Terminome-Seq library preparation protocol

Plant Material

Arabidopsis thaliana Col-0 and *pnp1-1* seeds were germinated on MS medium with 16 hrs of light per day at 23°C. Three-week old leaf material was flash-frozen in liquid nitrogen, and total RNA was isolated using TRI Reagent® according to the manufacturer's instructions (www.sigmaaldrich.com).

Terminome Library Synthesis and analysis

All libraries were produced from 1 µg of DNase I (www.neb.com)-treated RNA, and for TAP-treated samples, tobacco acid phosphorylase (www.epibio.com) was used according to the manufacturer's instructions. Library synthesis was carried out using the Illumina TruSeq Small RNA library preparation kit (www.illumina.com) with minor modifications depending on whether a native 5' or 3' end was the target (Figure 1). Libraries intended for native 3' end capture underwent an extra step between the initial 3' adapter ligation and ligation of the 5' adapter, namely RNA fragmentation using a Covaris sonicator (www.covaris.com), with a target size of 200 nt. Libraries intended for native 5' end capture required further adjustments. First, the order of adapter ligation was reversed: 5' adapter ligation – sonication – 3' adapter ligation. This reversal, however, generated unwanted primer dimers that were preferentially amplified during library amplification (PCR1) due to their small size (~133 bp) (Figure 2, left panel). Therefore, size selection was performed on the products from PCR1, retaining only products over 200 bp using Pippin Prep (www.sagescience.com). A second PCR amplification (PCR2) was executed on these products and gave libraries suitable for sequencing (Figure 2, right panel). Quality control was performed with an Agilent BioAnalyzer (www.agilent.com).

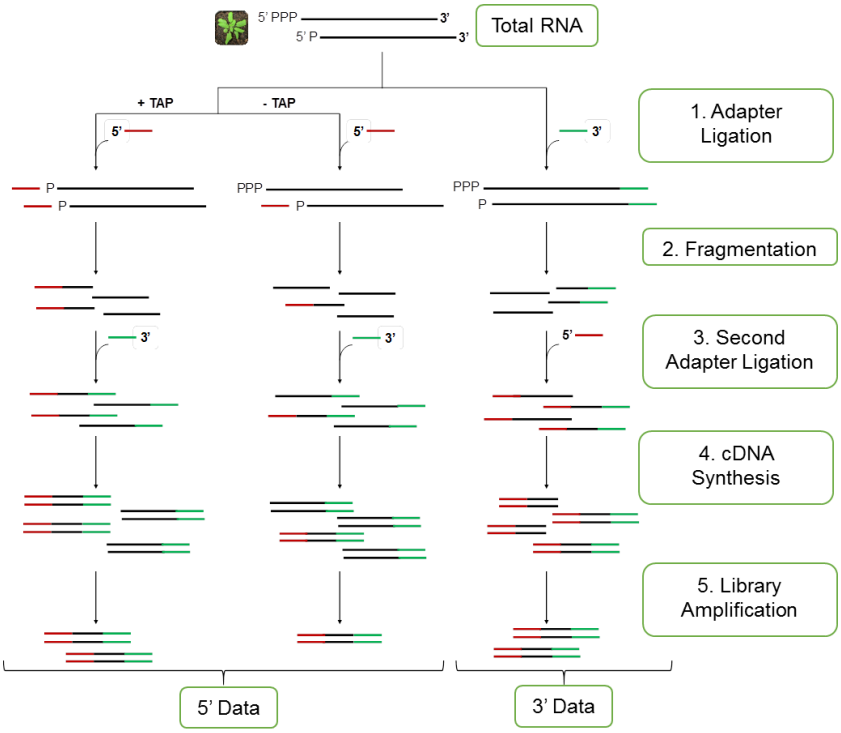

Figure 1 Terminome-Seq strategy

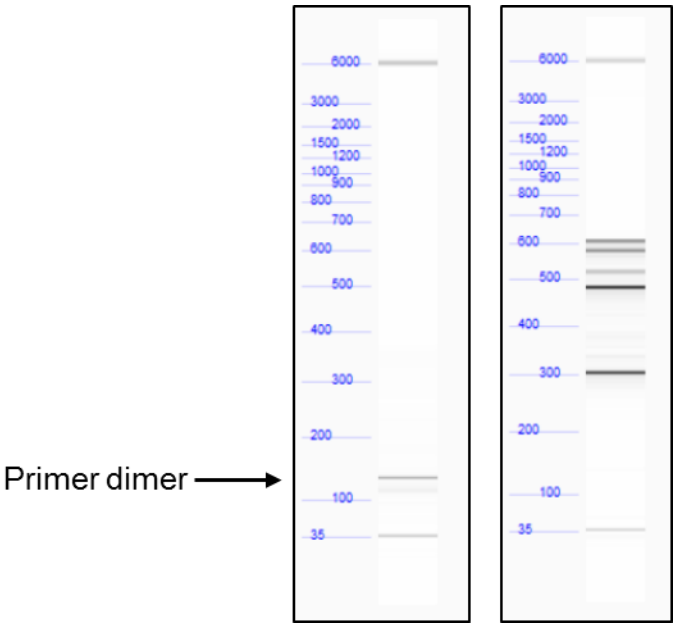

Figure 2 Bioanalyser results for a representative 5' end sequencing library before (left) and after (right) size selection and a second PCR amplification step

All library purifications were completed using magnetic AMPure beads (www.beckman.com).

Libraries were pooled and sequenced on a NextSeq500 Sequencer (www.illumina.com) using the v3 kit, and 2 x 75 bp paired-end reads were generated with 40 bp long R1 reads and 35 bp long R2 reads for all libraries. R1 reads are only of use for libraries generated to obtain 5' related data, while R2 reads contain data related to 3' ends and therefore only relevant for libraries generated to obtain 3' data. Raw sequences have been deposited on the SRA database with the number PRJNA533962 and can be accessed here https://www.ncbi.nlm.nih.gov/sra/PRJNA533962. Sequencing has been performed at the Biotechnology Resource Center from Cornell University, Ithaca, NY, USA.

Reads were quality trimmed using fastq-mcf (https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md) using default parameters. Alignment to the chloroplast genome was performed using TopHat2 (https://ccb.jhu.edu/software/tophat/index.shtml) allowing up to 2 alignments to account for chloroplast inverted repeat (-g 2). The GFF file containing the gene coordinates and the indexed genome files were downloaded from https://github.com/BenoitCastandet/chloroseq/tree/master/TAIR10_ChrC_files. This corresponds to the TAIR10 version modified to add the first exon of the chloroplast gene *ycf3* that is missing from the annotation. The novel splice site discovery option was disabled (-no-novel-juncs). Alignment statistics are available in Table 1.

Table 1 Terminome-Seq alignment summary

sample_name	library_number	Read[a]	Mappable reads[b]	Mapped reads[c]	Mapped/Mappable (%)
At_WT_noTAP_5'_1	library_14	R1	20879044	1933298	9,3
At_WT_noTAP_5'_2	library_15	R1	33452268	8754128	26,2
At_WT_TAP_5'_1	library_23	R1	15959028	2599072	16,3
At_WT_TAP_5'_2	library_24	R1	32794671	8252609	25,2
At_WT_3'_1	library_1	R2	13379429	3275122	24,5
At_WT_3'_2	library_2	R2	15105169	3036259	20,1
At_pnp_noTAP_5'_1	library_16	R1	52776856	9968852	18,9
At_pnp_noTAP_5'_2	library_17	R1	47539576	6056011	12,7
At_pnp_3'_1	library_7	R2	29219035	6977328	23,9
At_pnp_3'_2	library_8	R2	19628552	4621698	23,5

[a] Sequencing was 2 x 75 bp paired-end with 40 bp long R1 reads and 35 bp long and only R1 reads were used for 5' ends and R2 reads for 3' ends analyses.

[b] Mappable reads were selected after quality control using fastq-mcf.

[c] Mapped reads represent the mappable reads that aligned to the Arabidopsis chloroplast genome.

Native transcript ends were defined as the first nucleotide of each aligned read. These positions were extracted and counted using homemade bash scripts run_find_ends_5_R1 and run_find_ends_3_R2 to find 5' and 3' ends respectively. Coverage obtained was then normalized to Reads per Million (RPM) using the mapped reads indicated in Table 1 and the coverages for the two replicates were averaged. The full coverage at single nucleotide resolution for 5' (+/–TAP) and 3' ends generated for this manuscript can be accessed in Supplementary Table S1.