Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules

Camille Sessegolo[1,4], Corinne Cruaud[2], Corinne Da Silva[2], Marion Dubarry[2], Thomas Derrien[3], Vincent Lacroix[1,4], Jean-Marc Aury[2]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558 F-69622 Villeurbanne, France.
[2] Genoscope, Institut de biologie François-Jacob, Commissariat a l'Energie Atomique (CEA), Université Paris-Saclay, F-91057 Evry, France
[3] Univ Rennes, CNRS, IGDR (Institut de génétique et développement de Rennes) - UMR 6290, F-35000, Rennes,France
[4] EPI ERABLE - Inria Grenoble, Rhône-Alpes, France.

Abstract

Background: Our vision of DNA transcription and splicing has changed dramatically with the introduction of short-read sequencing. These high-throughput sequencing technologies promised to unravel the complexity of any transcriptome. Generally gene expression levels are well-captured using these technologies, but there are still remaining caveats due to the limited read length and the fact that RNA molecules had to be reverse transcribed before sequencing. Oxford Nanopore Technologies has recently launched a portable sequencer which offers the possibility of sequencing long reads and most importantly RNA molecules.
Results: Here we generated a full mouse transcriptome from brain and liver using the Oxford Nanopore device. As a comparison, we sequenced RNA (RNA-Seq) and cDNA (cDNA-Seq) molecules using both long and short reads technologies. In addition, we tested the TeloPrime preparation kit, dedicated to the enrichment of full-length transcripts.
Conclusions: Using spike-in data, we confirmed that expression levels are efficiently captured by cDNA-Seq using short reads. More importantly, Oxford Nanopore RNA-Seq tends to be more efficient, while cDNA-Seq appears to be more biased. We further show that the cDNA library preparation of the Nanopore protocol induces read truncation for transcripts containing stretches of A's. Furthermore, bioinformatics challenges remain ahead for quantifying at the transcript level, especially when reads are not full-length. Accurate quantification of processed pseudogenes also remains difficult, and we show that current mapping protocols which map reads to the genome largely over-estimate their expression, at the expense of their parent gene.
Keywords: cDNA-Seq, RNA-Seq, long-reads, Nanopore, gene expression, transcriptome, mouse.

Background

To date our knowledge of DNA transcription is brought by the sequencing of RNA molecules which have been first reverse transcribed (RT). This RT step is prone to skew the transcriptional landscape of a given cell and erase base modifications. The sequencing of these RT-libraries, that we suggest to call cDNA-Seq, has became popular with the introduction of the short-read sequencing technologies [25], [19]. Recently, the Oxford Nanopore Technologies (ONT) company commercially released a portable

sequencer which is able to sequence very long DNA fragments [7] and enable now the sequencing of complex genomes ([9], [4] and [21]). Moreover this device (namely MinION) is also able to sequence native RNA molecules [8] representing the first opportunity to generate genuine RNA-Seq data.

Furthermore, even if short-read technologies offer a deep sequencing and were helpful to understand the transcriptome complexity and to improve the detection of rare transcripts, they still present some limitations. Indeed, read length is a key point to address complex regions of a studied transcriptome. Depending on the evolutionary history of a given genome, recent paralogous genes can lead to ambiguous alignment when using short reads. In a same way, processed pseudogenes generated by the retrotranscription of RNAs back into the genomic DNA are challenging to quantify using short reads. In addition to sequencing technologies and bioinformatics methods, preparation protocols have a significant impact on the final result as they can incorporate specific biases [2],[24]. The generation of data rely on a high number of molecular and computational steps which evolve at a fast pace. These changes in the protocol generally modified the appearance of the data. As an example, data produced with protocols based on oligo-dT or random primers in the RT step show differences in how they cover transcripts[2].

Here we produce a complete transcriptome dataset, containing both cDNA-Seq and RNA-Seq, using the Illumina and Nanopore technologies. RNAs were sampled from brain and liver tissues of mice and were mixed with Lexogen's Spike-In RNA Variants (SIRVs) as a control for quantification of RNAs. We follow the protocols recommended by the manufacturers to generate the three following datasets on each tissue: Illumina cDNA-Seq, Nanopore cDNA-Seq and Nanopore RNA-Seq. The first was sequenced using the Illumina platform (TruSeq_SR) and the last two using the MinION device (PCS108_LR and RNA001_LR). Additionally, the second was also sequenced using the Illumina platform (PCS108_SR). The presence of this replicate enables us to clarify which differences are due to the preparation protocol and which are due to the sequencing platform in itself. Moreover, we generated a Lexogen's TeloPrime library on both tissues (TELO_LR), this preparation kit is an all-in-one protocol for generating full-length cDNA from total RNA (Figure 1).

Data Description

Biological material

We used total RNA extracted from mouse brain (Cat # 636601) and liver (Cat # 636603) from Clontech (Mountain View, CA, USA).

Libraries preparation

Illumina cDNA library

RNA-Seq library preparations were carried out from a mix of 250 ng total RNA and 0.25 ng Spike-in RNA Variant Control Mix E2 (Lexogen, Vienna, Austria) using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA), which allows mRNA strand orientation. Ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington, MA, USA), and libraries profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Illumina on Nanopore cDNA library

250 ng of cDNA prepared using the "cDNA-PCR Sequencing" protocol (see "Nanopore cDNA library" below) were sonicated to a 100- to 1000-bp size using the E220 Covaris instrument (Covaris,

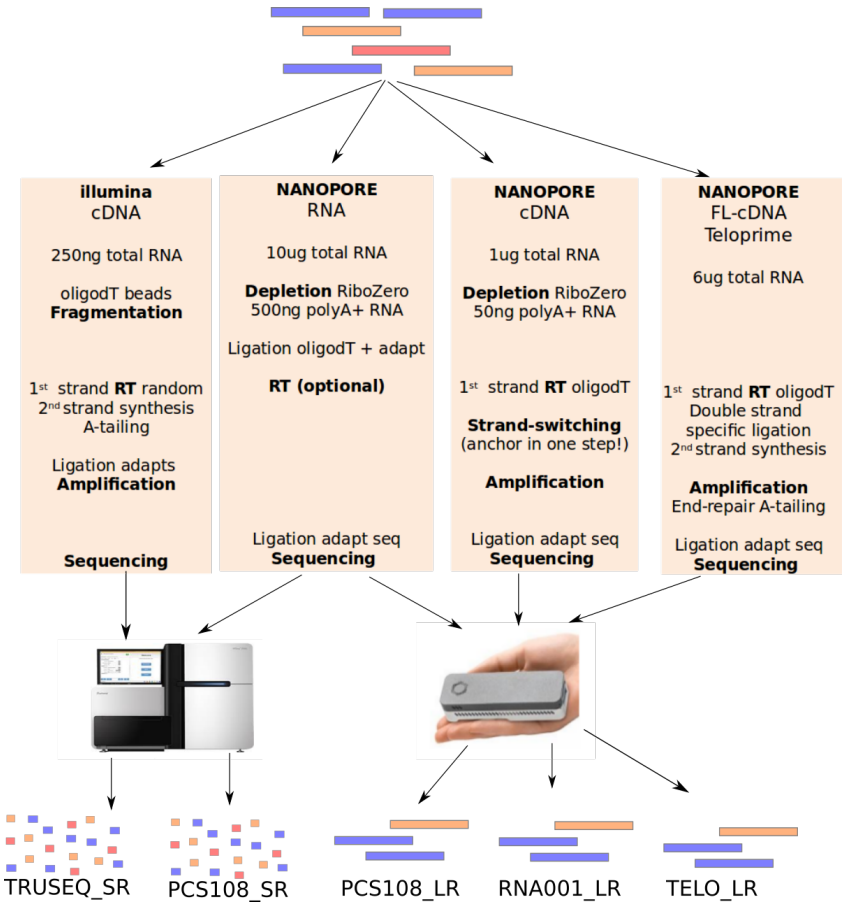

Figure 1: **Experimental design.** Five datasets have been generated on each tissue. Two were based on short-reads with the TruSeq protocol (TRUSEQ_SR) and the ONT library preparation (PCS108_SR) and the three others were based on long-reads with the ONT cDNA-Seq protocol (PCS108_LR), the ONT RNA-Seq protocol (RNA001_LR) and the Teloprime protocol (TELO_LR) have been used. (RT : retrotranscription).

Woburn, MA, USA). Fragments were end-repaired, then 3'-adenylated, and NEXTflex PCR free barcodes adapters (Bioo Scientific, Austin, TX, USA) were added using NEBNext Sample Reagent Module (New England Biolabs, Ipswich, MA, USA). Ligation products were amplified using Illumina adapter-specific primers and KAPA HiFi Library Amplification Kit (KapaBiosystems, Wilmington, MA, USA) and then purified with AMPure XP beads (Beckmann Coulter, Brea, CA, USA). Ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems), and libraries profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies).

Nanopore cDNA library

Total RNA was first depleted using the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina). RNA was then purified and concentrated on a RNA Clean Concentrator^{TM}-5 column (Zymo Research, Irvine, CA, USA). cDNA libraries were performed from a mix of 50ng RNA and 0.5 ng Spike-in RNA Variant Control Mix E2 (Lexogen) according to the Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) protocol "cDNA-PCR Sequencing" with a 14 cycles PCR (8 minutes for elongation time). ONT adapters were ligated to 650 ng of cDNA.

Table 1: Standard metrics of the generated datasets for both tissues : brain and liver.

	TruSeq_SR	PCS108_SR	PCS108_LR	RNA001_LR	TELO_LR
	Brain				
Number of reads	53,128,934	153,610,181	1,267,830	571,098	1,691,454
Cumulative size (Gb)	15.42	45.88	1.30	0.43	1.31
Average Size (bp)	151	151	1,028.94	758.05	775.77
N50 (bp)	151	151	1,283	1,357	896
Number reads >1Kb	NA	NA	522,422	154,735	389,468
Accession number	ERX2695239	ERX2695237	ERX2695238	ERX2695236	ERX2850744
	Liver				
Number of reads	49,270,153	178,019,939	3,043,572	418,102	2,668,975
Cumulative size (Gb)	14.16	53.43	3.30	0.34	2.64
Average Size (bp)	151	151	1,083.85	823.30	989.85
N50 (bp)	151	151	1,264	1,153	1,116
Number of reads >1Kb	NA	NA	1,218,569	117,047	884,008
Accession number	ERX2695241	ERX2695242	ERX2695243	ERX2695240	ERX2850745

Nanopore RNA library

RNA libraries were performed from a mix of 500ng RNA and 5ng Spike-in RNA Variant Control Mix E2 (Lexogen) according to the ONT protocol "Direct RNA sequencing". We performed the optional reverse transcription step to improve throughput, but cDNA strand was not sequenced.

Nanopore TeloPrime library

Three cDNA libraries were performed from 2μg total RNA for each RNA sample according to the TeloPrime Full-Length cDNA Amplification protocol (Lexogen). A total of 5 PCR were carried out with 30 to 40 cycles for the brain sample and 30 cycles for the liver sample. Amplifications were then pooled and quantified. Nanopore libraries were performed from respectively 560ng and 1000ng of cDNA using the SQK-LSK108 kit according to the Oxford Nanopore protocol.

Sequencing and reads processing

Illumina datasets

Both cDNA libraries, prepared with the TruSeq (TruSeq_SR) and Nanopore (PCS108_SR) protocols, were sequenced using 151 bp paired end reads chemistry on a HiSeq4000 Illumina sequencer (Table 1). After the Illumina sequencing, an in-house quality control process was applied to the reads that passed the Illumina quality filters. The first step discards low-quality nucleotides ($Q < 20$) from both ends of the reads. Next, Illumina sequencing adapters and primer sequences were removed from the reads. Then, reads shorter than 30 nucleotides after trimming were discarded. The last step identifies and discards read pairs that mapped to the phage phiX genome, using SOAP [18] and the phiX reference sequence (GenBank: NC_001422.1). These trimming and removal steps were achieved using in-house-designed software as described in [3].

Nanopore datasets

Nanopore libraries were sequenced using a MinION Mk1b with R9.4.1 (PCS108_LR and RNA001_LR) or R9.5 flowcells (TELO_LR). The data were generated using MinKNOW 1.11.5 and basecalled with Albacore 2.1.10 (Table 1).

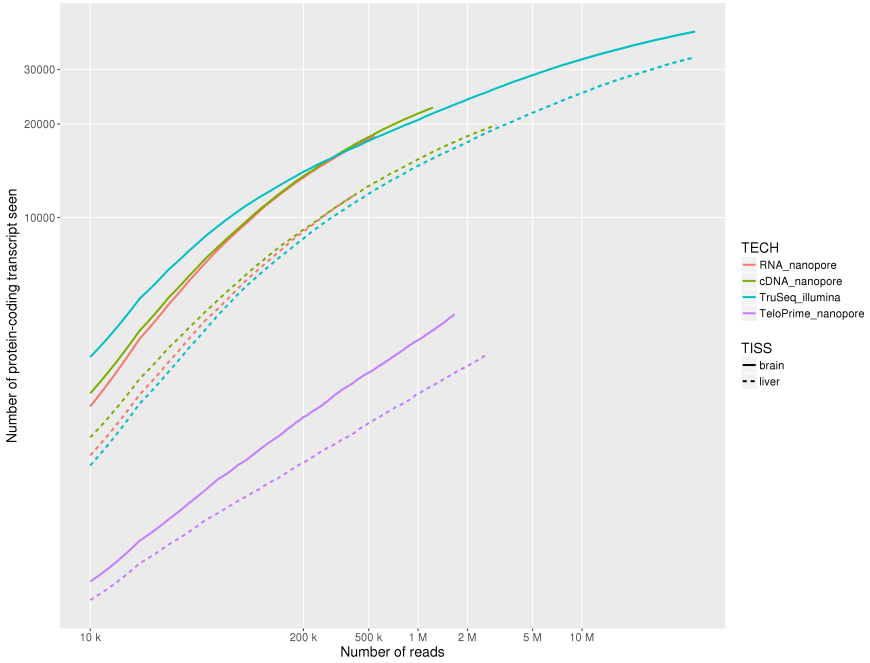

Figure 2: **Saturation curve.** Number of protein-coding transcripts seen by each technology at various sequencing depth. Solid and dashed lines correspond respectively to brain and liver samples.

Reads alignment

Long reads were mapped to the spike-in transcripts using Minimap2 (version 2.14) [17] (-ax map-ont) and we used the number of aligned reads as a proxy of the expression of a given transcript. Short reads were mapped to the spike-in transcripts using bowtie [14] and quantified using RSEM [16]. The quantification obtained is given in TPM (transcript per million).
We then assessed the mouse transcripts expression and mapped the long reads against the mouse transcripts (Ensembl 94) using Minimap2 (with the following options -ax map-ont and -uf for direct RNA reads). Supplementary alignments and reads aligned on less than 80% of their length were filtered out. Expression was directly approximated by the number of reads which mapped on a given transcript. Short reads were mapped to the reference genome (release Grcm38.p6) using STAR with the gtf option (annotation Ensembl 94). Gene expression levels (TPM) were then derived using Feature Counts (Subread package version 1.6.0).

Spliced alignment and error rate

The error rate of ONT reads is still around 10% and complicates the precise detection of splice sites. Mouse splice sites are often canonical, as observed when aligning reference annotation (coding genes from Ensembl 94) using BLAT, 98.5% of introns were GT-AG. Here minimap2 was able to detect only 80.7% of GT-AG introns when ONT RNA-Seq reads were used as input. Interestingly, the proportion of cannonical splice sites is lower when using ONT cDNA-Seq (67.7%). In fact ONT RNA-Seq reads are strand-specific which is of high value for the alignment and splice site detection. When using high quality sequences (coding genes from Ensembl 94) instead of reads, minimap2 retrieved 96.4% of GT-AG introns. These results show that the detection of splice sites using long but noisy reads is challenging and that dedicated aligners still need some improvements.

Table 2: Long reads coverage of single-isoform genes.

	PCS108_LR	RNA001_LR	TELO_LR
	Brain		
# of mapped reads	99,192	45,485	200,884
Avg coverage	61.18%	70.97%	76.37%
Median coverage	64.62%	83.84%	84.65%
Full-length reads (>80%)	40.26%	53.32%	60.18%
	Liver		
# of mapped reads	344,362	48,032	381,005
Avg coverage	73.86%	73.48%	79.09%
Median coverage	83.24%	84.57%	88.23%
Full-length reads (>80%)	55.63%	54.77%	60.02%

General comparison of sequencing technologies

RNA-Seq is a powerful method that provides a quantitative view of the transcriptome and the number of sequenced fragments is a key point to capture thoroughly the expression of genes. The Illumina technology is able to generate billions of short tags, and unsurprisingly allows to access a largest number of genes/transcripts. However with the same number of reads, both Illumina and Nanopore technology are able to uncover the same number of transcripts (Figure 2). As a comparison, the high-throughput PromethION device will be able to generate up to 60 millions of long reads and will compete directly with the Illumina technology.

Long reads sequencing offers the possibility to capture full-length RNAs. When looking at single isoform genes, we found that in average reads cover between 61 and 74% of the messenger RNAs (Table 2). But even though this horizontal coverage is quite high, the proportion of reads that covered more than 80% of the transcript remains low (near 55% except for cDNA-Seq of the brain sample). RNA degradation can obviously explain a proportion of these fragmented reads, and it has been shown more recently that a software artifact may truncate reads (around 20%) during the base-calling process [1].

Improving the proportion of full-length reads

The proportion of full-length transcripts can be improved by using a dedicated library preparation protocol. Here we tested the TeloPrime amplification kit, commercialized by the Lexogen company. This protocol is selective for full-length RNA molecules that are both capped and polyadenylated. Using this protocol we were able to slightly improve the proportion of full-length reads (which covered at least 80% of a given transcript, Table 2). However, in return, we captured a lower number of genes, lowering the interest of such a protocol in most applications (Figure 2).

Sequencing biases of transcripts containing long homopolymers of A's

Since cDNA synthesis is initiated with a poly-dT primer, a relevant question is whether transcripts which internally contain stretches of A's are correctly sequenced. We find that using cDNA-Seq, cDNA molecules stemming from such transcripts are often truncated, resulting in reads covering only the 3' end of the transcripts. As an example, the *Ncs1* gene contains a stretch of 15 A's and ONT cDNA-Seq reads are systematically interrupted at the location of this stretch of A's (Figure 3c, tracks 2 and 3), while this is not the case for Illumina Truseq (Figure 3c, track 4) and ONT RNA-Seq (Figure 3c, track 1).This gene is not an exception and a systematic analysis of all transcripts covered by at least 20 reads

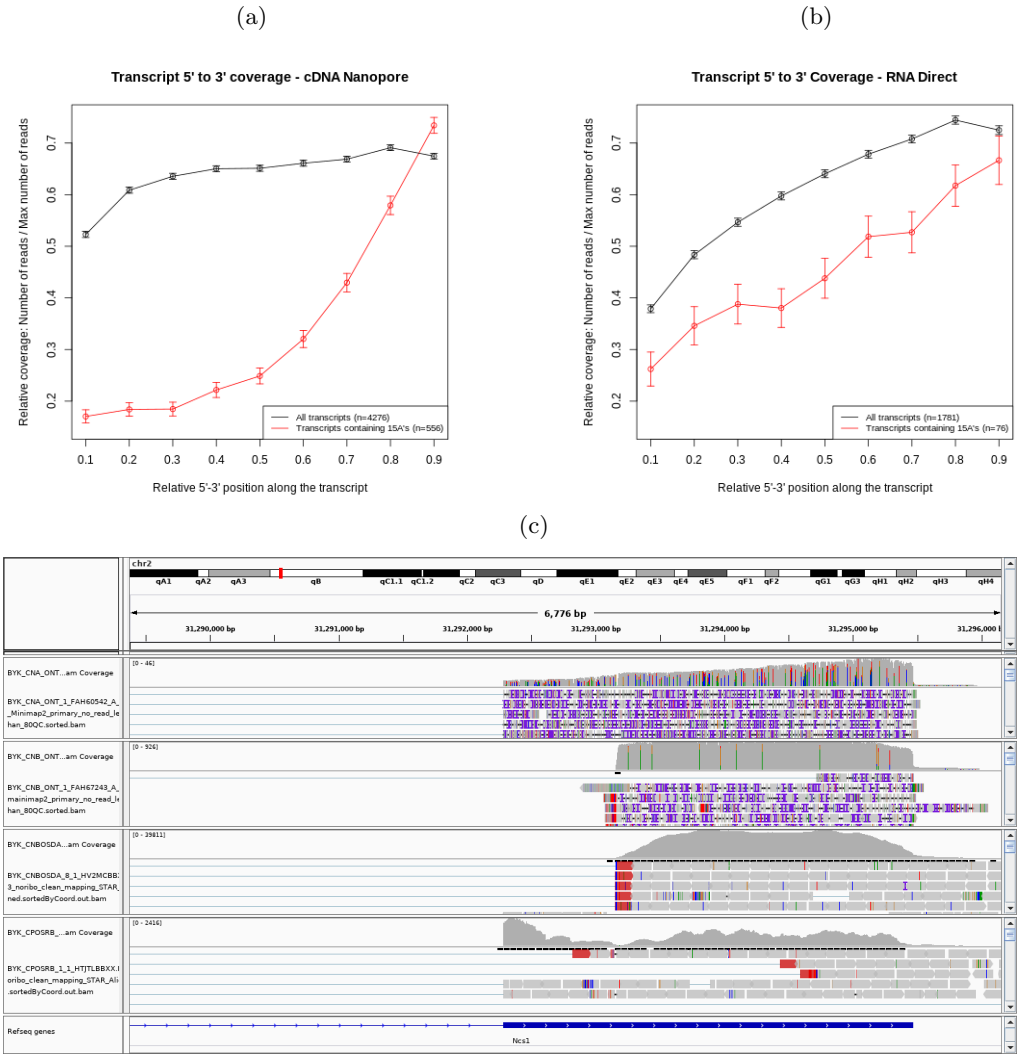

Figure 3: **Truncated reads** (a) Relative coverage of transcripts for the ONT cDNA-Seq dataset for transcripts covered by at least 20 reads. Error bars represent the standard error of the observed coverage. Transcripts containing stretches of at least 15 A's are less covered in 5' than others transcripts. It indicates that these transcripts are covered by a high proportion of truncated reads.(b) Relative coverage of transcripts for the ONT RNA-Seq dataset for transcripts covered by at least 20 reads. Error bars represent the standard error of the observed coverage. Contrary to the cDNA dataset transcripts containing stretches of at least 15 As are not truncated. Coverage of all the transcripts decrease because many reads are not full-length. (c) Example of *Ncs1* visualized with IGV. Truncated reads are in tracks 2 (ONT cDNA-Seq) and 3 (Illumina, Nanopore protocol). Non-truncated reads are in tracks 1 (ONT RNA-Seq) and 4 (Illumina Truseq). The region where the truncation occurs is a polyA stretch.

and containing a stretch of at least 15 A's reveals that read truncation is common (Figure 3a). This bias is also present in other published Nanopore RNA-Seq dataset [27] (Gene *Mettl2b*, Supplementary figure 1), but has remained overseen so far. It however concerns at least 5% of transcripts in mouse and human, as can be estimated from RNA-Seq data. The estimation is even higher and reaches 10% when

using cDNA-Seq data. The true expression level of these transcripts is however not easy to capture and might be over-estimated by the ONT cDNA protocol (Supplementary Figure 1a) compared to the ONT RNA-Seq protocol (Supplementary Figure 1b). Overall, the ONT cDNA-Seq protocol produces reads that are shorter and more numerous for this class of transcripts. This bias seems to be associated to the use of a polydT primer, and could therefore also affect Pacbio cDNA-Seq data. Although it had been searched for in previous work [13], it may have been largely underestimated because the authors were then focusing on stretches of 20 A's.

Evaluation of the accuracy of the gene expression quantification using spike-in data

In order to assess which protocol was best to quantify gene expression, we analysed the spike-ins contained in the brain datasets. Since we exactly know which transcripts are present in the sample, the quantification is rather straightforward. The best quantifications were obtained for the ONT RNA-Seq (ρ =0.86) and Illumina Truseq (ρ =0.80) protocols (Figure 4). In contrast, cDNA-Seq (sequenced using Illumina or Nanopore) produced more unprecise quantifications (ρ =0.53 and ρ =0.66). For all protocols, the precision of the estimation was poorer for weakly expressed transcripts, an expected result.

Although lexogen spike-in have been used to evaluate the quantification obtained with ONT cDNA-Seq [26] or ONT RNA-Seq [8] protocol separately, we are the first to compare the quantification obtained with ONT cDNA-Seq, RNA-Seq and Illumina cDNA-Seq. Our experimental design allows us to compare the effect of the sequencing platform and of the library preparation protocol independently.

In order to assess the quality of the quantification in a more realistic context where we do not know which transcripts are present in the sample, we also mapped the reads to a modified set of transcripts corresponding either to an over-annotation or an under-annotation (as provided by Lexogen). In both cases, the correlations were overall poorer than before, but the order was maintained, with ONT RNA-Seq and then Illumina cDNA-Seq being the more faithful protocols (Supplementary Figures 2 and 3).

Quantification of the expression level of mouse genes

Figure 5 summarizes the correlations in terms of transcript quantification of our datasets. Given that with the spike in, the best quantification were obtained with ONT RNA-Seq, we compared the quantifications obtained with this protocol with the ones obtained with the ONT and Illumina cDNA-Seq protocols. Comparing the Illumina cDNA-Seq and the ONT RNA-Seq protocols we obtain a spearman coefficient of correlation $\rho = 0.51$. The correlation is lower in the liver sample (Supplementary Figure 4) probably because of a lower number of RNA-seq reads and a shorter read length.

Comparing the ONT RNA-Seq and cDNA-Seq quantification, we obtain a higher correlation ($\rho = 0.75$), suggesting that read length strongly influences transcript quantification. Indeed, in the comparison between Illumina cDNA-Seq and ONT RNA-Seq dataset, the lack of correlation comes from one main cause. Discriminating transcripts of a same gene that share common sequences with short reads is difficult. Longer reads are clearly helpful, however they do not always enable to discriminate transcripts. Indeed, in the case where a read only covers the 3' end of a transcript, and not the full length, it may be ambiguously assigned to several transcripts.

For example, for the *Swi5* gene, although several rare (lowly expressed) transcripts are seen only with Illumina, the other ones are harder to quantify (red dots in figure 5a). RSEM uses the unique part of each transcript to proportionally allocate the reads that mapped equally on the common part of the transcript. In the case where a transcript has no read which uniquely maps to it, its expression cannot be computed and is set to 0. This is the case for the transcript ENSMUST00000050410 (*Swi5*-201, Supplementary figure 5) of *Swi5*, whose expression is underestimated (0 TPM), an issue already seen in spike-ins (SIRV-311). Conversely, some transcripts are underestimated by ONT RNA-Seq. This is

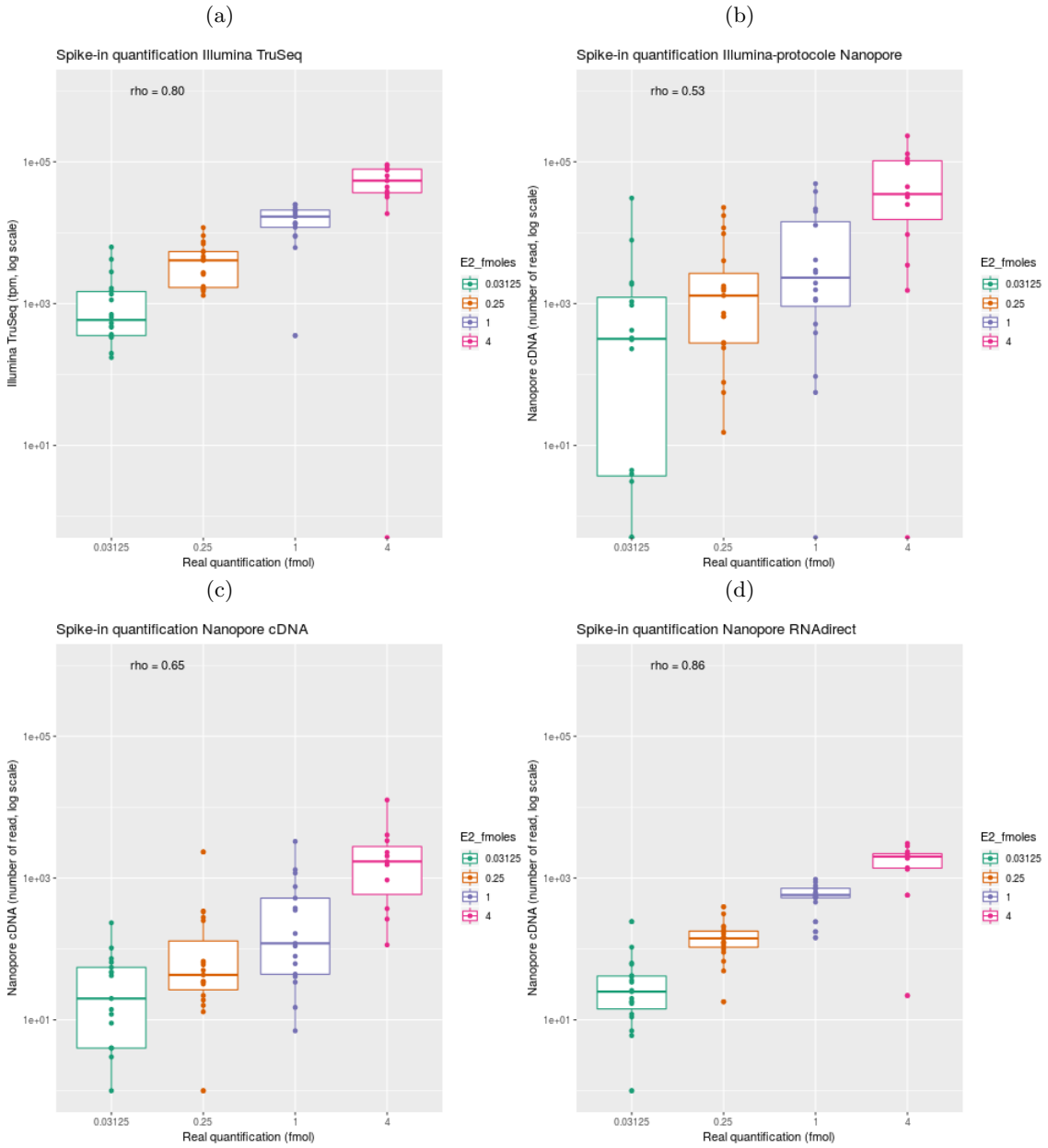

Figure 4: **Evaluation of quantification using the SIRV E2 spike-in mix.** Reads were mapped against the SIRV transcriptome and quantifications were computed at the transcript level. The observed quantifications are correlated with the known theoretical quantifications of the spike in. (a) Correlation obtained for Illumina cDNA-Seq (Spearman's $\rho = 0.80$, n=327386). (b) Correlation obtained for Illumina with the ONT cDNA-Seq protocol (Spearman's ρ =0.53, n=7981494 reads). (c) Correlation obtained for ONT cDNA-Seq (Spearman's $\rho = 0.65$, n=42559 reads). (d) Correlation obtained for ONT RNA-Seq (Spearman's $\rho = 0.86$, n=35513 reads).

the case of *Swi5*-204, whose unique region is located at the 5' end of the gene, and is therefore poorly covered by long reads.

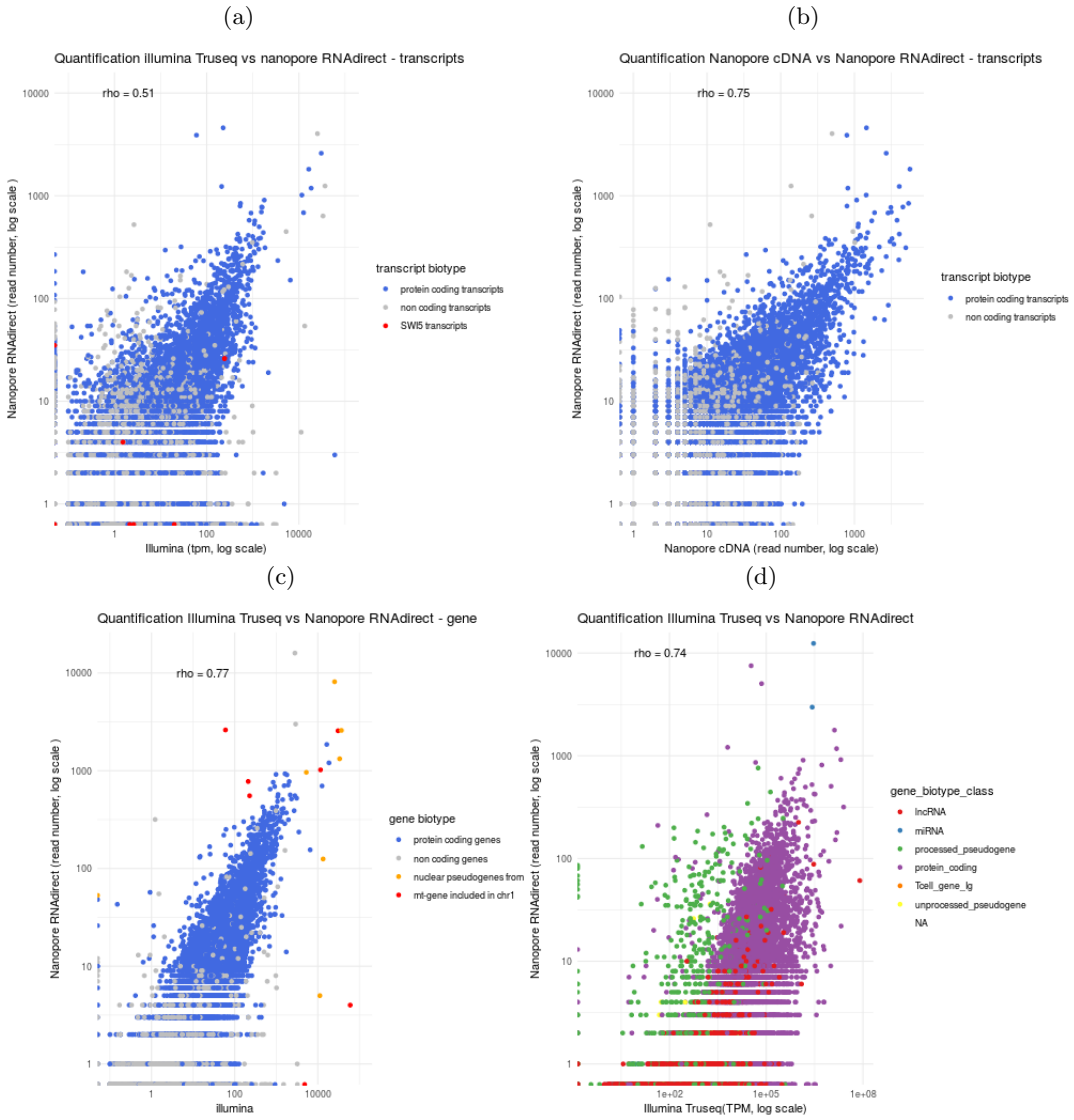

Figure 5: **Comparison of quantifications.** Transcripts or genes annotated as protein-coding are in blue. Spearman's ρ has been computed for all transcripts or genes. (a) Comparison of ONT RNA-Seq and Illumina cDNA-Seq quantifications at the transcript level (Spearman's $\rho = 0.51$). Red points correspond to the transcripts of the *Swi5* gene. (b) Comparison of ONT RNA-Seq and ONT cDNA-Seq quantifications at the transcript level (Spearman's $\rho = 0.75$). (c) Comparison of ONT RNA-Seq and Illumina cDNA-Seq quantifications at the gene level (transcript quantification were summed for each gene, Spearman's $\rho = 0.77$). Red points correspond to the processed pseudogene produced by the introgression of the mitochondrial chromosome on the chromosome 1 and orange points corresponds to the mitochondrial parent genes. (d) Reads were mapped against the mouse reference genome and quantifications computed at gene level. We compared the ONT RNA-Seq and the Illumina cDNA-Seq protocols (Spearman's $\rho = 0.74$). Green points correspond to processed pseudogenes.

To avoid the difficult step of correctly assigning a read to a transcript, we summed the quantification of all transcripts for each gene. Figure 5c shows the quantification at gene level. As reported in other

papers [8] [5] [22] the correlation at gene level is quite good (ρ =0.78). However inter-genes repeats remains a cause of mis-quantified genes. For example, a large part of the mitochondrial chromosome had been recently integrated in the mouse's chromosome 1 [15]. As a consequence, 7 genes are present in 2 copies in the genome, one copy annotated as functional on the mitochondrial chromosome and another one, annotated as pseudogene on chromosome 1 (shown in red in figure 5c). Since this integration is recent, the copies did not diverge yet. They are therefore difficult to quantify due to multimapping, even when using long reads, since the repeat is larger than the full transcript. More generally, the quantification of repeat-containing genes is challenging. Long reads are particulary useful for quantifying these genes, in particular long non coding RNAs, which are enriched in transposable elements. Mapping long reads to the correct location remains however challenging, in particular for processed pseudogenes.

Quantification of processed pseudogenes

These are particular cases of pseudogenes which come from the retrotranscription and reintegration in the genome of one of the transcript of their parent gene[11]. After their integration, they have no intron and, without any selective pressure, they diverge from their parent gene proportionally with their age. Some of them are expressed [6] and are annotated as transcribed processed pseudogenes although the vast majority of pseudogenes are not expressed [11]. Correctly assigning the reads to the parent gene and not the pseudogene is not trivial.

Figure 5d shows that mapping long reads to the genome with Minimap 2 (-ax splice) (as used in [27]) results in the mis-quantification of processed pseudogenes. The expression of most of them is over-estimated by the ONT RNA-Seq protocol (this is also the case of ONT cDNA-Seq). It can be explained by two main reasons.

First, it can come from the fact that if a mapper has to choose between two genomic locations, one with gaps (introns of the parent gene), and one with no gaps (the processed pseudogene), it will tend to favour the gapless mapping, because it is easier to find. We note that the scoring system of minimap2 consists in selecting the max-scoring sub-segment, excluding introns, and therefore not explicitly favouring the gapless mapping. However, this requires that splice sites are correctly identified in the first place, a task which remains difficult with noisy long reads.

A second reason explaining the overestimation of processed pseudogenes is related to polyA tails. Processed pseudogenes originate from transcripts which contained a polyA tail, which was then integrated in the genome, downstream the pseudogene. Many of the ONT reads originating from the parent gene also contain this polyA tail, favoring the alignment at the processed pseudogene genomic location. The alignment will be longer thanks to the polyA tail. An example is shown in figure 6. The processed pseudogene *Rpl17-ps8* differs from its parent gene by two bases (A to G at the position chrX:96,485,078 and A to G at the position chrX:96,485,267). These divergences are marked in red in the figure. At these two positions we observe that reads differ from the reference genome : they have a G instead of an A. This means that these reads come from the parent gene and we mistakenly aligned them onto the pseudogene because it is intronless and contains a polyA tail. We therefore strongly recommend to map reads on the reference transcriptome and not on the genome, as reference transcripts do not contain introns, nor polyA tails.

Quantification of genes overlapping transposable elements

Another example of repeat-associated gene biotype is given by the long non-coding RNAs (lncRNAs) which are highly enriched in transposable elements (TEs). These TEs are sometimes considered as the functional domains of lncRNAs [10], and it has been estimated that 66% of mouse lncRNA overlap at least one TE [12]. Our experimental design allows us to assess the impact of read lengths on non-coding (versus protein-coding) gene quantifications for different levels of TE coverage. Given that lncRNAs are lowly expressed, for this specific analysis, we restricted to cDNA-Seq and did not apply our 80% query coverage filter. Using annotated TEs from the RepeatMasker database [23],

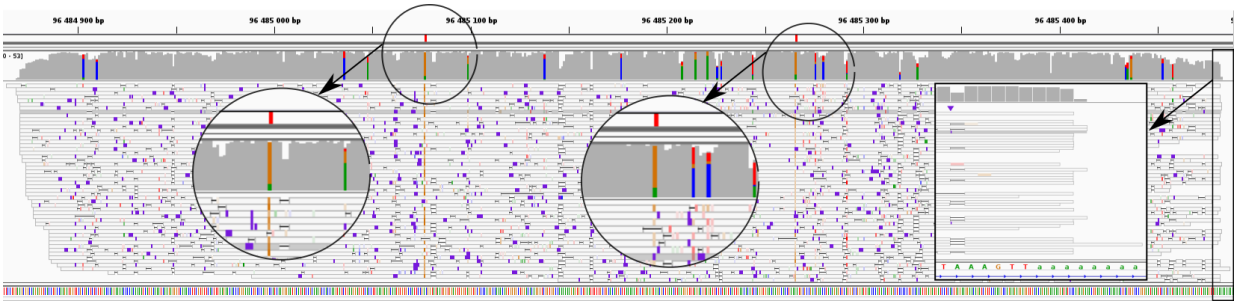

Figure 6: **Example of a processed pseudogene whose expression is overestimated : *Rpl17-ps8* (retro Rpl17)** Alignment visualization with IGV of *Rpl17-ps8*. The positions of divergence between Rpl17-ps8 and Rpl17 are shown in red in the first track. Second track is ONT RNA-Seq coverage, third track is ONT RNA-Seq reads. Colored positions in the coverage track correspond to mismatches. Most reads contain mismatches at the exact position of the divergences with the parent gene. They are therefore incorrectly mapped, partly because they overlap the polyA tail which is integrated in the genome downstream the pseudogene.

we classified lncRNAs and mRNAs based on their TE coverage in 4 categories (with the "0%" class corresponding to genes without any exonic-overlapping TE and conversely, the class of ">66-100%" for genes highly enriched in exonic TE) (Supplementary Figure 6a). For each expressed gene, we further computed the ratio between Nanopore cDNA versus Illumina TruSeq gene quantifications with respect to their TE categories (Supplementary Figure 6a). As expected, the higher the TE content of a gene, the larger the difference in quantification between long and short-read sequencing technologies. Although this tendency is observed both for protein-coding and lncRNAs biotypes, lncRNAs are more impacted given that they are more prone to be enriched in TEs. One interesting example is given by the known imprinted lncRNA *KCNQ1OT1* (ENSMUSG00000101609) which is specifically expressed from the paternal allele in opposite direction to the *KCNQ1* protein-coding gene [20] (Supplementary Figure 6b). About 41% of the *KCNQ1OT1* transcript sequence is composed of TE elements and its quantification using Illumina TruSeq versus Nanopore cDNA protocols highlights contrasting values (TPM = 0.36 and Nanopore cDNA = 118).

Conclusion

We find that ONT long-reads can be used to quantify transcripts and that the most faithful quantification was obtained with the ONT RNA-Seq protocol. While the cDNA-seq data was more biased, in particular for transcripts containing long stretches of A's.
Quantifying transcripts and not genes is still challenging, and requires the development of dedicated bioinformatics methods, in particular when the annotation is not complete, or for non-model species where no annotation is available. Determining the abundance of processed pseudogenes is challenging, and current mapping protocols largely overestimate their expression. This overestimation is especially true when a reference genome is used instead of a reference transcriptome. The limitation is however that, even for model species, there are many transcripts which have not yet been annotated. Reads stemming from them are therefore unmapped, or incorrectly assigned to alternative transcripts.
Improvements in the technology and library preparation protocol to obtain more reads and more full-length reads are also expected to be very helpful in obtaining precise quantification of all genes and transcripts. The recent launch of the PromethION device will allow a deep sequencing of transcriptomes which should enable to overcome the limitations of the MinION device.

Availability of supporting data

The Illumina and MinION data are available in the European Nucleotide under the following accession number PRJEB27590.

Competing interests

The authors declare that they have no competing interests. CC and JMA are part of the MinION Access Programme (MAP) and JMA received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

Acknowledgements

The authors are grateful to Oxford Nanopore Technologies Ltd for providing early access to the MinION device through the MAP, and we thank the staff of Oxford Nanopore Technology Ltd for technical help.

Funding

This work was supported by French National research agency (ANR project ANR-16-CE23-0001 'ASTER'), the INRIA, the Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and France Génomique (ANR-10-INBS-09-08).

References

[1] Angela Brooks (Nanopore RNA Consortium) - Native RNA sequencing of human polyadenylated transcripts. https://nanoporetech.com/resource-centre/native-rna-sequencing-human-polyadenylated-transcripts, 2018. [Accessed 25 Fev 2019].

[2] Adriana Alberti, Caroline Belser, Stéfan Engelen, Laurie Bertrand, Céline Orvain, Laura Brinas, Corinne Cruaud, Laurène Giraut, Corinne Da Silva, Cyril Firmo, Jean-Marc Aury, and Patrick Wincker. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics*, 15(1):912, oct 2014.

[3] Adriana Alberti, Julie Poulain, Stefan Engelen, Karine Labadie, Sarah Romac, Isabel Ferrera, Guillaume Albini, Jean-Marc Aury, Caroline Belser, Alexis Bertrand, Corinne Cruaud, Corinne Da Silva, Carole Dossat, Frédérick Gavory, Shahinaz Gas, Julie Guy, Maud Haquelle, E'krame Jacoby, Olivier Jaillon, Arnaud Lemainque, Eric Pelletier, Gaëlle Samson, Mark Wessner, Pascal Bazire, Odette Beluche, Laurie Bertrand, Marielle Besnard-Gonnet, Isabelle Bordelais, Magali Boutard, Maria Dubois, Corinne Dumont, Evelyne Ettedgui, Patricia Fernandez, Espérance Garcia, Nathalie Giordanenco Aiach, Thomas Guerin, Chadia Hamon, Elodie Brun, Sandrine Lebled, Patricia Lenoble, Claudine Louesse, Eric Mahieu, Barbara Mairey, Nathalie Martins, Catherine Megret, Claire Milani, Jacqueline Muanga, Céline Orvain, Emilie Payen, Peggy Perroud, Emmanuelle Petit, Dominique Robert, Murielle Ronsin, Benoit Vacherie, Silvia G. Acinas, Marta Royo-Llonch, Francisco M. Cornejo-Castillo, Ramiro Logares, Beatriz Fernández-Gómez, Chris Bowler, Guy Cochrane, Clara Amid, Petra Ten Hoopen, Colomban De Vargas, Nigel Grimsley, Elodie Desgranges, Stefanie Kandels-Lewis, Hiroyuki Ogata, Nicole Poulton, Michael E. Sieracki, Ramunas Stepanauskas, Matthew B. Sullivan, Jennifer R. Brum, Melissa B. Duhaime, Bonnie T. Poulos, Bonnie L. Hurwitz, Silvia G. Acinas, Peer Bork, Emmanuel Boss, Chris Bowler, Colomban De Vargas, Michael Follows, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone,

Olivier Jaillon, Stefanie Kandels-Lewis, Lee Karp-Boss, Eric Karsenti, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen Raes, Christian Sardet, Michael E. Sieracki, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Shinichi Sunagawa, Patrick Wincker, Stéphane Pesant, Eric Karsenti, Patrick Wincker, Eric Karsenti, and Patrick Wincker. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data*, 4:170093, aug 2017.

[4] Caroline Belser, Benjamin Istace, Erwan Denis, Marion Dubarry, Franc-Christophe Baurens, Cyril Falentin, Mathieu Genete, Wahiba Berrabah, Anne-Marie Chèvre, Régine Delourme, Gwenaëlle Deniot, France Denoeud, Philippe Duffé, Stefan Engelen, Arnaud Lemainque, Maria Manzanares-Dauleux, Guillaume Martin, Jérôme Morice, Benjamin Noel, Xavier Vekemans, Angélique D'Hont, Mathieu Rousseau-Gueutin, Valérie Barbe, Corinne Cruaud, Patrick Wincker, and Jean-Marc Aury. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, 4(11):879–887, nov 2018.

[5] Ashley Byrne, Anna E Beaudin, Hugh E Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M Dubois, E Camilla Forsberg, Mark Akeson, and Christopher Vollmers. ARTICLE Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, 8, 2017.

[6] Francesco Nicola Carelli, Takashi Hayakawa, Yasuhiro Go, Hiroo Imai, Maria Warnefors, and Henrik Kaessmann. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome research*, 26(3):301–14, mar 2016.

[7] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, may 2016.

[8] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J Heron, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, jan 2018.

[9] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, jan 2018.

[10] R. Johnson and R. Guigo. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, 20(7):959–976, jul 2014.

[11] Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, 10(1):19–31, jan 2009.

[12] David Kelley and John Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology*, 13(11):R107, 2012.

[13] Richard I. Kuo, Elizabeth Tseng, Lel Eory, Ian R. Paton, Alan L. Archibald, and David W. Burt. Normalized long read rna sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics*, 18(1):323, Apr 2017.

[14] Ben Langmead. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics*, Chapter 11:Unit 11.7, dec 2010.

[15] Dario Leister and Erik Richly. NUMTs in Sequenced Eukaryotic Genomes. *Molecular Biology and Evolution*, 21(6):1081–1084, 06 2004.

[16] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, dec 2011.

[17] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[18] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, aug 2009.

[19] Doron Lipson, Tal Raz, Alix Kieu, Daniel R Jones, Eldar Giladi, Edward Thayer, John F Thompson, Stan Letovsky, Patrice Milos, and Marie Causey. Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnology*, 27(7):652–658, jul 2009.

[20] D. Mancini-DiNardo, Scott J S Steele, John M Levorse, Robert S Ingram, and Shirley M Tilghman. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & Development*, 20(10):1268–1282, may 2006.

[21] Maximilian H-W Schmidt, Alexander Vogel, Alisandra K Denton, Benjamin Istace, Alexandra Wormit, Henri van de Geest, Marie E Bolger, Saleh Alseekh, Janina Maß, Christian Pfaff, Ulrich Schurr, Roger Chetelat, Florian Maumus, Jean-Marc Aury, Sergey Koren, Alisdair R Fernie, Dani Zamir, Anthony M Bolger, and Björn Usadel. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant cell*, 29(10):2336–2348, oct 2017.

[22] Masahide Seki, Eri Katsumata, Ayako Suzuki, Sarun Sereewattanawoot, Yoshitaka Sakamoto, Junko Mizushima-Sugano, Sumio Sugano, Takashi Kohno, Martin C Frith, Katsuya Tsuchihara, and Yutaka Suzuki. Evaluation and application of rna-seq by minion. *DNA Research*, page dsy038, 2018.

[23] Maja Tarailo-Graovac and Nansheng Chen. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, Chapter 4:Unit 4.10, mar 2009.

[24] Erwin L. van Dijk, Yan Jaszczyszyn, and Claude Thermes. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, 322(1):12–20, mar 2014.

[25] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, jan 2009.

[26] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6:100, feb 2017.

[27] Rachael E Workman, Alison Tang, Paul S. Tang, Miten Jain, John R Tyson, Philip C Zuzarte, Timothy Gilpatrick, Roham Razaghi, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen Jones, Terrance P Snutch, Nicholas James Loman, Benedict Paten, Matthew W Loose, Jared T Simpson, Hugh E. Olsen, Angela N Brooks, Mark Akeson, and Winston Timp. Nanopore native rna sequencing of a human poly(a) transcriptome. *bioRxiv*, 2018.