**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-treated RNA (www.neb.com), and for TAP-treated samples, tobacco acid phosphorylase (TAP; www.epibio.com) was used according to the manufacturer's instructions with heat inactivation at the end of the incubation period. Library synthesis was carried out using the Illumina TruSeq Small RNA library preparation kit (www.illumina.com) intended to capture the RNA population containing a 5' phosphate and 3' hydroxyl group. Minor modifications were made to the protocol depending on whether a native 5' or 3' end was the target (Figure 1). Libraries intended for native 3' end capture followed the protocol with initial 3' adapter ligation using T4 RNA ligase 2, a deletion mutant that can only ligate a 3' hydroxyl group to a 5' adenylated RNA, consistent with the 3' RNA adapter chemistry. After ligation, the RNA was fragmented using a Covaris sonicator (www.covaris.com), with a target size of 200 nt, followed by ethanol precipitation for concentration and 5' adapter ligation with T4 RNA ligase. Libraries intended for native 5' end capture required further adjustments. First, the order of adapter ligation was reversed: 5'

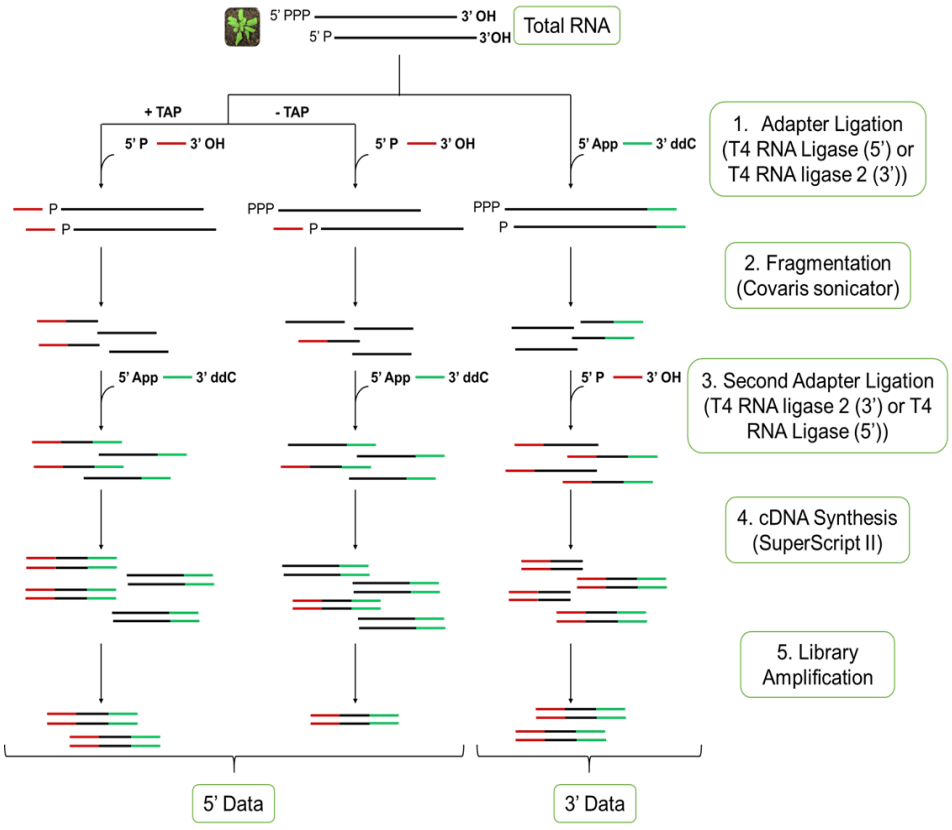


**Figure 1 Terminome-Seq strategy**

adapter ligation (with T4 RNA ligase) – sonication – ethanol precipitation – 3' adapter ligation (with T4 RNA ligase 2). In this case, excess 5' adapter remaining following the sonication and ethanol precipitation could ligate to added 3' adapter, but not to any new 5' ends created through sonication as the new 5' ends would not be adenylated. This resulted in unwanted adapter 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 after Pippin size selection and before library submission for sequencing using an Agilent BioAnalyzer (www.agilent.com). The final cDNA libraries were purified using magnetic AMPure beads (www.beckman.com) following the manufacturer's protocol.

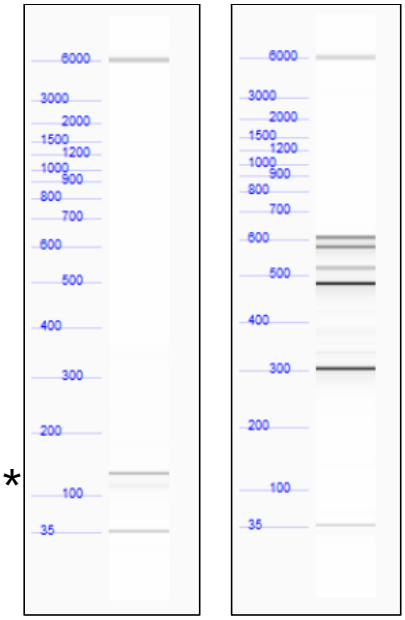


**Figure 2 Bioanalyzer results for a representative 5' end sequencing library before (left) and after (right) size selection and a second PCR amplification step.** * - adapter dimers

Multiple steps in the above protocol, including fragmentation, ethanol precipitation, Pippin size selection (5' libraries only), and AMPure purification of cDNA libraries, resulted in a bias towards the retention, and therefore sequencing, of fragments >67 nt. As a consequence, ends of small RNAs (smRNAs) and smaller tRNAs would be expected to be underrepresented in the results. An additional bias was introduced because the RT primer is fully complementary to the 3' adapter, ending in a sequence complementary to TGG. Illegitimate priming by the adapter resulted in an estimated 52 additional 3' ends terminating in TGG which are included in our data as they cannot be distinguished from legitimate 3' termini ending in T(U)GG (Figure 3

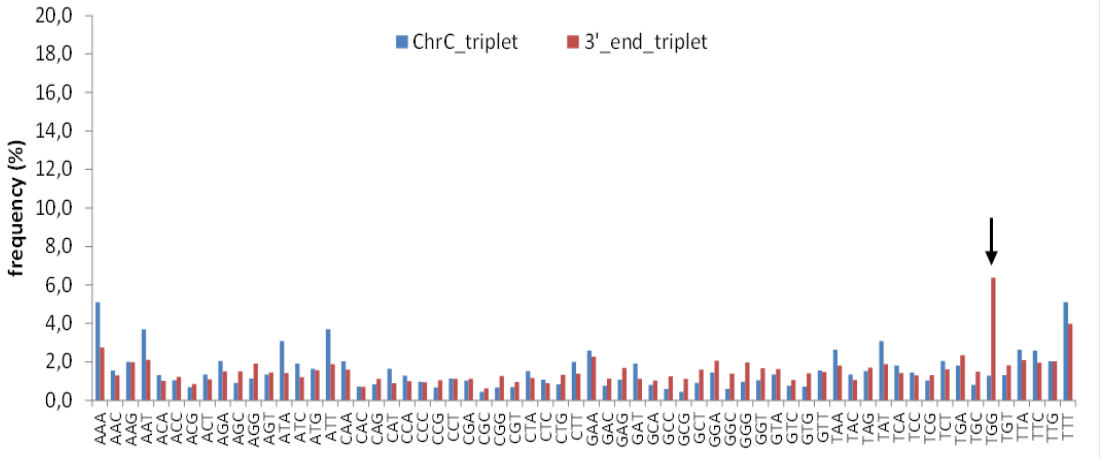


**Figure 3: Comparison between the frequencies of the plastome trinucleotide and the triplet downstream of 3' ends.** TGG triplets are highlighted by an arrow.

illustrates the bias). While this bias was inevitable, the overall interpretation of the results was not affected.

Libraries were pooled and sequenced on a NextSeq500 Sequencer (www.illumina.com) using the v3 kit with paired-end reads generating 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 are 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) with the 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 the 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 paired-end with 40 bp long R1 reads and 35 bp long R2 reads. 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 a single nucleotide resolution for 5' (+/–TAP) and 3' ends generated for this manuscript can be accessed in Supplementary Table S1.

Additionally, two other alignment strategies were used as a control. First, both R1 and R2 reads were aligned to the chloroplast genome leaving the post-alignment strategy the same; an end was defined as the first nucleotide of uniquely mapped reads. As illustrated in Figure 4, the results are similar to the ones obtained when only R1 or R2 reads were used for the alignment. This additionally allowed us to evaluate the fragment size for the different libraries (Figure 5).

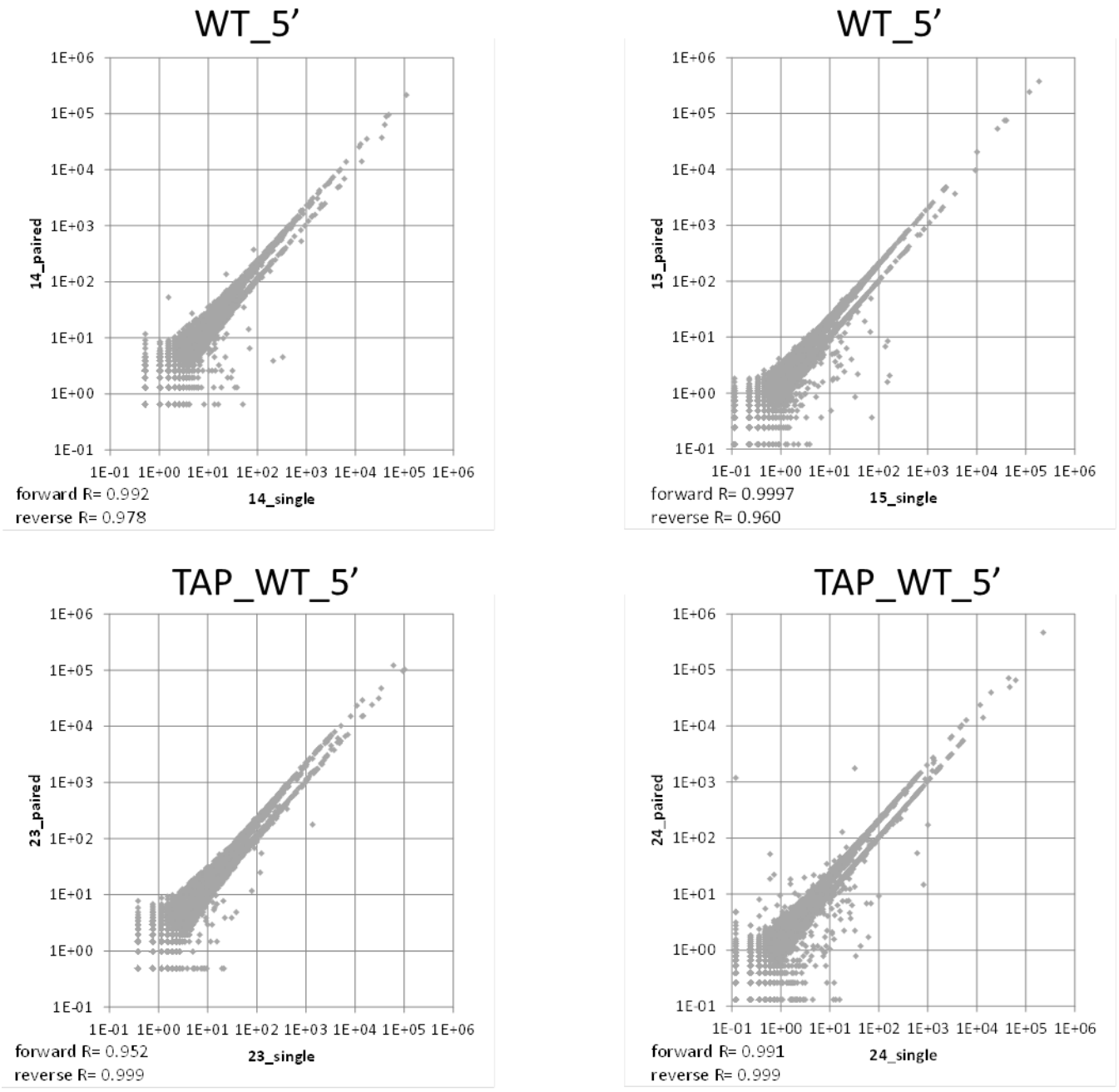


**Figure 4 Comparison of 5' ends obtained when only R1 reads were aligned to the chloroplast genome (single – x-axis) or when both R1 and R2 reads were aligned to the chloroplast genome (paired – y-axis)**.

RPM for all ends were graphed for the 4 WT 5' libraries based on the alignment strategy (single R1 reads or paired R1 and R2 reads). Correlation coefficients for each strand are given to the bottom left of each graph.

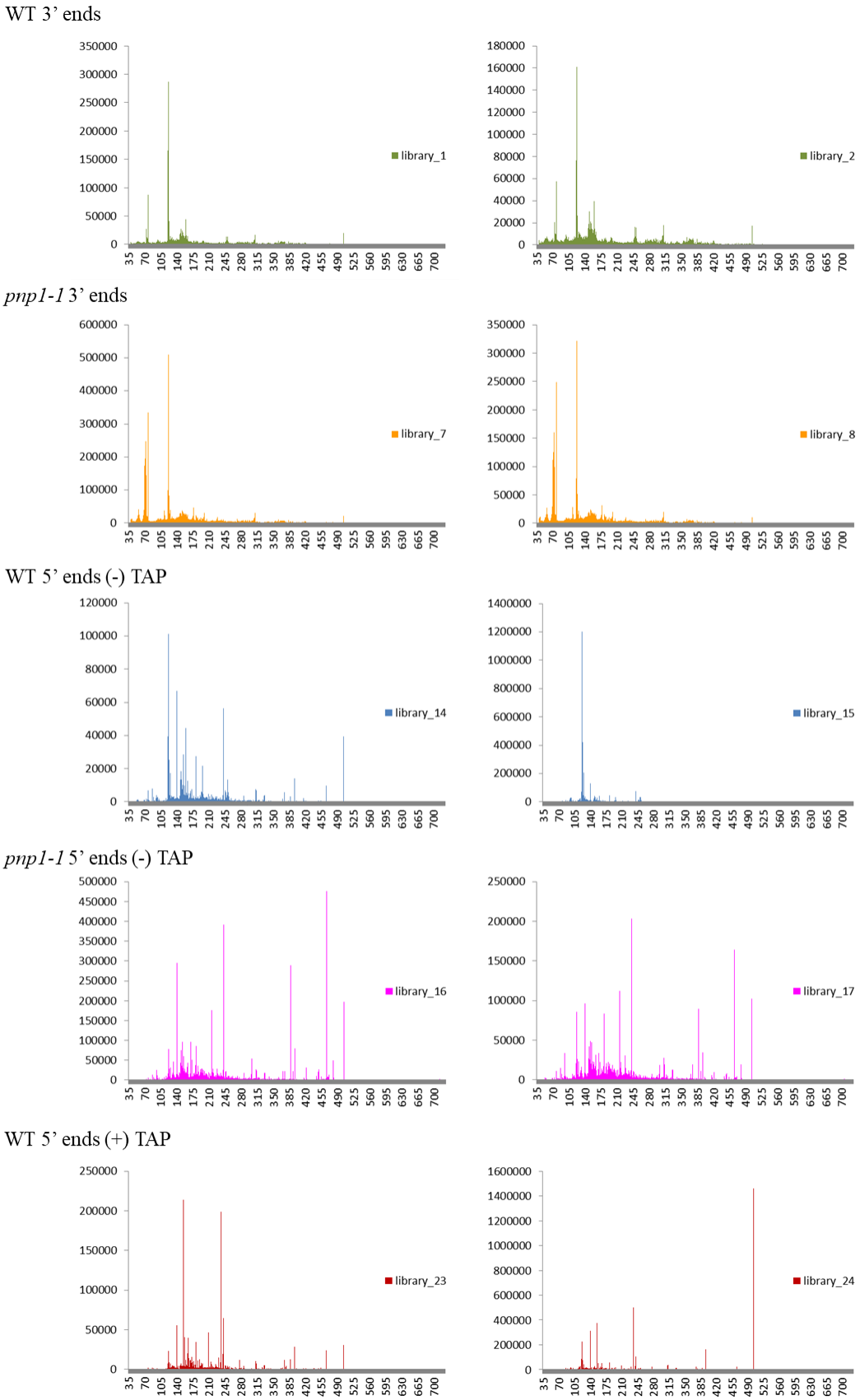


**Figure 5 Fragment size of the different Terminome-Seq libraries.**

In the last alignment strategy used, both R1 and R2 reads were aligned to the full *Arabidopsis col0* genome (including the recently deposited *col0* mitochondrial genome, Sloan et al, 2018, *Plant Cell*, Vol. 30: 525–527), allowing reads potentially matching both the nucleus and chloroplast genomes to be removed prior to alignment to the chloroplast. The post alignment strategy was the same; an end was defined as the first nucleotide of uniquely mapped reads. As illustrated in Figure 6, the results are almost identical to those obtained when both R1 and R2 reads aligned to the chloroplast genome only.

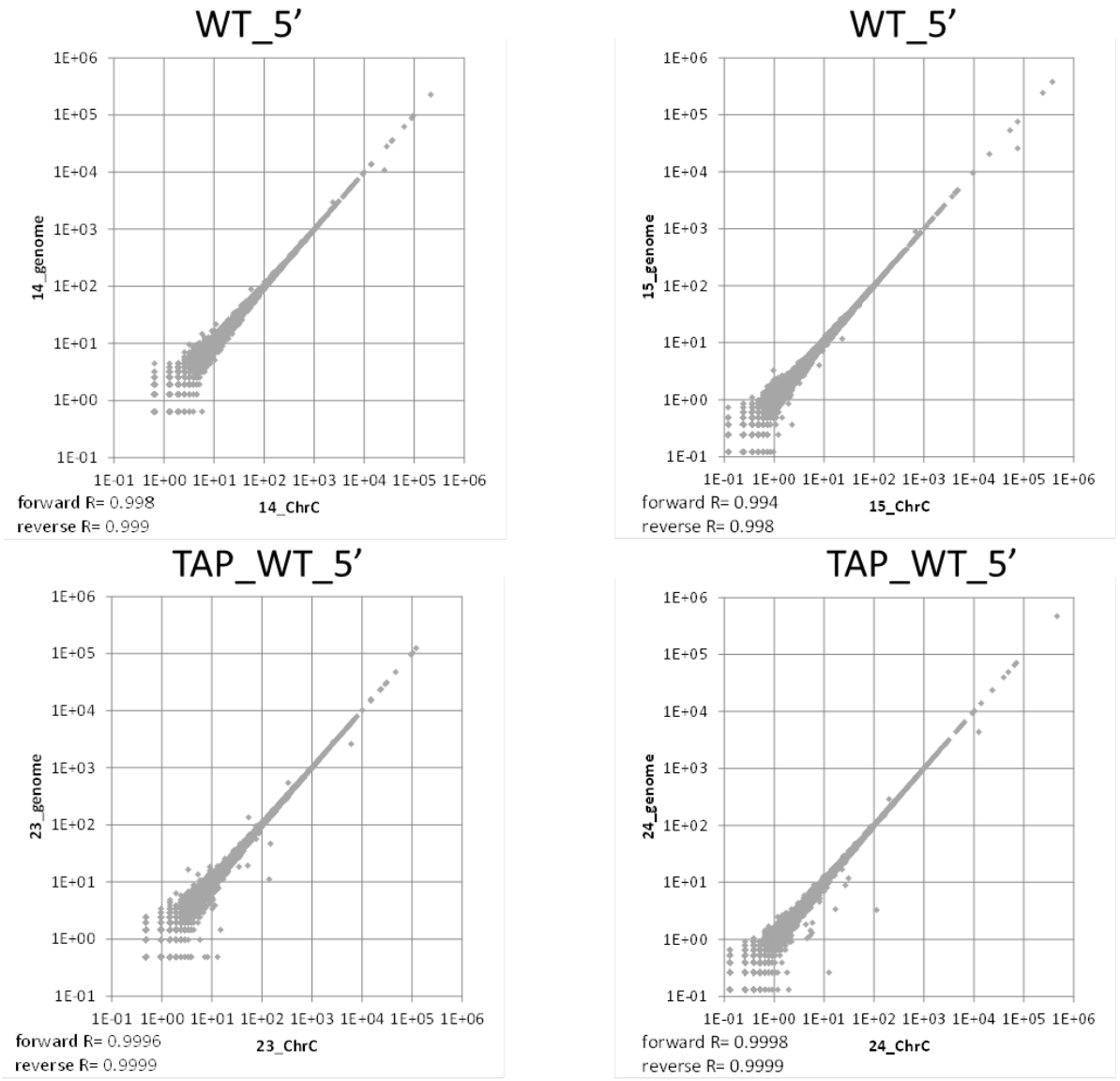


**Figure 6 Comparison between the 5' ends obtained when R1 and R2 reads were aligned to the chloroplast genome only (ChrC – x-axis) or to the full *Arabidopsis* genome (genome – y-axis).**

Finally, to confirm that the alignment strategy did not introduce any major biases, the influence of the 3 strategies on the resulting +TAP/-TAP ratio used to infer TSS in the three "hotspots", *ndhF*, as_*ndhF* and *clpP_i1* was evaluated (Figure 7). No substantial difference was noted, indicating that any of the three described alignment strategies may be used.

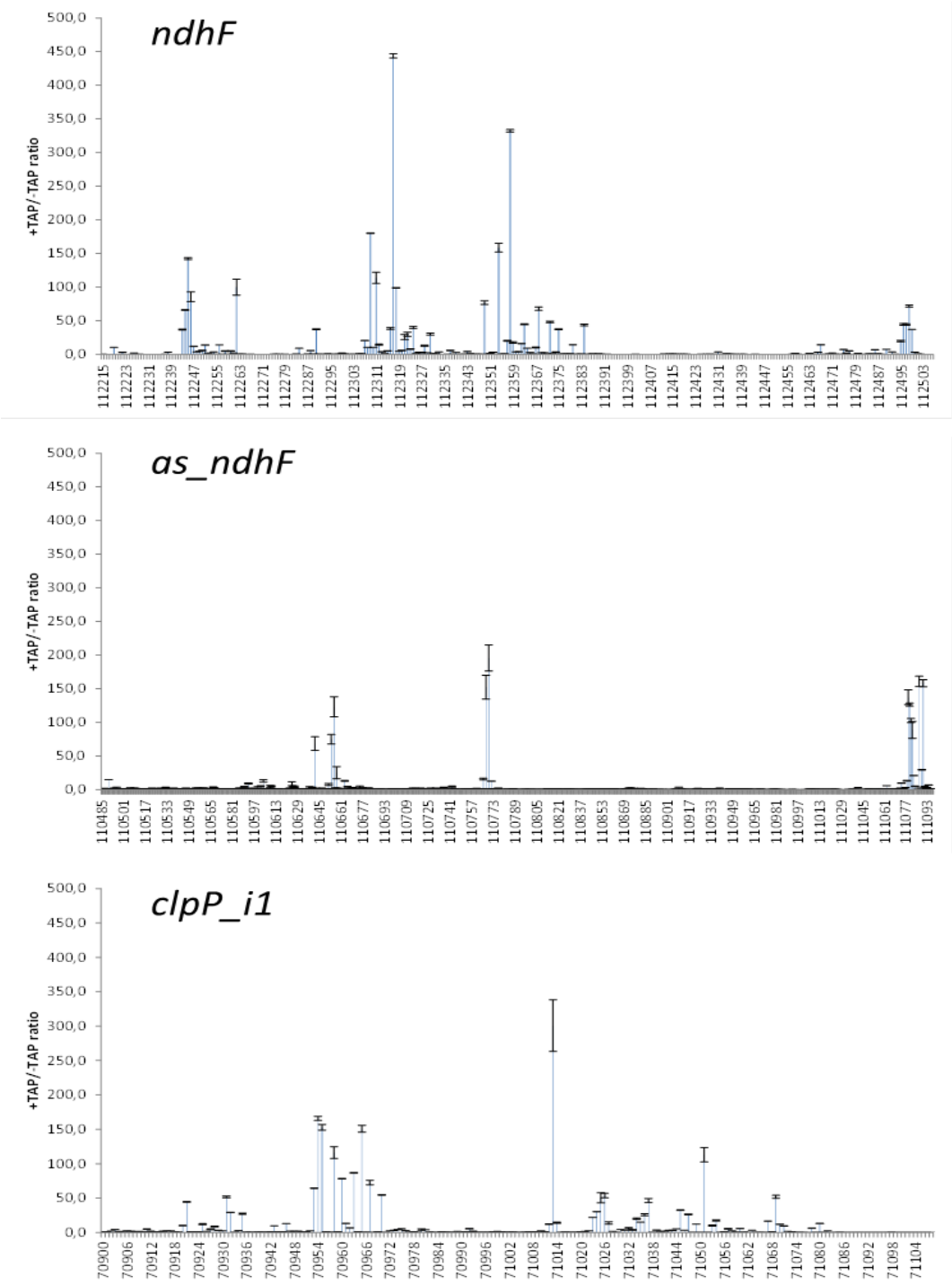


**Figure 7 Average +TAP/TAP ratios obtained from three different alignment strategies in TSS hotspots.** Error bars indicate the standard deviation in the +TAP/-TAP ratio from the three different strategies.