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

Jean-Marc Aury, Camille Sessegolo, Corinne Cruaud, Corinne Da Silva, Audric Cologne, Marion Dubarry, Thomas Derrien, Vincent Lacroix

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

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

[1] Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622, Villeurbanne, France. EPI ERABLE - Inria Grenoble, Rhône-Alpes, 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

@J_M_Aury jmaury@genoscope.cns.fr

The entire dataset (fastq and bam files) is available from the following website:
http://www.genoscope.cns.fr/ont_mouse_rna/

Introduction

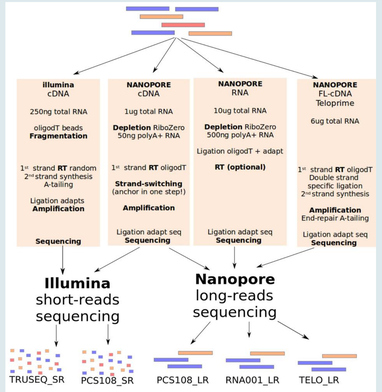
To date our knowledge of DNA transcription is brought by the sequencing of RNA molecules which have been first reverse transcribed (RT). This RT step is prone to skew the transcriptional landscape of a given cell and erase base modifications. The sequencing of these RT-libraries, that we suggest to call cDNA-Seq, has became popular with the introduction of the short-read sequencing technologies[1,2]. Recently, the Oxford Nanopore Technologies (ONT) company commercially released a portable sequencer which is able to sequence native RNA molecules[3] 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.

Experimental design

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 E2 spike-In RNA Variants (SIRVs[4]) 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). From the brain tissue, we generated biological (two brain RNA samples, C1 and C2) and technical replicates (R1 and R2) for the three datasets.

Additionally, the second was also sequenced using the Illumina platform (PCS108_SR). This 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[5] library on both tissues (TELO_LR), this preparation kit is an all-in-one protocol for generating full-length cDNA from total RNA.



Brain sample 1	C1R1	X	X	X	X	X
Brain sample 2	C2R1	X		X	X	
	C2R2	X		X	X	
Liver Sample 1	C1R1	X	X	X	X	X

Experimental design. Five protocols have been used 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). (RT: Reverse Transcription). For the brain, two biological replicates, C1 and C2, have been generated and two technical replicates, R1 and R2, have been generated for the second biological replicate. For the first biological replicate all the five protocols were used whereas the TRUSEQ_SR, the PCS108_LR and the RNA001_LR were used for the second biological replicates.

Overview of the sequencing data

Overview of the ONT data from brain samples

RNA Samples	PCS108_LR			RNA001_LR			TELO_LR
	C1	C2		C1	C2		C1
	R1	R1	R2	R1	R1	R2	R1
Number of reads	1,267,830	5,834,882	3,003,844	571,098	364,041	210,654	1,691,454
Cumulative size (Gb)	1.30	7.03	3.28	0.43	0.38	0.20	1.31
Average Size (bp)	1,028.94	1,204.54	1,091.85	758.05	1,032.10	957.17	775.77
N50 (bp)	1,283	1,749	1,591	1,357	1,492	1,417	896
Number reads >1 Kb	522,422	2,869,633	1,339,489	154,735	141,970	73,232	389,468
Full-length reads >80% single isoform	40.25%	41.08%	40.73%	53.31%	57.19%	60.61%	60.18%
Accession number	ERX2695238	ERX3387950	ERX3387952	ERX2695236	ERX3387949	ERX3387951	ERX2850744

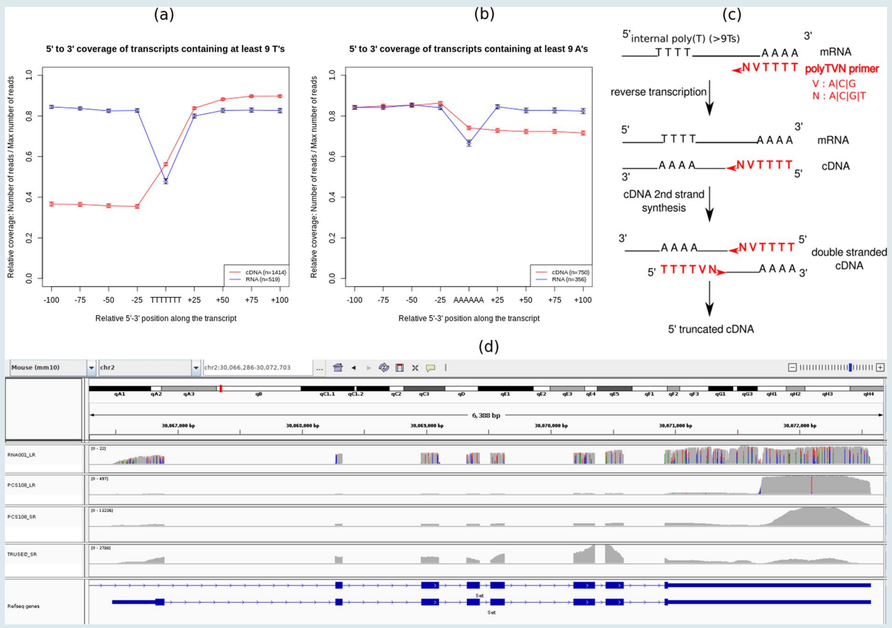
Overview of the Illumina data from brain samples

RNA Samples	TruSeq_SR			PCS108_SR
	C1	C2		C1
	R1	R1	R2	R1
Number of reads	53,128,934	41,562,993	45,719,216	153,610,181
Cumulative size (Gb)	15.42	12.36	13.60	45.88
Read Size (bp)	151	151	151	151
Accession number	ERX2695239	ERX3387947	ERX3387948	ERX2695237

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

Sequencing biases of transcripts containing internal runs of poly(T)

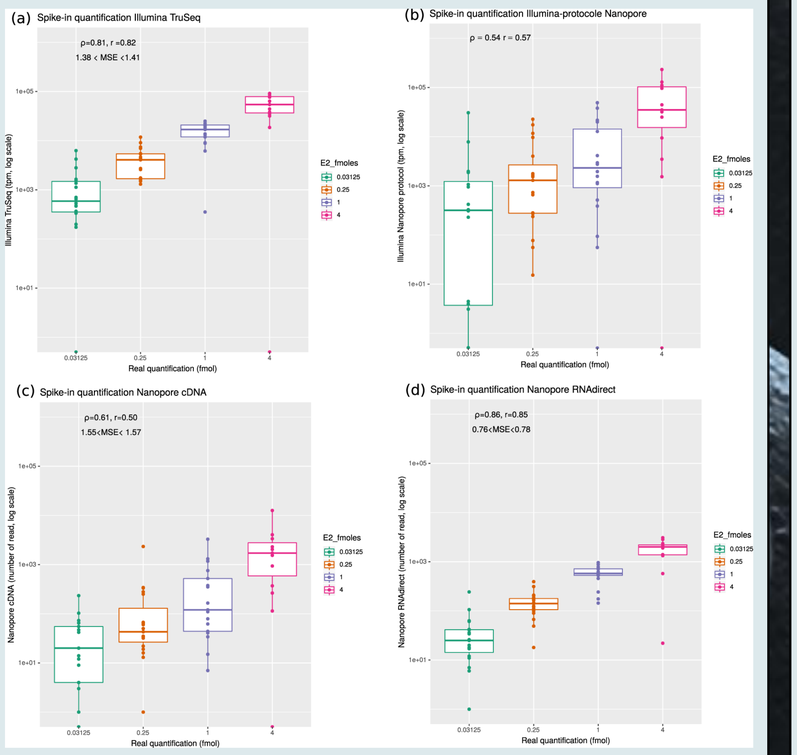
Since cDNA synthesis is initiated with an anchored poly-dT primer (poly-TVN), a relevant question is whether transcripts containing internal runs of poly(A) or poly(T) are correctly sequenced. This bias has remained unreported so far, but it is also present in other published Nanopore dataset[6]. It however concerns a large fraction of transcripts, 27% of transcripts expressed with at least one read in mouse brain contain at least 9 T's. Importantly, the bias not only affects read length, but also transcript quantification. In mouse brain, 35% of cDNA-Seq reads map to transcripts with at least 9 T's, compared to 14% of RNA-Seq reads. This suggests that the abundance of these transcripts is over-estimated when using cDNA-Seq, at the expense of the other transcripts.



Truncated reads.
(**a**) Relative coverage of transcripts for the ONT cDNA-Seq dataset and the ONT RNA-Seq dataset for transcripts covered by at least 10 reads around a poly(T). With the ONT cDNA-Seq dataset, transcripts containing internal runs of at least 9 T's are less covered in 5'. The coverage deficit observed in the ONT RNA-seq dataset is due to indel sequencing errors associated to homopolymers.
(**b**) Relative coverage of transcripts for the ONT cDNA-Seq dataset and the ONT RNA-Seq dataset for transcripts covered by at least 10 reads around a poly(A). Using the ONT cDNA-Seq dataset, transcripts containing stretches of at least 9 A's are less covered in 3'. Again, the coverage deficit observed in the ONT RNA-seq dataset is due to indel sequencing errors associated to homopolymers.
(**c**) Mechanism explaining why internal runs of T's are causing 5' truncated reads. The PolyTVN primer binds to the internal run of poly(A) of the cDNA so that the second cDNA strand is 5' truncated.
(**d**) Example of a gene named *Set* 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 poly(T).

Accuracy evaluation of the gene expression quantification

We aligned reads to the reference transcriptome, used RSEM[7] for short reads, and counted the number of primary minimap2[8] alignments for long reads. The best quantifications were obtained for the ONT RNA-Seq and Illumina TruSeq protocols.



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 (n=67 transcripts). The observed quantifications are correlated with the known theoretical quantifications of the spike in.
(**a**) Correlation obtained for Illumina cDNA-Seq.
(**b**) Correlation obtained for Illumina with the ONT cDNA-Seq protocol.
(**c**) Correlation obtained for ONT cDNA-Seq.
(**d**) Correlation obtained for ONT RNA-Seq.

Conclusion

We generated a dataset which we think should be of general interest for the community. This dataset consists of RNA and cDNA sequencing of the same samples using both Illumina and ONT technologies. Importantly, we also mixed Lexogen spike-in together with our mouse samples.

Using the spike-in data, we find that the ONT RNA-Seq protocol is the most accurate, slightly better than the widely used Illumina TruSeq protocol. In contrast, the cDNA-Seq data was more biased and yielded a poorer quantification.

We further found that transcripts with internal runs of poly(T) tend to be truncated and over-sampled when using the ONT cDNA-Seq protocol. Biases associated to internal runs of poly(T) had remained undetected, although they may affect more than 20% of expressed transcripts in mouse.

Quantifying transcripts and not genes is still challenging. We therefore strongly recommend to map reads on the reference transcriptome and not on the genome, as reference transcripts do not contain introns, nor poly(A) tails. The development of dedicated bioinformatics tools, as well as the improvement of alignment tools, seem essential to correctly handle processed pseudogenes and quantify transcripts.

1. Lipson, D. *et al.* **Quantification of the yeast transcriptome by single-molecule sequencing**. *Nature Biotechnology* **27**, 652–658, issn: 1087-0156 (July 2009).
2. Wang, Z., Gerstein, M. & Snyder, M. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews. Genetics* **10**, 57–63, issn: 1471-0064 (Jan. 2009).
3. Garalde, D. R. *et al.* **Highly parallel direct RNA sequencing on an array of nanopores**. *Nature Methods* **15**, 201–206, issn: 1548-7091 (Jan. 2018).
4. https://www.lexogen.com/sirvs/
5. https://www.lexogen.com/teloprime-full-length-cdna-amplification/
6. Workman, R.E., Tang, A.D., Tang, P.S. *et al.* **Nanopore native RNA sequencing of a human poly(A) transcriptome**. *Nat Methods* **16,** 1297–1305 (2019).
7. Li, B. & Dewey, C. N. **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* **12**, 323, issn: 1471-2105 (Dec. 2011).
8. Li, H. **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* **34**, 3094–3100 (2018).