Methods 178 (2020) 89–95

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data

Pedro Prudêncio[a,b,*,1], Kenny Rebelo[a,1], Ana Rita Grosso[a,d], Rui Gonçalo Martinho[a,b,c], Maria Carmo-Fonseca[a,*]

[a] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[b] Center for Biomedical Research, Universidade do Algarve, Faro, Portugal
[c] iBiMED, Departamento de Ciências Médicas, Universidade de Aveiro, Aveiro, Portugal
[d] UCIBIO, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

ARTICLE INFO

Keywords:
mNET-seq
RNA polymerase II
Nascent RNA
Co-transcriptional splicing

ABSTRACT

Mammalian Native Elongating Transcript sequencing (mNET-seq) is a recently developed technique that generates genome-wide profiles of nascent transcripts associated with RNA polymerase II (Pol II) elongation complexes. The ternary transcription complexes formed by Pol II, DNA template and nascent RNA are first isolated, without crosslinking, by immunoprecipitation with antibodies that specifically recognize the different phosphorylation states of the polymerase large subunit C-terminal domain (CTD). The coordinate of the 3′ end of the RNA in the complexes is then identified by high-throughput sequencing. The main advantage of mNET-seq is that it provides global, bidirectional maps of Pol II CTD phosphorylation-specific nascent transcripts and coupled RNA processing at single nucleotide resolution. Here we describe the general pipeline to prepare and analyse high-throughput data from mNET-seq experiments.

1. Introduction

The advent of high-throughput sequencing combined with innovative and diversified techniques to capture RNA molecules has enabled a new generation of genome-wide studies of transcription and co-transcriptional RNA processing. Native Elongating Transcript sequencing (NET-seq) was originally established in yeast to visualize the genomic position of the active site of RNA polymerase II (Pol II) by identifying the 3′ ends of the nascent RNA [1]. Because only the coordinates of the 3′ end nucleotides are recorded, single nucleotide resolution is achieved. NET-seq relies on the intrinsic stability of ternary transcription complexes (formed by Pol II, DNA template and nascent RNA) to isolate Pol II elongation complexes by immunoprecipitation without crosslinking. In the original study, a 3xFLAG epitope tag was added to the C-terminus of the third Pol II subunit (Rpb3), yeast cells were lysed and a crude whole-cell extract was used for immunoprecipitation using anti-FLAG antibodies [1,2]. Application of NET-seq to yeast revealed pervasive polymerase pausing and backtracking throughout gene transcription [1] and advanced our understanding of promoter directionality [1,3]. The NET-seq strategy was also used in bacteria to map the density of RNA polymerase, leading to the identification of novel pause sites across the genome [4,5].

In contrast to yeast and bacteria, solubilisation of Pol II complexes under native conditions is typically incomplete in metazoan cells [6]. A practical solution to this problem was found by Nojima and Proudfoot [7,8], who solubilized isolated native chromatin by extensive micrococcal nuclease (MNase) digestion and then immunoprecipitated elongation complexes using antibodies that specifically recognize the different phosphorylation states of the polymerase large subunit C-terminal domain (CTD). Although initially applied to mammalian cells, and thus termed mNET-seq, the procedure can be adapted to any organism, as recently illustrated in plants [9].

In the mNET-seq method, RNA is purified from the immunoprecipitated Pol II complexes and used to prepare a cDNA library for high-throughput Illumina sequencing (Fig. 1A). To enable directional sequencing, the 5′ hydroxyl (OH) generated by MNase digestion of RNA is first converted to a 5′ phosphate by T4 polynucleotide kinase. RNAs isolated from Pol II complexes are then size-selected on denaturing polyacrylamide gels before subsequent adapter ligation for PCR-based preparation of the cDNA library. Specific adapters are then ligated to the 5′ P and 3′ OH ends of each RNA fragment (Fig. 1A). The adapters consist of sequences used to amplify the library by PCR using generic forward

* Corresponding authors.
E-mail address: pprudencio@medicina.ulisboa.pt (P. Prudêncio).
[1] These authors have contributed equally to the work.

https://doi.org/10.1016/j.ymeth.2019.09.003
Received 3 May 2019; Received in revised form 17 July 2019; Accepted 1 September 2019
Available online 05 September 2019

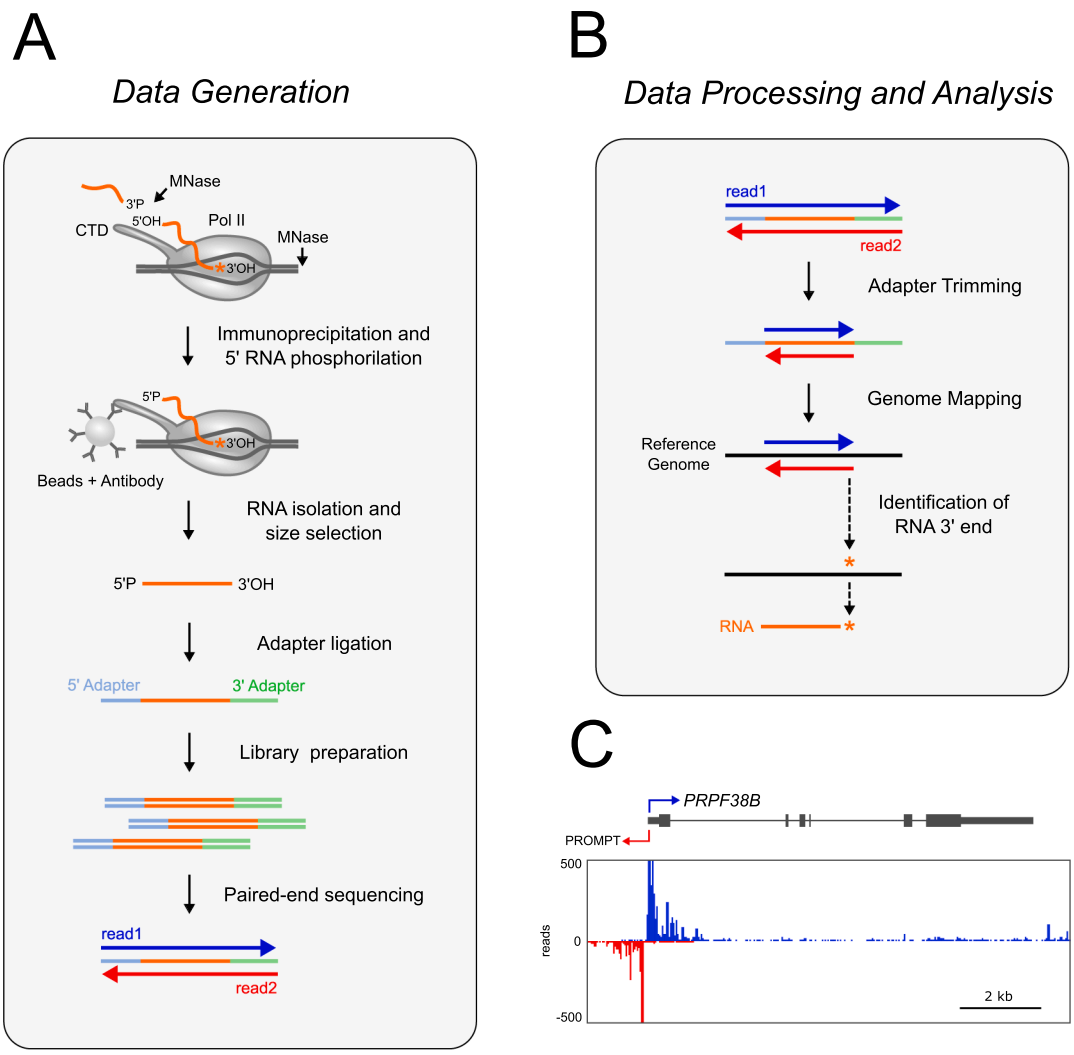

Fig. 1. Overview of mNET-seq. (A) Isolation of Pol II elongation complexes and library preparation. (B) Data processing. The orange asterisk denotes the nucleotide at the RNA 3′ end. (C) Visualization of mNET-seq profile along the *PRPF38B* gene. Data from HeLa cells 8WG16 mNET-seq replica1 [7] was aligned to the hg38 genome reference (GENCODEv28). Data visualized with UCSC genome browser. Blue and red arrows denote promoter bi-directionality. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and reverse primers, as well as sequences needed to associate the target nucleic acids with the sequencing instrument (e.g. the flowcell in Illumina sequencers) and, optionally, barcode sequences. After high-throughput sequencing, data must be prepared for analysis. This includes trimming of adapter sequences, mapping high quality reads to the reference genome, identification of the 3′ end nucleotide in each RNA fragment, and selection of genomic regions to be analyzed (Fig. 1B, C). In the following sections, we describe and discuss the primary analysis pipeline that we apply to data from mNET-seq experiments.

2. Methods

2.1. Quality control and adapter trimming

We use FastQC [10] for quality analysis of mNET-seq raw reads. As an example, we use the following FastQC (version 0.11.7) command to analyze data deposited in GSE106881 (GSM2856674 and GSM2856677):

```
fastqc mNET_Long_S5P_rep1_1.fastq
    fastqc mNET_Long_S5P_rep1_2.fastq
```

We further use FastQC to assess GC content, over-abundance of adapters and over-represented sequence, from which an indication of PCR duplication rate may be inferred [11]. Removal of adapter sequences and low quality reads is performed with Cutadapt [12]. We use Cutadapt (version 1.18) with an error rate of 0.05, and allow it to match 'N's in the reads to the adapter sequence; reads that are shorter than 10 bases are discarded, and the adapter is removed only once from each read. Example of Cutadapt command:

```
cutadapt -a TGGAATTCTCGGGTGCCAAGG -A GATCGTCGGACT
GTAGAACTCTGAAC –m
    10 -e 0.05 –match-read-wildcards -n 1 -o mNET_Long_
S5P_rep1_1_tr.fastq.gz –p
    mNET_Long_S5P_rep1_2_tr.fastq.gz    mNET_Long_S5P_rep1_1.
fastq
    mNET_Long_S5P_rep1_2.fastq
```

2.2. Mapping of reads to the reference genome

We initially used TopHat2 [13] for aligning mNET-seq reads to the reference human genome [7]. However, the currently most popular

mappers for RNA-seq data are STAR [14] and HISAT2 [15]. For STAR index generation, we set STAR (version 2.6.0b) to detect chimeric alignments with the minimum mapped length of at least 20nt on each end.

STAR index generation:

```
STAR –runMode genomeGenerate –genomeDir ./starIndex/
–genomeFastaFiles
    /genomes/human/hg38/GRCh38.primary.genome.fa
–-sjdbGTFfile
    /genomes/human/hg38/gencode.v28.annotation.gtf
```

Aligning paired-end reads:

```
STAR –runMode alignReads –genomeDir /genomes/human/
hg38/star/ –readFilesIn
    ./mNET_Long_S5P_rep1_1_tr.fastq.gz ./mNET_Long_S5P_rep1_
2_tr.fastq.gz–chimSegmentMin
    20 –outSAMtype BAM Unsorted –readFilesCommand gunzip -
c –outFileNamePrefix
    /alignments/mNET_Long_S5P_rep1_
```

The following command is used to obtain uniquely mapped reads with SAMtools (version 1.7):

```
samtools view -H mNET_Long_S5P_rep1_Aligned.out.bam
 > mNET_Long_S5P_rep1_header.sam
    samtools view -q 255 mNET_Long_S5P_rep1_Aligned.out.
bam >
    mNET_Long_S5P_rep1_unique.sam
    cat mNET_Long_S5P_rep1_header.sam mNET_Long_S5P_rep1_
unique.sam >
    mNET_Long_S5P_rep1_unique_H.sam
    samtools view -Sb -h mNET_Long_S5P_rep1_unique_H.sam >
    mNET_Long_S5P_rep1_unique.bam
    rm -f mNET_Long_S5P_rep1_unique_H.sam
    rm -f mNET_Long_S5P_rep1_unique.sam
```

An important mapping quality parameter is the percentage of mapped reads, which should always be higher than 70% [16]. We further use the RSeQC tool for quality control after mapping [17]. Only uniquely mapped reads are considered for further analysis.

2.3. Identification of RNA 3′ ends

NET-seq achieves single nucleotide resolution by mapping exclusively the nucleotide at the 3′ end of each immunoprecipitated RNA fragment. The full-length read sequences are discarded and only the coordinates of the 3′ end nucleotides are recorded as 1 M CIGAR strings. The RNA 3′ end corresponds to the 5′ nucleotide of read 2 in each sequencing pair, with the directionality indicated by read 1 (Fig. 1B). We do not use read 1 information because in sequencing-by-synthesis techniques accuracy decreases towards the 3′ end [18].

We developed a Python script for this purpose that is available in (https://github.com/kennyrebelo/mNET_snr). Briefly, the input alignment file (.bam) provided together with the -f argument are converted to .sam to ease subsequent parsing. For each pair of reads in the .sam file only read 2 is considered; any read that contains deletions, insertions or soft clipping information in the CIGAR string is disregarded. After obtaining a SAM flag that identifies the nucleotide at the 3′ end position, the CIGAR string is turned into “1M”. Finally, the single nucleotide resolution .sam file is converted into a .bam file. 3′ end nucleotides are obtained with Python (version 2.7.12) using the command:

```
python get_SNR_bam_ignoreSoftClip.py -f mNET_Long_S5P_rep1_
unique.bam –s
    mNET_Long_S5P_rep1 -d ./
```

2.4. Identification and removal of PCR internal priming and duplication events

Occasionally, the primer for reverse transcription (RT) anneals to the RNA fragment rather than to the adapter (Fig. 2A). When such internal priming occurs, the genomic sequence adjacent to the aligned read is complementary to the primer (NTGG in Fig. 2A, right panel). In order to remove reads that result from internal priming events, we developed a script that identifies the presence of the 3′ OH adapter sequence (for example, TGGAATTCTCGGGTGCCAAGG) downstream of the aligned reads (Fig. 2B). The script, which was developed and tested with SAMtools v1.7 and bedtoolsv2.27.1-1-gb87c465, is available in (https://github.com/kennyrebelo/Filtering_InternalPriming). For single-end reads, the input .bam alignment file is converted into a .bed file. /; for paired-end reads, the second read from each pair is extracted and added to a new .bam file that will then be converted to a .bed file. Iterations through the .bed file reveal the genome coordinates downstream of the 3′ OH position (i.e., the last nucleotide of single reads or the first nucleotide of the read 2 in paired reads). The number of nucleotides analysed downstream of the 3′ OH position corresponds to the adapter length provided together with the –a parameter. The next step is extracting nucleotide sequence for each .bed entry (converting the .bed file into a .fasta file). All entries matching the adapter sequence are discarded. The remaining read IDs are saved into a .txt file and are used to extract internal priming-free reads from the original alignment file. Example run for paired reads:

```
python Filter_InternalPriming.py -f /alignments/mNET_Long_
S5P_rep1_unique_sorted.bam –s
    paired -a TGG.. -g /genomes/human/hg38/GRCh38.primary.
genome.fa
```

We empirically determined that restricting the filter to the first three nucleotides of the adapter sequence (TGG in Fig. 2) detected the highest number of internal priming events. This is probably because base pairing of only a few nucleotides is sufficient for priming.

Using a pool of adapters having randomized sequences (barcodes) helps reducing PCR overamplification bias, as reads that align to the same genomic position and contain identical barcodes are likely the result of PCR duplication events. Duplicate reads can be removed using BBmap/clumpify.sh [19].

2.5. Distinguishing Pol II density profiles from co-transcriptional splicing

Because the final (3′ OH end) nucleotide of a nascent RNA lies at the active site of the polymerase, NET-seq provides nucleotide resolution profiles of RNA Pol II along the genome (Fig. 3A). However, the mNET-seq technique additionally detects the 3′ OH end of RNAs that are not located at the polymerase active site but associate with the Pol II elongation complex and are therefore co-immunoprecipitated [7,20]. This includes the 3′ OH ends generated by co-transcriptional cleavage of splice sites (Fig. 3B) and the free 3′ OH ends of spliceosome snRNAs (Fig. 3C). NET-seq reads mapping precisely to the last nucleotide of spliced exons correspond to splicing intermediates that are formed by cleavage at the 5′ splice site after the first splicing reaction, and reads mapping to the last nucleotide of introns correspond to released intron

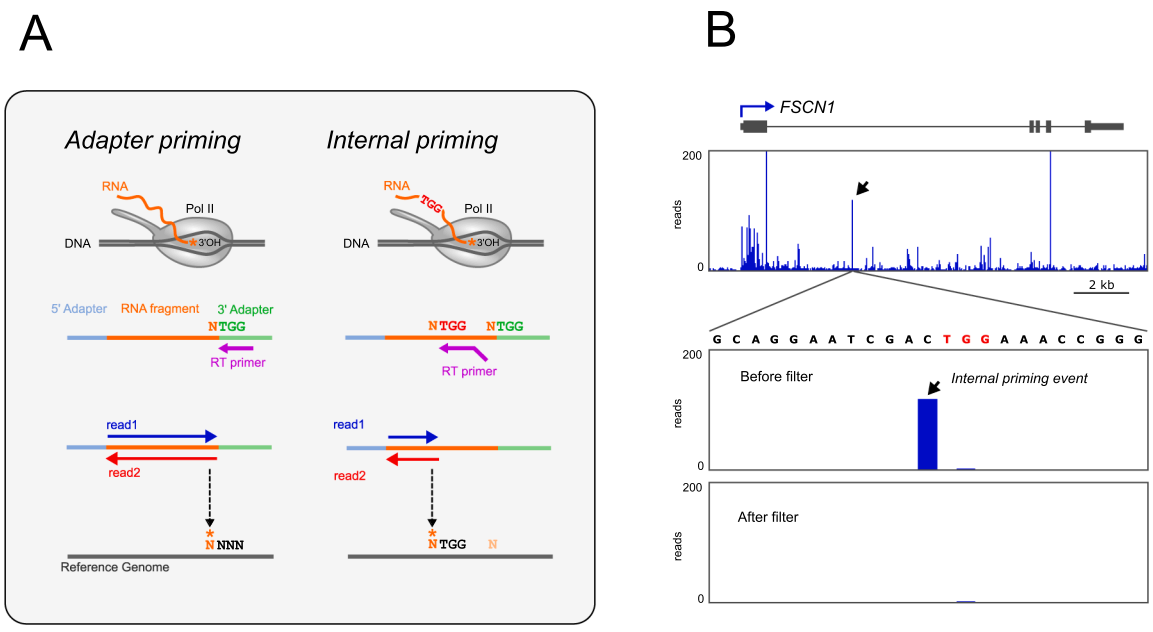

Fig. 2. Identification and removal of internal priming events. (A) Diagram depicts the expected base-pair complementarity between the RT primer and the adapter (left panel). Internal priming occurs when the RT primer hybridizes to the RNA sequence (right panel). (B) Visualization of mNET-seq profile along the *FSCN1* gene. The arrow denotes a spike resulting from internal priming. Data from HeLa cells long reads S5P mNET-seq replica1 [20].

lariats after completion of the splicing reaction (Fig. 3B). NET-seq reads mapping to the end of snRNA genes correspond to mature snRNAs engaged in co-transcriptional spliceosome assembly (Fig. 3C).

In order to determine RNA Pol II density profiles, only 3′ OH ends corresponding to the nucleotide at the active site of the polymerase are considered. To exclude signal from co-transcriptional splicing, reads that map to the very last nucleotide of introns and exons are discarded and the corresponding genomic positions are not considered. These reads can be removed using bedtools (version v2.27.1-1-gb87c465) together with SAMtools (version 1.7):

```
intersectBed -a mNET_Long_S5P_rep1_SNR.bam -b exons_lastNT.bed -wa -v | samtools view - > mNET_Long_S5P_rep1_SNR_noLastNT_temp.sam
    cat mNET_Long_S5P_rep1_header.sam mNET_Long_S5P_rep1_SNR_noLastNT_temp.sam > mNET_Long_S5P_rep1_SNR_noLastNT.sam
    samtools view -bS mNET_Long_S5P_rep1_SNR_noLastNT.sam > mNET_Long_S5P_rep1_SNR_noLastNT.bam
    rm -f mNET_Long_S5P_rep1_SNR_noLastNT_temp.sam
    rm -f mNET_Long_S5P_rep1_SNR_noLastNT.sam
```

2.6. Selection of transcriptionally active genes

One approach to identify which genes are being transcribed in a particular cell type is to use RNA-seq data of polyadenylated RNA. Alternatively, mNET-seq read density over genes can be used to measure transcriptional activity. However, because many inactive genes maintain high levels of Pol II paused near the promoter, thus generating promoter-proximal reads, quantification of mNET-seq signal should be restricted to the gene body region. In order to identify transcriptionally active genes based on mNET-seq signal, we use a strategy adapted from GRO-seq analysis [21] that relies on read density in gene desert regions as background reference for absence of transcription. Very large intergenic regions (gene deserts) are divided into 500 kb windows, and read densities are calculated by dividing read counts in each window by the window length (in bp). Read counts per window are obtained with bedtools (version v2.27.1-1-gb87c465) using the command:

```
coverageBed -a intergenic_regions_500kb_Windows.bed –b
    Filter_IP/mNET_Long_S5P_rep1_noInternalPriming.bam
-counts >
    intergenic_regions_500kb_Windows_cov.bed
```

A density threshold is arbitrarily defined as the 90th percentile of the total read density (Fig. 4A). This threshold is then used to identify which annotated genes are transcriptionally active, based on mNET-seq signal (in RPKM) over the gene body (Fig. 4B).

2.7. Data visualization

Several visualization tools are available to depict mNET-seq profiles in specific genomic regions and with strand directionality. These include VING [22], IGV [23] and UCSC genome browser [24].

2.8. Metagene analysis

Metagene plots provide visual representations of the average mNET-seq signal at specific genomic regions. For example, to visualize Pol II density at the intron-exon boundary, we use deepTools [25] to integrate mNET-seq signal over the last 200 nucleotides of introns and the first 200 nucleotides of adjacent exons (Fig. 5A). By defining a window of 200 bp upstream and downstream of the intron-exon junction (i.e., the 3′ splice site), the analysis must be restricted to introns and exons longer than 200 nucleotides. Further biological constraints can be imposed on the genomic regions selected for metagene analysis, such as only intron-exon boundaries of transcriptionally active genes or only intron-exon boundaries of constitutively spliced exons (identified in RNA-seq poly(A) data). For comparisons between regions (for example, intron-exon boundaries of spliced versus non-spliced exons) or experimental samples (for example, intron-exon boundaries in control cells versus cells treated with a drug that inhibits splicing), normalizations should be implemented. For normalization, we divide the number of reads at each nucleotide by the total number of reads in the entire genomic region under analysis. These values are then used to calculate

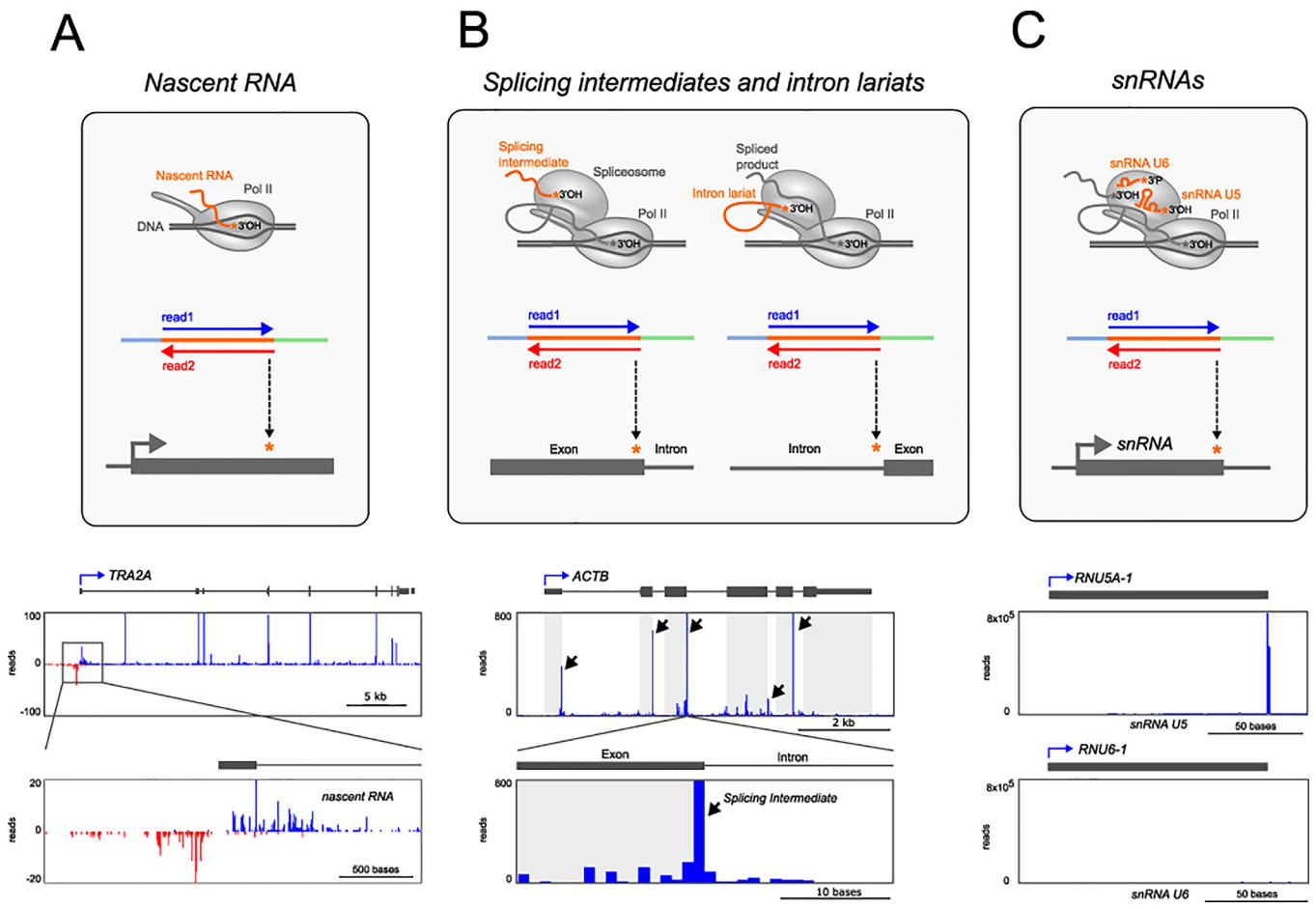

Fig. 3. Distinguishing mNET-seq signal from nascent RNA and co-transcriptional splicing. (A) A large fraction of mNET-seq signal corresponds to 3′ OH ends of nascent RNAs (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *TRA2A* gene. Data from HeLa cells short reads S5P mNET-seq replica1 [20]. (B) A fraction of mNET-seq signal corresponds to 3′ OH ends of either splicing intermediates or excised intron lariats (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *ACTB* gene using data from HeLa cells long reads S5P mNET-seq replica1 [20]. (C) An additional fraction of mNET-seq signal corresponds to 3′ OH ends of spliceosome snRNAs (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *snRNA U5* gene. Note that in contrast to the *snRNA U5* profile, no accumulation of mNET-seq signal is detected at the end of *snRNA U6* gene, as expected since U6 snRNA contains a phosphate terminal group at the 3′ end. Data from HeLa cells long reads S5P mNET-seq replica1 [20]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

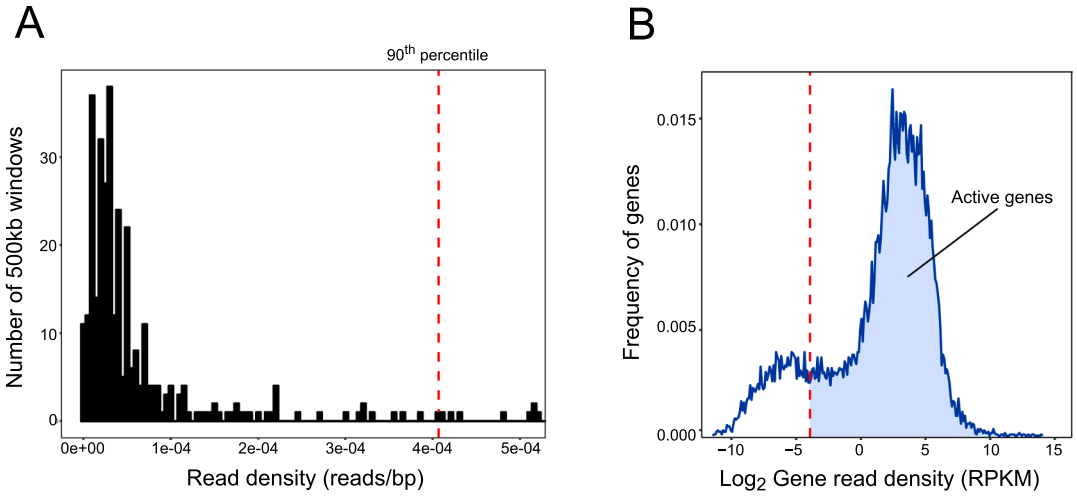

Fig. 4. Identification of transcriptionally active genes based on mNET-seq signal along the gene body. (A) Frequency distribution of read density in intergenic regions (gene deserts). A total of 724 windows (each 500 Kb in length) were defined in the RefSeq NCBI hg38 downloaded from UCSC table browser (accessed on the 3rd March 2019). The red dashed line represents the 90th percentile of read density in all regions analysed. (B) Frequency distribution of gene read density (RPKM) represented in Log2 scale. The 90th percentile of read density over gene deserts is set as threshold (red dashed line). A total of 13,426 genes are classified as transcriptionally active (blue area). Dataset from HeLa cells long reads S5P mNET-seq replica1 [20] (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

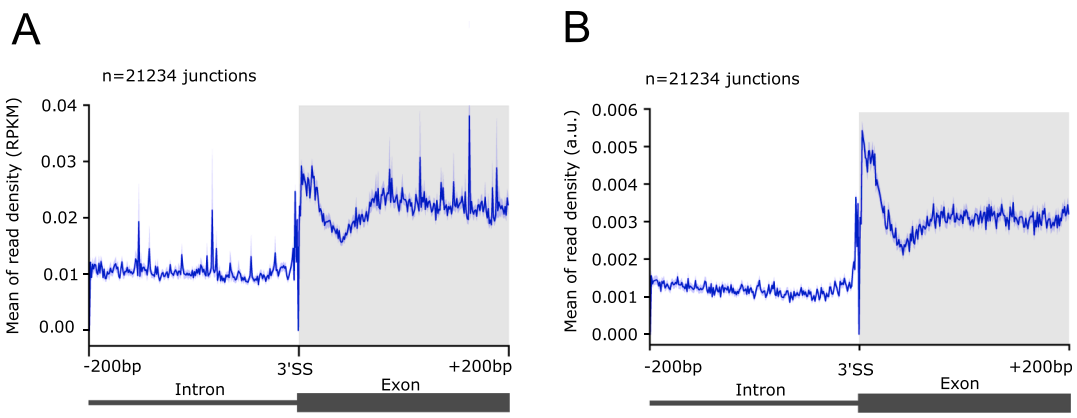

Fig. 5. Metagene analysis. (A) Metagene analysis for 21,234 intron-exon boundaries centred at the 3′ splice site (3′ ss). The mean of read density in RPKM was calculated for each nucleotide. (B) Read density was normalized and expressed in arbitrary units (a.u.). Dataset from HeLa cells long reads S5P mNET-seq replica1 [20]

the mean for each nucleotide, and the results are plotted in an arbitrary unit ranging from 0 to 1 (Fig. 5B).

2.9. Peak calling

Spikes in the density of 3′ ends at the active site of the polymerase are indicative of Pol II pausing. To identify Pol II pause positions along any given gene, Churchman and Weissman developed an algorithm that finds nucleotides where the NET-seq read density is at least three standard deviations above the mean in a local region [1]. As we systematically found significant accumulation of mNET-seq signal at the last nucleotide of spliced exons (Fig. 3B, bottom panel), we adapted this peak identifier strategy to quantify the prevalence of splicing intermediates and excised introns at genome-wide level [20]. The algorithm that we designed compares the number of reads mapping to the last nucleotide of exons and introns to the mean read density across the corresponding exon or intron. Only positions with coverage of at least 4 reads are considered. The script that is available in (https://github.com/kennyrebelo/NET_snrPeakFinder), was used with the following command:

```
python NET_snrPeakFinder.py -i mNET_Long_S5P_rep1_SNR.bam
-g gene_list.bed -s paired
```

Using this algorithm on mNET-seq datasets from HeLa cells, we found that approximately 90% of efficiently spliced exons had a 5′ splicing intermediate peak, whereas 3′ splice site peaks were rare [20]. Accumulation of mNET-seq signal corresponding to 5′ splicing intermediates and excised introns can be viewed as a proxy for co-transcriptional splicing kinetics, as the intensity of each peak depends on the lifetime of that particular RNA product. According to this view, mNET-seq reveals a significant time lag between the first and second splicing steps, whereas excised introns are rapidly degraded or dissociated from Pol II after completion of splicing.

3. Conclusions and perspectives

To date, the mNET-seq technique has been used in human [7,20,26], mouse [20] and plant cells [9]. In all cases, antibodies specific for Pol II with the CTD phosphorylated on serine 5 residues (S5P) immunoprecipitated abundant RNA fragments mapping precisely to the last nucleotide of exons, as expected for intermediates formed after the first splicing reaction. This suggests that the catalytically active spliceosome forms a tight complex with S5P Pol II and highlights the discovery potential of mNET-seq compared to previous techniques such as ChIP. Indeed, based largely on ChIP data, the established view was that the CTD was phosphorylated on serine 5 near the transcription start site, but shortly after transcription initiation and RNA capping these phosphates were removed [27]. According to mNET-seq data, serine 5 phosphorylation is not restricted to transcription initiation but is also present during elongation and co-transcriptional splicing. Noteworthy, a splicing-related accumulation of S5P Pol II along gene bodies was observed by ChIP in HeLa cells [28] and in yeast [29,30].

In addition to Pol II, other antibodies could in principle be used for mNET-seq. For example, mNET-seq analysis of complexes immunoprecipitated with antibodies to RNA processing factors could be useful for further studies on transcription-coupled pre-mRNA processing. Antibodies to Pol I and Pol III could also be used to determine the genomic distribution and nascent transcript profiles of these polymerases.

As an antibody-based technique, mNET-seq relies on the availability of antibodies that are specific and capable of precipitating the protein of interest. Problems related to antibody binding to off-target epitopes and differential accessibility of epitopes should always be considered, namely when investigating nascent transcripts associated with different phospho-isoforms of RNA Pol II. Another limitation of mNET-seq is treatment with MNase, which in our hands digests nascent RNA into 30–60 nucleotide-long fragments [20]. Developing new approaches to preserve longer stretches of nascent RNA associated with specific Pol II complexes is crucial for understanding how splicing kinetics is coordinated with transcription. Another challenge will be to adapt nascent RNA analysis to third generation sequencing technologies, which allow sequencing long molecules and avoid biases introduced by PCR amplification.

Acknowledgements/Funding

We are grateful to Nick Proudfoot and Taka Nojima (University of Oxford) for sharing mNET-seq datasets and for insightful discussion.

M.C.-F. acknowledges funding from Fundação para a Ciência e Tecnologia (FCT)/ Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through Fundos do Orçamento de Estado (UID/BIM/50005/2019), and FCT/FEDER/POR Lisboa 2020, Programa Operacional Regional de Lisboa, PORTUGAL 2020 (LISBOA-01-0145-FEDER-016394).. R.G.M. is supported by FCT grants PTDC/BEX-BID/0395/2014, PTDC/BIA-BID/28441/2017, and UID/BIM/04773/2013 CBMR 1334.

References

[1] L.S. Churchman, J.S. Weissman, Nascent transcript sequencing visualizes transcription at nucleotide resolution, Nature 469 (2011) 368–373, https://doi.org/10.

1038/nature09652.
[2] L.S. Churchman, J.S. Weissman, Native Elongating Transcript Sequencing (NET-seq), Curr. Protoc. Mol. Biol. 98 (2012) 14.4.1–14.4.17, https://doi.org/10.1002/0471142727.mb0414s98.
[3] Y. Jin, U. Eser, K. Struhl, L.S. Churchman, The ground state and evolution of promoter region directionality, Cell 170 (2017) 889–898.e10, https://doi.org/10.1016/j.cell.2017.07.006.
[4] M.H. Larson, R.A. Mooney, J.M. Peters, T. Windgassen, D. Nayak, C.A. Gross, S.M. Block, W.J. Greenleaf, R. Landick, J.S. Weissman, A pause sequence enriched at translation start sites drives transcription dynamics in vivo, Science 344 (2014) 1042–1047, https://doi.org/10.1126/science.1251871.
[5] I.O. Vvedenskaya, H. Vahedian-Movahed, J.G. Bird, J.G. Knoblauch, S.R. Goldman, Y. Zhang, R.H. Ebright, B.E. Nickels, Interactions between RNA polymerase and the "core recognition element" counteract pausing, Science 344 (2014) 1285–1289, https://doi.org/10.1126/science.1253458.
[6] H. Kimura, Y. Tao, R.G. Roeder, P.R. Cook, Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure, Mol. Cell. Biol. 19 (1999) 5383–5392, https://doi.org/10.1128/MCB.19.8.5383.
[7] T. Nojima, T. Gomes, A.R.F. Grosso, H. Kimura, M.J. Dye, S. Dhir, M. Carmo-Fonseca, N.J. Proudfoot, Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing, Cell. 161 (2015) 526–540, https://doi.org/10.1016/j.cell.2015.03.027.
[8] T. Nojima, T. Gomes, M. Carmo-Fonseca, N.J. Proudfoot, Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide, Nat. Protoc. 11 (2016) 413–428, https://doi.org/10.1038/nprot.2016.012.
[9] J. Zhu, M. Liu, X. Liu, Z. Dong, RNA polymerase II activity revealed by GRO-seq and pNET-seq in arabidopsis, Nat. Plants 4 (2018) 1112–1123, https://doi.org/10.1038/s41477-018-0280-0.
[10] S. Andrews, FASTQC. A quality control tool for high throughput sequence data. n.d. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (accessed April 15, 2019).
[11] R.M. Leggett, R.H. Ramirez-Gonzalez, B.J. Clavijo, D. Waite, R.P. Davey, Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics, Front. Genet. 4 (2013), https://doi.org/10.3389/fgene.2013.00288.
[12] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.J. 17 (2011) 10–12, https://doi.org/10.14806/ej.17.1.200.
[13] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (2013) R36, https://doi.org/10.1186/gb-2013-14-4-r36.
[14] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21, https://doi.org/10.1093/bioinformatics/bts635.
[15] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (2015) 357–360, https://doi.org/10.1038/nmeth.3317.
[16] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szcześniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, Genome Biol. 17 (2016) 13, https://doi.org/10.1186/s13059-016-0881-8.
[17] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics 28 (2012) 2184–2185, https://doi.org/10.1093/bioinformatics/bts356.
[18] C.W. Fuller, L.R. Middendorf, S.A. Benner, G.M. Church, T. Harris, X. Huang, S.B. Jovanovich, J.R. Nelson, J.A. Schloss, D.C. Schwartz, D.V. Vezenov, The challenges of sequencing by synthesis, Nat. Biotechnol. 27 (2009) 1013–1023, https://doi.org/10.1038/nbt.1585.
[19] Bushnell, Brian, BBMap: A Fast, Accurate, Splice-Aware Aligner, 2014. https://www.osti.gov/servlets/purl/1241166.
[20] T. Nojima, K. Rebelo, T. Gomes, A.R. Grosso, N.J. Proudfoot, M. Carmo-Fonseca, RNA polymerase II phosphorylated on CTD serine 5 interacts with the spliceosome during co-transcriptional splicing, Mol. Cell 72 (2018) 369–379.e4, https://doi.org/10.1016/j.molcel.2018.09.004.
[21] L.J. Core, J.J. Waterfall, J.T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, Science 322 (2008) 1845–1848, https://doi.org/10.1126/science.1162228.
[22] M. Descrimes, Y.B. Zouari, M. Wery, R. Legendre, D. Gautheret, A. Morillon, VING: a software for visualization of deep sequencing signals, BMC Res. Notes 8 (2015) 419, https://doi.org/10.1186/s13104-015-1404-5.
[23] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26, https://doi.org/10.1038/nbt.1754.
[24] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006, https://doi.org/10.1101/gr.229102.
[25] F. Ramírez, D.P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A.S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis, Nucleic Acids Res. 44 (2016) W160–W165, https://doi.org/10.1093/nar/gkw257.
[26] M. Schlackow, T. Nojima, T. Gomes, A. Dhir, M. Carmo-Fonseca, N.J. Proudfoot, Distinctive patterns of transcription and RNA processing for human lincRNAs, Mol. Cell 65 (2017) 25–38, https://doi.org/10.1016/j.molcel.2016.11.029.
[27] D. Eick, M. Geyer, The RNA polymerase II carboxy-terminal domain (CTD) code, Chem. Rev. 113 (2013) 8456–8490, https://doi.org/10.1021/cr400071f.
[28] E. Batsché, M. Yaniv, C. Muchardt, The human SWI/SNF subunit Brm is a regulator of alternative splicing, Nat. Struct. Mol. Biol. 13 (2006) 22–29, https://doi.org/10.1038/nsmb1030.
[29] R.D. Alexander, S.A. Innocente, J.D. Barrass, J.D. Beggs, Splicing-dependent RNA polymerase pausing in yeast, Mol. Cell 40 (2010) 582–593, https://doi.org/10.1016/j.molcel.2010.11.005.
[30] K.T. Chathoth, J.D. Barrass, S. Webb, J.D. Beggs, A splicing-dependent transcriptional checkpoint associated with prespliceosome formation, Mol. Cell 53 (2014) 779–790, https://doi.org/10.1016/j.molcel.2014.01.017.