicated in disease by disrupting the amount of functional transcripts and the protein they encode.

Conversely, for other applications a more minimal representation of the transcriptional output of the locus is beneficial, where many transcripts are annotated at a gene capturing alternative TSS, poly(A) sites and splicing, this may lead to the annotation of multiple alternative coding sequences (CDSs), perhaps with multiple translation initiation sites and termini. This can be problematic for the annotation/interpretation of variation data where the base(s) affected by a variant may have multiple possible functional consequences assigned. The same genomic position can be assigned as affecting a CDS, 5' or 3' UTR, core or proximal splice site sequence or intronic sequence depending on the number and characteristics of the transcripts annotated at a gene. Indeed, the annotation of multiple CDSs in different frames could lead to the same variant being called as synonymous, non-synonymous or as a loss-of-function variant (Frankish et al. 2015).

Similarly, although a comprehensive annotation would be beneficial for transcript identification and gene-level quantification, it may instead hinder accurate isoform-level quantification by causing over-dispersion of counts between structurally similar isoforms, or even assigning reads to rare isoforms not expressed in the analysed sample. Therefore, using as references a minimal representation of transcriptional output and personal genome sequences may yield more accurate quantification results for isoform-level quantification.

Another practical consideration of capturing large numbers of transcripts is the effect it may have on visualising genomic data. Genome browsers such as Ensembl (Martin et al. 2023) and UCSC (Nassar et al. 2023) are frequently the primary access point for interrogating the intersection between genomic data and reference annotation, including gene and transcript annotation. Significant inflation in the number of transcripts annotated at a locus can negatively affect the browsing experience even where compression options are available. Even using browsers like IsoVis (Wan et al. 2024), specifically developed to deal with high

amounts of isoforms, may eventually run into the same issues. In simple terms, where there are many transcripts it reduces the space on the screen to display other data tracks and can also make interpretation more difficult as tracking relationships between features may be affected by the increased space between them and the presence of interposed transcript models.

Our ability to describe a transcriptome has historically been constrained by the requirement to map all our transcriptomic data to a reference genome. However, such reference genomes have many loci that do not accurately represent actual transcriptional output on the reference allele or haplotype. To alleviate reference bias effects, high quality sequences are being generated for pangenome projects for many species; for example the human reference pangenome project (~1200 human genomes) (Liao et al. 2023), with some at telomere-to-telomere quality (Schneider et al. 2017) as well as the model organisms mouse (Keane et al. 2011), rat (de Jong et al. 2024), farmed animals such as cow (Smith et al. 2023) and pig (Miao et al. 2024) and the crop plants rice (Shang et al. 2022) and brassicas (Golicz et al. 2016). These pangenome resources promise to revolutionize our ability to accurately map transcriptomic data to the haplotype from which it originates, supporting the confident identification of expression and splicing QTLs and allowing the creation of haplotype-specific representation of the transcriptome. Methods have already been developed to map RNA-seq data to pangenome graphs e.g. VG and RPVG (Sibbesen et al. 2023), and HISAT2 (Kim et al. 2019). These methods demonstrate the improvements in alignment achieved by using the graph over a single reference genome, and they are planned to be extended to support the mapping of lrRNA-seq data. This is likely to reveal haplotype-specific or haplotype-enriched or depleted splicing and transcripts. This presents a further challenge to the groups producing reference gene and transcript annotation. It will be necessary to maintain and improve the annotation of the reference genomes that are likely to remain very widely used for the foreseeable future but also to produce the haplotype-specific annotation that will be needed to harness the full potential of the

pangenome. Currently, new haplotypes are frequently annotated by mapping or projecting annotation from the reference genome, with some *de novo* annotation to add genes, but as transcriptomic datasets can be mapped to genetically identical (or at least very closely related) haplotypes it will be essential to record haploytpe-specific transcripts. As the pangenome and its associated pantranscriptome matures it will become both possible and necessary to accurately reflect haplotype-specific splicing in the reference gene and transcript annotation produced for the pangenome, but at least initially the annotation on the pangenome will have to contend with some of the challenges for annotation of transcripts on a single reference genome. Comprehensive annotation of all possible transcripts that could be considered TD has the potential to add large numbers of transcripts that are not relevant to the reference genome and many other haplotypes (and could be considered false positive errors in these haplotypic contexts), however, excluding transcripts from the set would lead to their subsequent exclusion (and false negative errors) when annotation from a single reference genome is projected to alternative haplotypes.

A reference annotation can take one of two broad approaches on how to deal with TD (it should always seek to exclude technical artefacts). Reference annotation can be inclusive and capture transcripts comprehensively or it can seek to exclude transcripts defined as TD. While this latter approach is both attractive and technically feasible, it presupposes that we have sufficient scientific knowledge and resolution in the data to exclude all biological noise and include all ‘real’ transcripts. An exclusionary approach may be able to achieve something quite close to an ideal minimal representation of transcripts, but it may not be future-proof to new understanding of genome biology, developments in experimental methods and data generation and computational tools. We can take a lesson from the historical annotation produced by the predecessor to Ensembl/GENCODE in the era of data scarcity at the time the original sequencing of the human genome was completed (Benson et al. 1997). At this time, while much attention was paid to the annotation of protein-coding transcripts, transcripts with premature termination codons likely to be subject to nonsense-

mediated decay (NMD) were also annotated along with transcripts that retained intronic sequence. A comprehensive annotation of pseudogenes of protein-coding genes was also produced and transcripts with no obvious protein-coding potential were also annotated. In all these cases, transcript features were recorded in the annotation without a clear understanding of their relevance to the function of the cell and in all cases it can be argued that their annotation was well founded and proved useful. NMD in particular but also intron-retention have been demonstrated as important in post-transcriptional gene regulation (Jacob and Smith 2017; Monteuuis et al. 2019), comprehensive pseudogene annotation has supported analysis of genome evolution and been practically useful in understanding mapping queries in more recent transcriptomic data (Frankish and Harrow 2014; Amaral et al. 2023), and the non-coding transcripts anticipated the then-nascent, now vast field of long non-coding RNA study (Ramilowski et al. 2020; Mattick and Rinn 2015). Not throwing data away but capturing and labelling it as well as possible has proved useful since the original annotation efforts. Even where transcript structures remain unchanged over the years (because the original decision of the starts and ends of exons was correctly determined early on), broader and deeper annotation and metadata may be layered on to the original models. This cannot happen if the transcript is not part of the annotation.

Given this, a goal for reference annotation should be to capture everything, every transcript structure and CDS, every TSS and polyadenylation site, and do it in a haplotype specific way across the pangenome (Figure 3B). Data generation methods that are currently available, along with improved computational methods will support movement towards this goal. If the transcript set is maximally inclusive, to support the broadest possible variety of analysis by downstream users then it must be maximally labelled to support the filtering or subsetting of the larger set to provide the transcripts that are best suited for a users specific analysis.

Simple subsets of transcripts already exist in reference annotation. For example, GENCODE has several sets of transcripts readily identifiable in the genome browsers and release files; GENCODE comprehensive (everything), GENCODE basic (subset of full length coding

transcripts and minimal representation of other gene biotypes), GENCODE primary (smaller subset of transcripts of likely functional significance based on expression and evolutionary conservation and constraint) (Frankish et al. 2023), MANE Select, MANE Plus Clinical (subsets of transcripts agreed with NCBI RefSeq based on expression, evolutionary conservation and constraint, developed to support the consistent reporting of clinical variation) (Morales et al. 2022), APPRIS principal isoforms (protein-centric analysis to determine likely functional isoforms) (Rodriguez et al. 2018). These somewhat naive subsets are useful in providing an initial set of transcripts for analysis, browser display and variant reporting but they point to the beginning of the possibility rather than an endpoint.

As the number of reads generated by lrRNA-seq experiments explodes, allowing us to detect many transcripts including those that are very rare, haplotype specific, cell-state or stimulation specific, etc, we will also have greater power to describe every transcript added to reference annotation. Every annotated transcript can be compared to reads from any sequencing experiment deposited in a public sequence archive to determine whether its expression was detected in that experiment, and if it was, extract what were the absolute and relative expression values in terms of reads or proportion of transcription from its locus represented by the transcript. In addition, the lack of detection of the transcript in an experiment may also be recorded. In future, analysis pipelines will be able to identify the most appropriate haplotype in the reference pangenome for mapping to ensure haplotype-specific transcripts are appropriately placed. These data can be accurately captured for every transcript, alongside metadata from the experiment such as tissue, cell type, activation state, developmental stage, age, sex, disease state, etc to be stored in a database (Figure 3B). The database could be updated with new transcript annotation and experimental data as they become available and permit customisation based on any captured metadata (e.g. tissue and abundance) allowing users flexibility to create the right dataset for their analysis or create automated subsets based on predetermined filter parameters. Integration with genome browsers might allow on the fly creation of annotation tracks for display while

archiving strategies and detailed filterset metadata could be included to allow the recreation of past filtered transcript sets at any time.

Figure 3: Genome annotation of transcriptome diversity. A) Benefits and challenges of annotating the transcriptome diversity revealed by long read sequencing methods. Challenges include storage, interpretation and visualization of vast amounts of transcripts, and are outweighed by the benefits of using a comprehensive reference for accurate gene quantification, transcript reconstruction and function discovery among others. B) Redefining the annotation paradigms. Defining haplotype-specific pantranscriptomes aligns with current pangenome efforts to describe genome diversity. Extensive metadata annotation of the wealth of data provided by LRS allows inclusion and reference customization via subsetting to accommodate diverse analysis scenarios.

Concluding remarks

Single-molecule long-read sequencing technologies have demonstrated an unprecedented capacity to uncover the vast diversity of both common and rare RNA molecules that constitute the transcriptomes. Traditionally, data analysis methods have focused on identifying and characterizing the consistent and functional components of the transcriptome. However, the wealth of new data from these advanced technologies suggests that rare but genuine transcripts can no longer be ignored, but at the same time, not every single new transcript found in a LRS experiment may require the same consideration. Embracing this potential necessitates the development of new analysis methods and a redefinition of existing paradigms. Innovative analytical procedures are required to effectively distinguish technical artifacts from biological noise and to assess their biological relevance in a variety of contexts. Concurrently, new strategies and protocols must be established to annotate the ever-growing diversity of transcriptomes in a manner that is useful for both current and future research. Rather than simplifying analysis, long-read sequencing presents exciting analytical challenges, demanding sophisticated approaches to manage the vast expanse of molecular data being discovered.

Competing interest statement

A.C. has received in-kind funding from Pacific Biosciences for library preparation and sequencing. A.C. collaborates with Oxford Nanopore in the Marie Skłodowska-Curie Actions Doctoral Network project LongTREC.

Data availability

All newly generated data used in this study is publicly available at the European Nucleotide Archive, study accession PRJEB85167.

Acknowledgements

We thank Alejandro Paniagua for bioinformatic analysis and scientific discussions, and Tianyuan Liu for graphic design and scientific discussions. We thank Isabel Fariñas and José Manuel Morante from the University of Valencia, and Luis Ferrández for mouse handling and data generation. This work was supported by a grant from the National Institutes of Health (1R21HG011280-01), the Spanish MICIN (PID2020-119537RB-I00) and the European Union's programme Horizon Europe under a Marie Skłodowska-Curie grant (101149931). The computations were performed on the high performance computing cluster Garnatxa at the Institute for Integrative Systems Biology (I2SysBio), I2SysBio is a joint research institute of the University of Valencia (UV) and Spanish National Research Council (CSIC).

References

Aguzzoli Heberle B, Brandon JA, Page ML, Nations KA, Dikobe KI, White BJ, Gordon LA, Fox GA, Wadsworth ME, Doyle PH, et al. 2024. Mapping medically relevant RNA isoform diversity in the aged human frontal cortex with deep long-read RNA-seq. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-024-02245-9.

Amaral P, Carbonell-Sala S, De La Vega FM, Faial T, Frankish A, Gingeras T, Guigo R, Harrow JL, Hatzigeorgiou AG, Johnson R, et al. 2023. The status of the human gene catalogue. *Nature* **622**: 41–47.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.

Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, Tsitsiridis G, Ansari M, Graf E, Strom T-M, et al. 2019. An atlas of the aging lung mapped by single cell

transcriptomics and deep tissue proteomics. *Nat Commun* **10**: 963.

Bartz J, Jung H, Wasiluk K, Zhang L, Dong X. 2023. Progress in Discovering Transcriptional Noise in Aging. *Int J Mol Sci* **24**. http://dx.doi.org/10.3390/ijms24043701.

Benson DA, Boguski MS, Lipman DJ, Ostell J. 1997. GenBank. *Nucleic Acids Res* **25**: 1–6.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**. http://dx.doi.org/10.1093/gigascience/giz100.

Castells-Roca L, García-Martínez J, Moreno J, Herrero E, Bellí G, Pérez-Ortín JE. 2011. Heat shock response in yeast involves changes in both transcription rates and mRNA stabilities. *PLoS One* **6**: e17272.

Chen Y-C, Chen C-Y, Chiang T-W, Chan M-H, Hsiao M, Ke H-M, Tsai IJ, Chuang T-J. 2023a. Detecting intragenic trans-splicing events from non-co-linearly spliced junctions by hybrid sequencing. *Nucleic Acids Res* **51**: 7777–7797.

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023b. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195.

de Jong TV, Pan Y, Rastas P, Munro D, Tutaj M, Akil H, Benner C, Chen D, Chitre AS, Chow W, et al. 2024. A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats. *Cell Genom* **4**: 100527.

de la Rubia I, Srivastava A, Xue W, Indi JA, Carbonell-Sala S, Lagarde J, Albà MM, Eyras E. 2022. RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *Genome Biol* **23**: 153.

Dong X, Chen R. 2020. Understanding aberrant RNA splicing to facilitate cancer diagnosis and therapy. *Oncogene* **39**: 2231–2242.

Dong X, Du MRM, Gouil Q, Tian L, Jabbari JS, Bowden R, Baldoni PL, Chen Y, Smyth GK, Amarasinghe SL, et al. 2023. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat Methods* **20**: 1810–1821.

Enge M, Arda HE, Mignardi M, Beausang J, Bottino R, Kim SK, Quake SR. 2017. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**: 321–330.e14.

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, et al. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**: D942–D949.

Frankish A, Harrow J. 2014. GENCODE pseudogenes. *Methods Mol Biol* **1167**: 129–155.

Frankish A, Uszczynska B, Ritchie GRS, Gonzalez JM, Pervouchine D, Petryszak R, Mudge JM, Fonseca N, Brazma A, Guigo R, et al. 2015. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* **16 Suppl 8**: S2.

Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, Wang Y, Kadash-Edmondson KE, Lin L, Xing Y. 2023. ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci Adv* **9**: eabq5072.

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**: 353–359.

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, et al. 2016. The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun* **7**: 13390.

Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB, Nielsen LK, Mattick JS, Mercer TR. 2016. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods* **13**: 792–798.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, et al. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**: S4.1–9.

Hayashizaki Y, Carninci P. 2006. Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet* **2**: e63.

Holmqvist I, Bäckerholm A, Tian Y, Xie G, Thorell K, Tang K-W. 2021. FLAME: long-read bioinformatics tool for comprehensive spliceome characterization. *RNA* **27**: 1127–1139.

Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**: 199–204.

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**: 199–208.

Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**: 1043–1057.

Jose BR, Gardner PP, Barquist L. 2019. Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem Soc Trans* **47**: 527–539.

Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, Cancer Genome Atlas Research Network, et al. 2018. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**: 211–224.e6.

Kaur G, Perteghella T, Carbonell-Sala S, Gonzalez-Martinez J, Hunt T, Mądry T, Jungreis I, Arnan C, Lagarde J, Borsari B, et al. 2024. GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing. *bioRxiv*. http://biorxiv.org/lookup/doi/10.1101/2024.10.29.620654.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289–294.

Keil N, Monzó C, McIntyre L, Conesa A. 2024. SQANTI-reads: a tool for the quality assessment of long read data in multi-sample lrRNA-seq experiments. *bioRxiv*. http://dx.doi.org/10.1101/2024.08.23.609463.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Kimmel JC, Penland L, Rubinstein ND, Hendrickson DG, Kelley DR, Rosenthal AZ. 2019. Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res* **29**: 2088–2103.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.

Križanovic K, Echchiki A, Roux J, Šikic M. 2018. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**: 748–754.

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**: 751.

Larsen PA, Lutz MW, Hunnicutt KE, Mihovilovic M, Saunders AM, Yoder AD, Roses AD. 2017. The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease. *Alzheimers Dement* **13**: 828–838.

Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov K, Blaxter ML, Marques Bonet T, et al. 2022. Standards recommendations for the Earth BioGenome Project. *Proc Natl Acad Sci U S A* **119**. http://dx.doi.org/10.1073/pnas.2115639118.

Levy R, Alter Regev T, Paes W, Gumpert N, Cohen Shvefel S, Bartok O, Dayan-Rubinov M, Alon M, Shmueli MD, Levin Y, et al. 2023. Large-Scale Immunopeptidome Analysis Reveals Recurrent Posttranslational Splicing of Cancer- and Immune-Associated Genes. *Mol Cell Proteomics* **22**: 100519.

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Börno S, Caiment F, Vingron M, Herwig R. 2023. IsoTools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics* **39**. http://dx.doi.org/10.1093/bioinformatics/btad364.

Maglott DR, Katz KS, Sicotte H, Pruitt KD. 2000. NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28**: 126–128.

Martinez-Jimenez CP, Eling N, Chen H-C, Vallejos CA, Kolodziejczyk AA, Connor F, Stojic L, Rayner TF, Stubbington MJT, Teichmann SA, et al. 2017. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**: 1433–1436.

Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J, et al. 2023. Ensembl 2023. *Nucleic Acids Res* **51**: D933–D941.

Marx V. 2023. Method of the Year 2022: long-read sequencing. *Nat Methods* **20**: 1.

Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, Chen R, Dean C, Dinger ME, Fitzgerald KA, et al. 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24**: 430–447.

Mattick JS, Rinn JL. 2015. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* **22**: 5–7.

Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, Nielsen LK, Dinger ME,

Mattick JS. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc* **9**: 989–1009.

Merlotti A, Sadacca B, Arribas YA, Ngoma M, Burbage M, Goudot C, Houy A, Rocañín-Arjó A, Lalanne A, Seguin-Givelet A, et al. 2023. Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer. *Sci Immunol* **8**: eabm6359.

Miao J, Wei X, Cao C, Sun J, Xu Y, Zhang Z, Wang Q, Pan Y, Wang Z. 2024. Pig pangenome graph reveals functional features of non-reference sequences. *J Anim Sci Biotechnol* **15**: 32.

Moll P, Ante M, Seitz A, Reda T. 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat Methods* **11**: i–iii.

Monteuuis G, Wong JJL, Bailey CG, Schmitz U, Rasko JEJ. 2019. The changing paradigm of intron retention: regulation, ramifications and recipes. *Nucleic Acids Res* **47**: 11497–11513.

Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva O, Farrell CM, et al. 2022. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**: 310–315.

Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT, et al. 2023. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**: D1188–D1195.

Nicolas D, Zoller B, Suter DM, Naef F. 2018. Modulation of transcriptional burst frequency by histone acetylation. *Proc Natl Acad Sci U S A* **115**: 7153–7158.

Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, Birol I. 2023. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nat Commun* **14**: 2940.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Oehler D, Spychala A, Gödecke A, Lang A, Gerdes N, Ruas J, Kelm M, Szendroedi J, Westenfeld R. 2022. Full-length transcriptomic analysis in murine and human heart reveals diversity of PGC-1α promoters and isoforms regulated distinctly in myocardial ischemia and obesity. *BMC Biol* **20**: 169.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–45.

Orabi B, Xie N, McConeghy B, Dong X, Chauve C, Hach F. 2023. Freddie: annotation-independent detection and discovery of transcriptomic alternative splicing isoforms using long-read sequencing. *Nucleic Acids Res* **51**: e11.

Oxford Nanopore Technologies. 2019. *The Value of Full-Length Transcripts without Bias*. https://a.storyblok.com/f/196663/x/8badf93497/rna-sequencing-white-paper.pdf.

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín

R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2024a. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**: 793–797.

Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, et al. 2024b. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods*. http://dx.doi.org/10.1038/s41592-024-02298-3.

Patowary A, Zhang P, Jops C, Vuong CK, Ge X, Hou K, Kim M, Gong N, Margolis M, Vo D, et al. 2024. Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms. *Science* **384**: eadh7688.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**. http://dx.doi.org/10.12688/f1000research.23297.2.

Petri AJ, Sahlin K. 2023. isONform: reference-free transcriptome reconstruction from Oxford Nanopore data. *Bioinformatics* **39**: i222–i231.

Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**: 915–918.

Rahimi K, Venø MT, Dupont DM, Kjems J. 2021. Nanopore sequencing of brain-derived full-length circRNAs reveals circRNA-specific exon usage, intron retention and microexons. *Nat Commun* **12**: 4825.

Ramilowski JA, Yip CW, Agrawal S, Chang J-C, Ciani Y, Kulakovskiy IV, Mendez M, Ooi JLC, Ouyang JF, Parkinson N, et al. 2020. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res* **30**: 1060–1072.

Ramnarine TJS, Grath S, Parsch J. 2022. Natural variation in the transcriptional response of Drosophila melanogaster to oxidative stress. *G3* **12**. http://dx.doi.org/10.1093/g3journal/jkab366.

Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, Rezaie N, Trout D, Razavi-Mohseni M, Jiang Y, et al. 2023. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *bioRxiv*. http://dx.doi.org/10.1101/2023.05.15.540865.

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. 2020. The full-length transcriptome of C. elegans using direct RNA sequencing. *Genome Res* **30**: 299–312.

Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. 2018. APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res* **46**: D213–D217.

Sahlin K, Medvedev P. 2020. De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. *J Comput Biol* **27**: 472–484.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849–864.

Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, Lin H, Hu M, Zhao F, Zhang C, et al. 2022. A

super pan-genomic landscape of rice. *Cell Res* **32**: 878–896.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014.

Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, Paten B. 2023. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat Methods* **20**: 239–247.

Singh P, Ahi EP. 2022. The importance of alternative splicing in adaptive evolution. *Mol Ecol* **31**: 1928–1938.

Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, Low WY, Pausch H, Demyda-Peyrás S, Prendergast J, et al. 2023. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome Biol* **24**: 139.

Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10**: 3359.

Su Y, Yu Z, Jin S, Ai Z, Yuan R, Chen X, Xue Z, Guo Y, Chen D, Liang H, et al. 2024. Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. *Nat Commun* **15**: 3972.

Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5’ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**: 542–561.

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**: 1438.

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, 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.

Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante CD, Rasmussen M, Snyder MP. 2015. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol* **33**: 736–742.

Tosti L, Hang Y, Debnath O, Tiesmeyer S, Trefzer T, Steiger K, Ten FW, Lukassen S, Ballke S, Kühl AA, et al. 2021. Single-Nucleus and In Situ RNA-Sequencing Reveal Cell Topographies in the Human Pancreas. *Gastroenterology* **160**: 1330–1344.e11.

Veiga DFT, Nesta A, Zhao Y, Deslattes Mays A, Huynh R, Rossi R, Wu T-C, Palucka K, Anczukow O, Beck CR, et al. 2022. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv* **8**: eabg6711.

Verwilt J, Mestdagh P, Vandesompele J. 2023. Artifacts and biases of the reverse transcription reaction in RNA sequencing. *RNA* **29**: 889–897.

Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. 2023. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with

Mandalorion. *Genome Biol* **24**: 167.

Wan CY, Davis J, Chauhan M, Gleeson J, Prawer YDJ, De Paoli-Iseppi R, Wells CA, Choi J, Clark MB. 2024. IsoVis - a webserver for visualization and annotation of alternative RNA isoforms. *Nucleic Acids Res* **52**: W341–W347.

Wang Y, Hu Z, Ye N, Yin H. 2021. IsoSplitter: identification and characterization of alternative splicing sites without a reference genome. *RNA* **27**: 868–875.

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* **6**: 100.

Wright CJ, Smith CWJ, Jiggins CD. 2022. Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet* **23**: 697–710.

Wu W, Syed F, Simpson E, Lee C-C, Liu J, Chang G, Dong C, Seitz C, Eizirik DL, Mirmira RG, et al. 2021. The Impact of Pro-Inflammatory Cytokines on Alternative Splicing Patterns in Human Islets. *Diabetes*. http://dx.doi.org/10.2337/db20-0847.

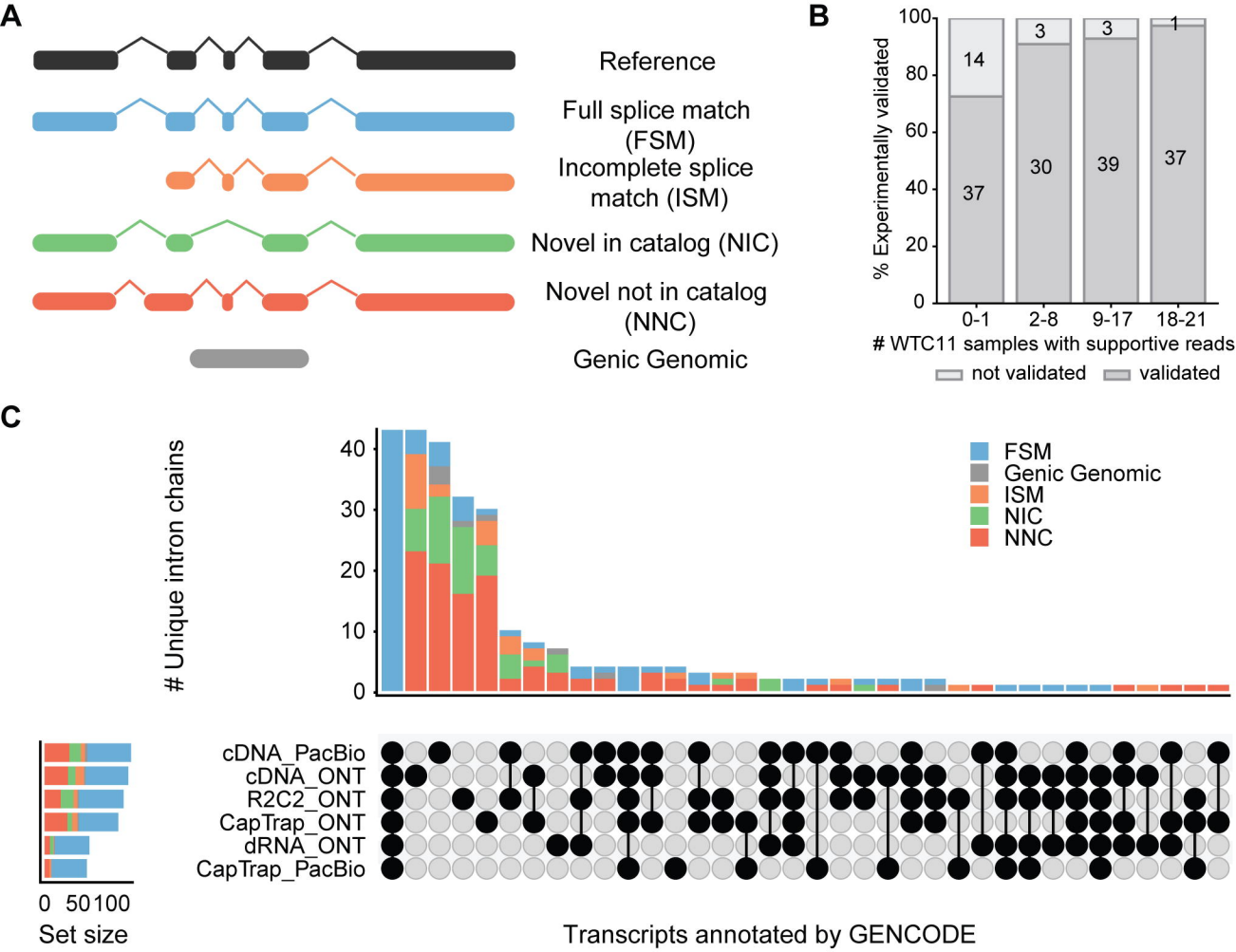
Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al. 2019. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. http://biorxiv.org/lookup/doi/10.1101/672931.

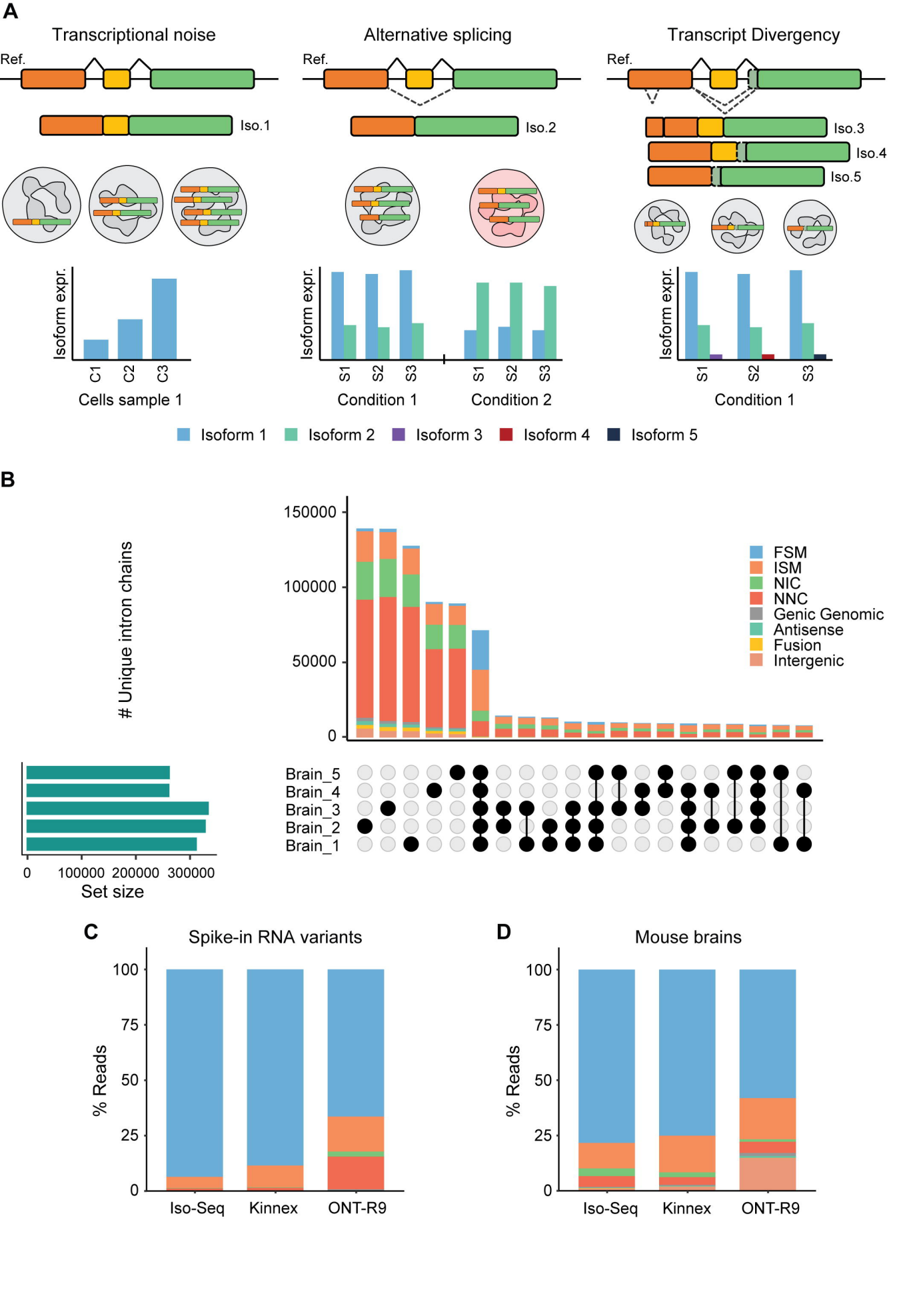
You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C, et al. 2015. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat Neurosci* **18**: 603–610.

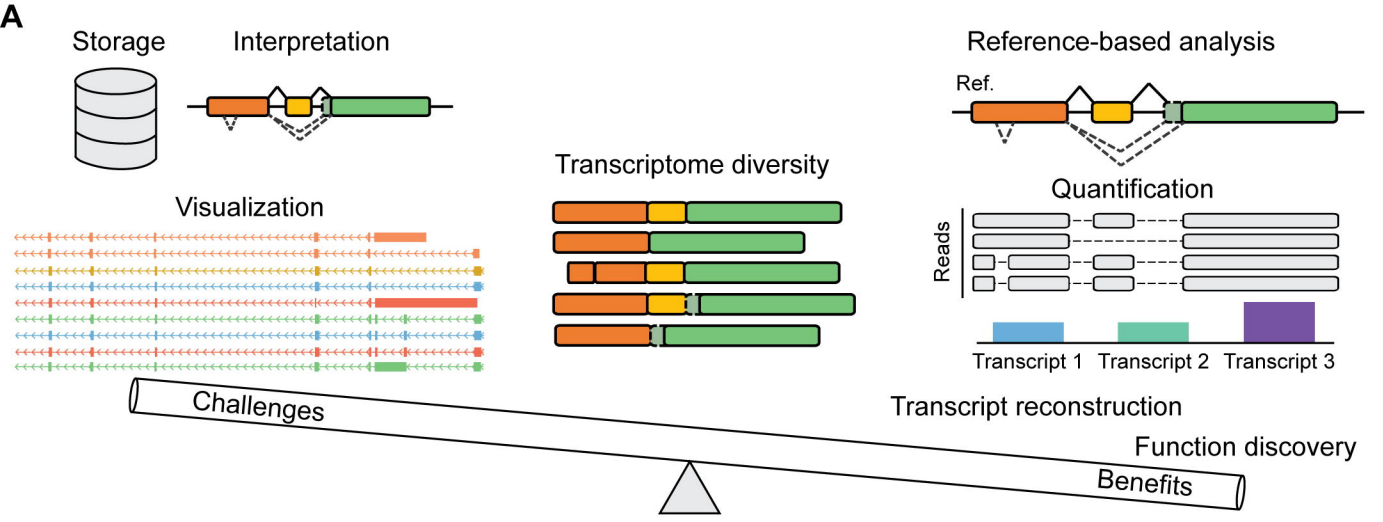
Zhang R, Kuo R, Coulter M, Calixto CPG, Entizne JC, Guo W, Marquez Y, Milne L, Riegler S, Matsui A, et al. 2022. A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biol* **23**: 149.

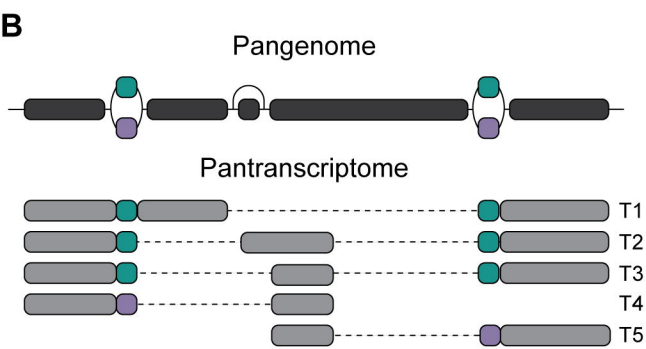
Fukasawa Y, Ermini L, Wang H, Carty K, Cheung MS. 2020. LongQC: a quality control tool for third generation sequencing long read data. *G3* **10**: 1193–1196.

Nanni A, Titus-McQuillan J, Bankole KS, Pardo-Palacios F, Signor S, Vlaho S, Moskalenko O, Morse A, Rogers RL, Conesa A, et al. 2024. Nucleotide-level distance metrics to quantify alternative splicing implemented in TranD. *Nucleic Acids Res* **5**: e28.


A
Reference
Full splice match (FSM)
Incomplete splice match (ISM)
Novel in catalog (NIC)
Novel not in catalog (NNC)
Genic Genomic
B
% Experimentally validated
100
80
60
40
20
0
14
37
3
30
3
39
1
37
0-1
2-8
9-17
18-21
WTC11 samples with supportive reads
not validated
validated
C
Unique intron chains
40
30
20
10
0
FSM
Genic Genomic
ISM
NIC
NNC
cDNA_PacBio
cDNA_ONT
R2C2_ONT
CapTrap_ONT
dRNA_ONT
CapTrap_PacBio
0 50100
Set size
Transcripts annotated by GENCODE


A
Transcriptional noise
Alternative splicing
Transcript Divergency
Ref.
Iso.1
Iso.2
Iso.3
Iso.4
Iso.5
Isoform expr.
C1
C2
C3
Cells sample 1
S1
S2
S3
Condition 1
Condition 2
Isoform 1
Isoform 2
Isoform 3
Isoform 4
Isoform 5
B
Unique intron chains
150000
100000
50000
0
FSM
ISM
NIC
NNC
Genic Genomic
Antisense
Fusion
Intergenic
Brain_5
Brain_4
Brain_3
Brain_2
Brain_1
0
100000
200000
300000
Set size
C
Spike-in RNA variants
% Reads
100
75
50
25
0
Iso-Seq
Kinnex
ONT-R9
D
Mouse brains
% Reads
100
75
50
25
0
Iso-Seq
Kinnex
ONT-R9





Transcript	Cell type	Disease	Haplotype	TD	...
T1	Neuron	No	1	No	...
T2	B cell	No	1	Yes	...
T3	Stem cell	No	1	No	...
T4	B cell	Yes	2	No	...
T5	Stem cell	Yes	2	No	...

```
SELECT * FROM ref_annotation WHERE Haplotype = 1
```