

Supplemental Information

**Integrator is a genome-wide attenuator
of non-productive transcription**

Søren Lykke-Andersen, Kristina Žumer, Ewa Šmidová Molska, Jérôme O. Rouvière, Guifen Wu, Carina Demel, Björn Schwalb, Manfred Schmid, Patrick Cramer, and Torben Heick Jensen

Figure S1

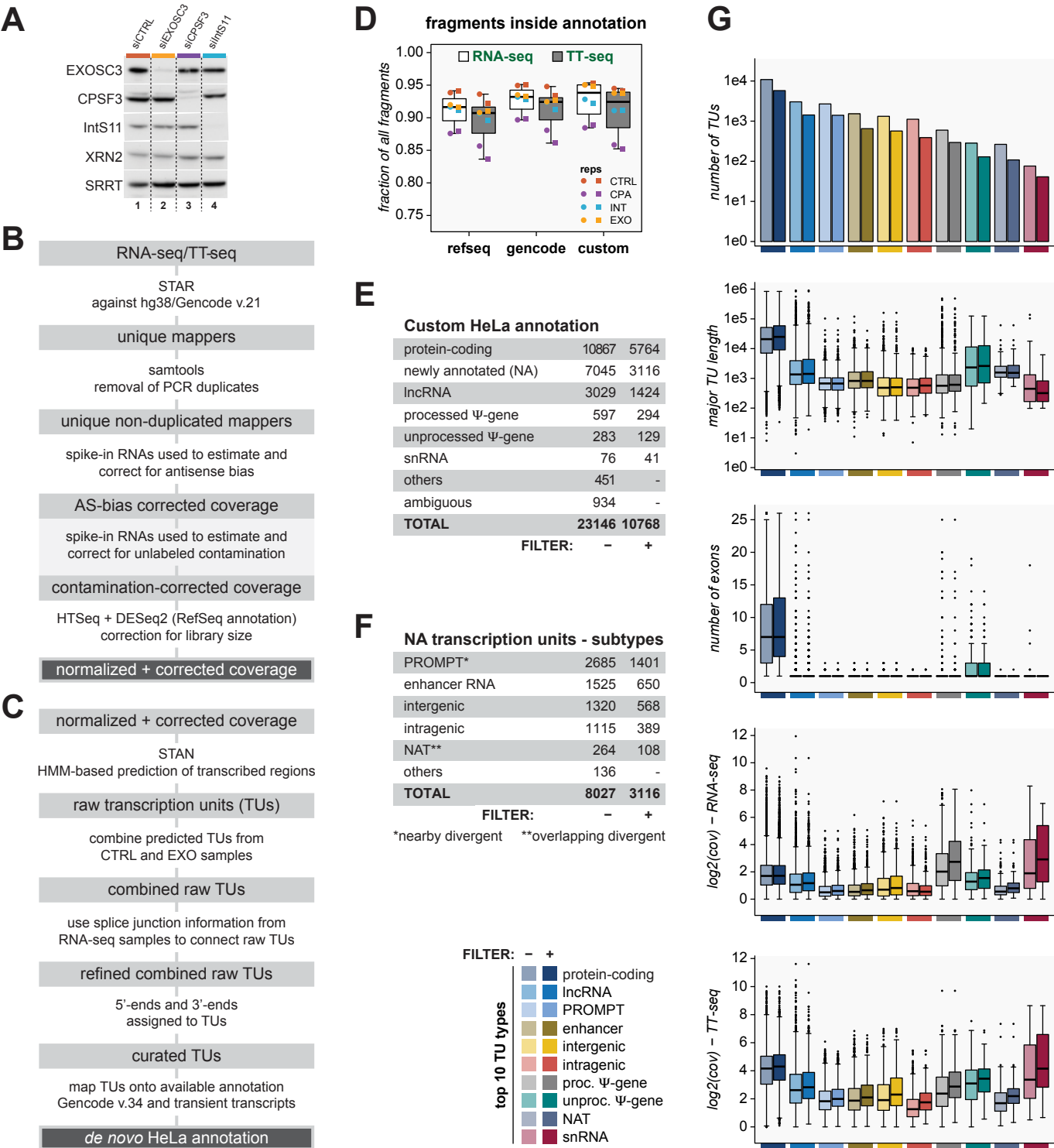


Figure S1, related to Figure 1. Exhaustive *de novo* annotation of HeLa cell TUs

(A) Western blotting analysis demonstrating efficient knockdown of indicated proteins. XRN2 and SRRT/ARS2 were probed as loading controls.

(B) Flow diagram illustrating the computational workflow for preprocessing the TT-seq and RNA-seq data to obtain coverage corrected for antisense-bias, contaminating unlabeled RNA (in the case of TT-seq) and library-size.

(C) Flow diagram illustrating the computational steps applied to TT-seq and RNA-seq data from siEGFP- and siEXOSC3-treated HeLa cells to annotate the HeLa cell transcriptome.

(D) Box plot displaying the fraction of RNA-seq (white) and TT-seq (gray) signal coverage from the indicated samples (siEGFP-, siCPSF3-, siIntS11- and siEXOSC3-treated samples, coined CTRL, CPA, INT and EXO, respectively), falling inside refseq (v.11), gencode (v.34) or custom *de novo* HeLa annotations. Box limits represent the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

(E) Overview of biotypes present in the HeLa transcriptome. NA=newly annotated and not in Gencode v.34. The 'FILTER -' column displays the number of identified TUs. The 'FILTER +' column displays the number of TUs used for downstream analysis in this study (see *STAR Methods* for details).

(F) Overview of subtypes of NA TUs. Columns as in (E).

(G) Characteristics of TU types and subtypes represented in the HeLa transcriptome annotation. For each TU type the values are shown for the complete annotated set and the subset used in downstream analyses on the left and right, respectively. From the top: (1) number of TUs within TU type-class (log10-scale), (2) length of TU (log10-scale), (3) the number of exons, (4) median log2-transformed RNA-seq coverage and (5) median log2-transformed TT-seq coverage. For clarity, the y-axis in (3) was restricted at 30 exons although there are several protein-coding TUs with more exons. Values are shown for the major transcript (see *STAR Methods* for details). Median, hinges and whiskers in box plots (panels 2-5) are shown as in (D).

Figure S2

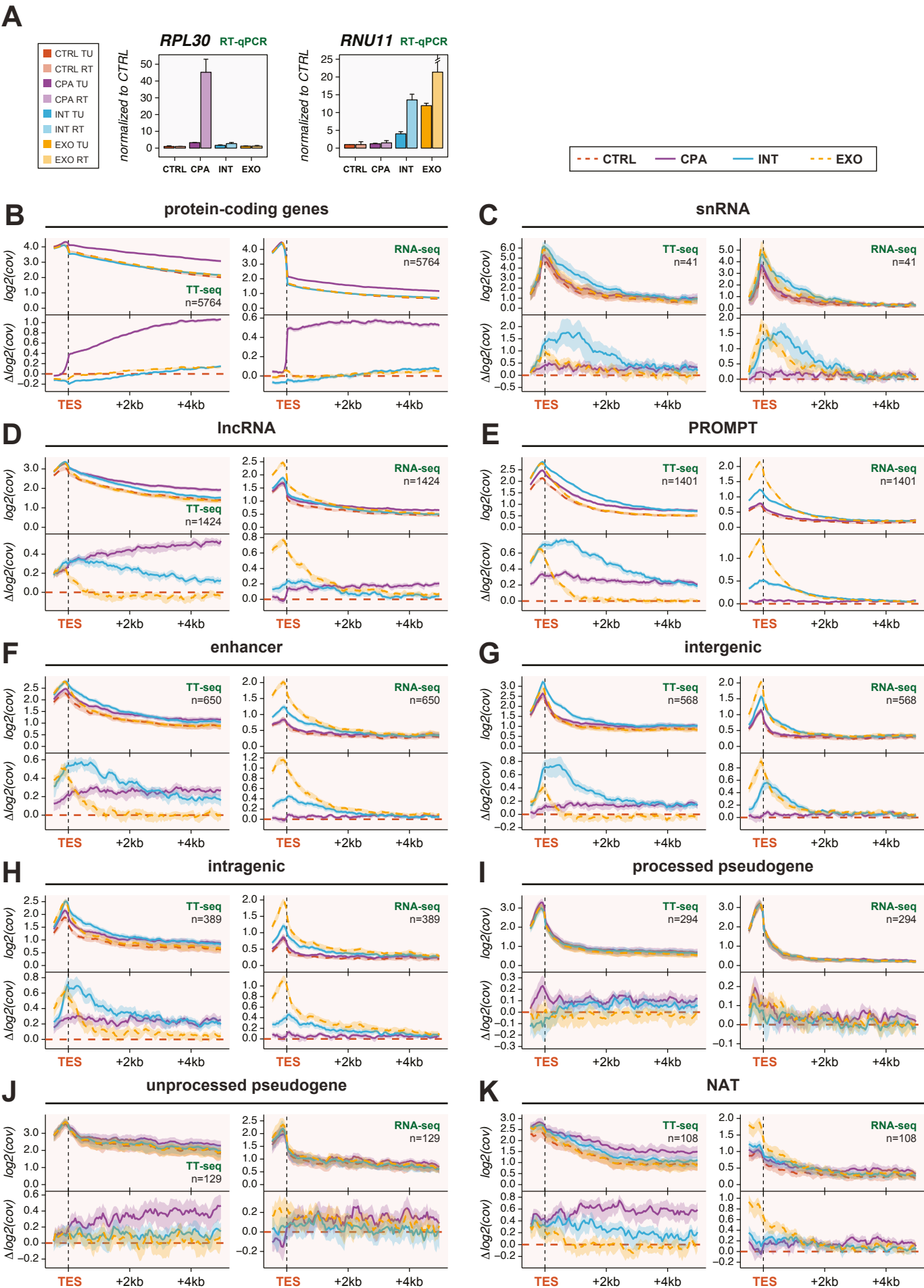


Figure S2, related to Figure 2. CPA and INT depletions display common and diverse global transcription termination phenotypes at TU ends

(A) RT-qPCR validations of bona-fide CPA and INT substrates, respectively. 'TU' and 'RT' denote the employed amplicons (positions depicted in Figure 2B).

(B)-(K) Aggregate plots of mean TT-seq (*left*) and RNA-seq (*right*) data aligned to TESs of TUs from major RNA TU type-classes represented in the HeLa transcriptome annotation: (B) protein-coding, (C) snRNA, (D) lncRNA, (E) PROMPT, (F) enhancer, (G) intergenic, (H) intragenic, (I) processed pseudogene, (J) unprocessed pseudogene and (K) NAT. Representation as in Figure 2E.

Figure S3

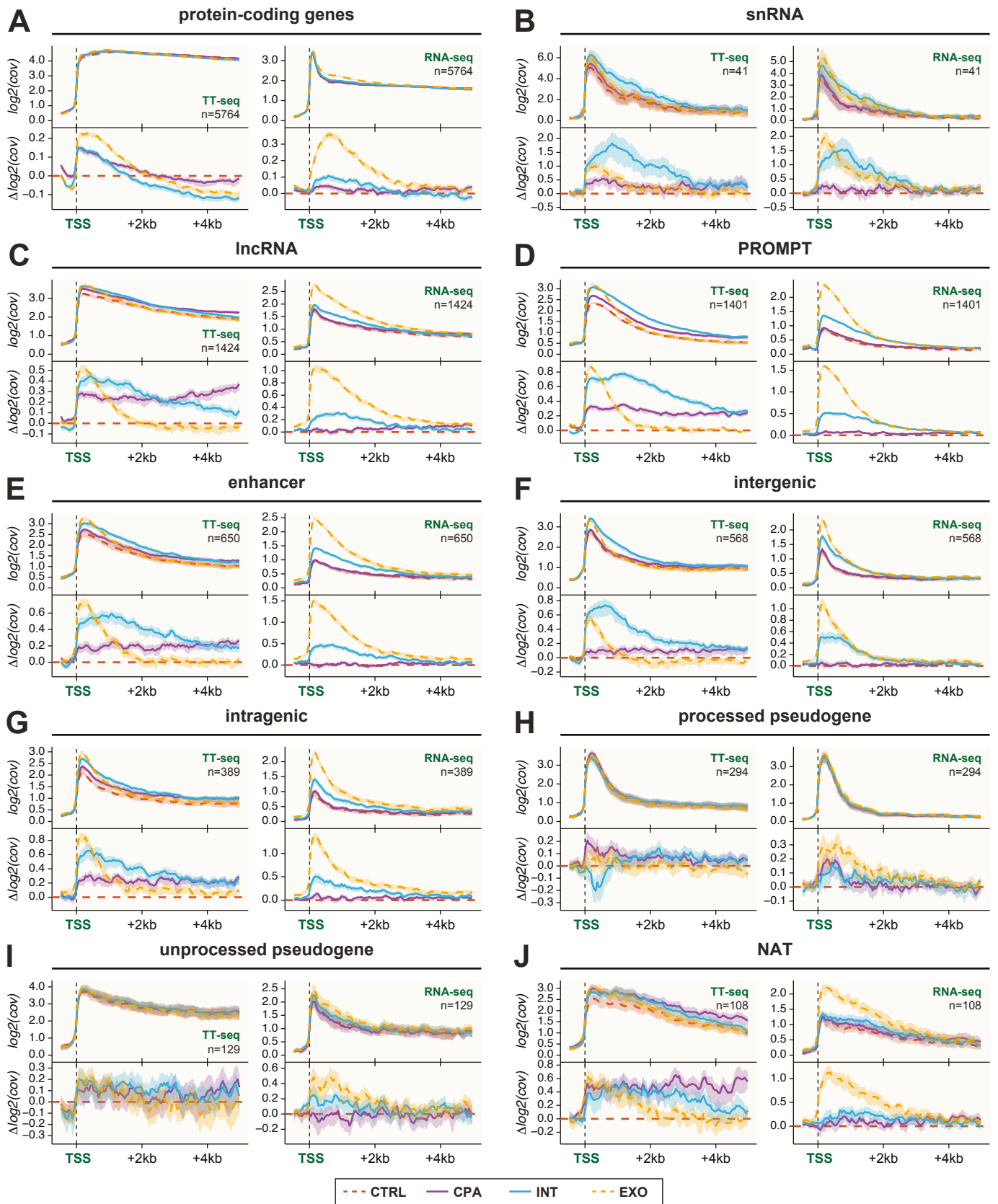


Figure S3, related to Figure 3. CPA and INT depletions cause increased promoter-proximal transcription globally

(A)-(J) Aggregate plots of mean TT-seq (*left*) and RNA-seq (*right*) data aligned to TSSs within TUs from major RNA TU type-classes represented in the HeLa transcriptome annotation: (A) protein-coding, (B) snRNA, (C) lncRNA, (D) PROMPT, (E) enhancer, (F) intergenic, (G) intragenic, (H) processed pseudogene, (I) unprocessed pseudogene and (J) NAT. Representation as in Figure 3A.

Figure S4

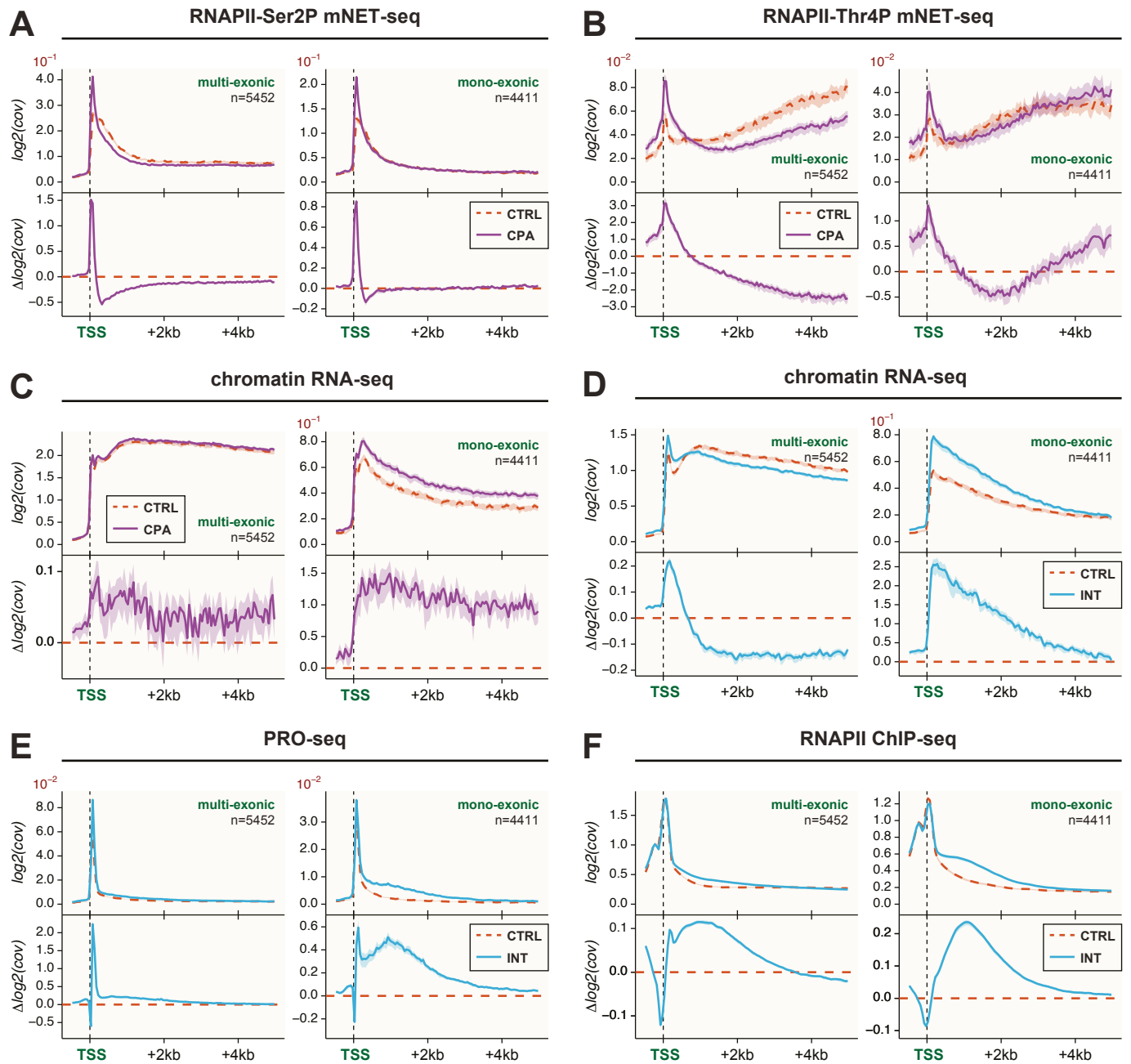


Figure S4, related to Figure 3. CPA and INT depletions cause increased promoter-proximal transcription globally

(A) Aggregate plots of mean Ser2P-RNAPII mNET-seq data (Nojima et al., 2015) for CPA depletion (*purple*) and CTRL (*red*) aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Mean values and 95% confidence intervals of log2-transformed coverage for 50-bp bins over all included TUs are displayed. The number of aggregated loci (n) is indicated.

(B) Aggregate plots of mean Thr4P-RNAPII mNET-seq data (Schlackow et al., 2017) for CPA depletion and CTRL aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Representation as in (A).

(C) Aggregate plots of mean chromatin RNA-seq data (Nojima et al., 2015) for CPA depletion and CTRL aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Representation as in (A).

(D) Aggregate plots of mean chromatin RNA-seq data (Lai et al., 2015) for INT depletion (*cyan*) and CTRL (*red*) aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Representation as in (A).

(E) Aggregate plots of mean PRO-seq data (Beckedorff et al., 2020) for INT depletion and CTRL aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Representation as in (A).

(F) Aggregate plots of mean RNAPII ChIP-seq data (Beckedorff et al., 2020) for INT depletion and CTRL aligned at the TSS of multi-exonic (*left*) and mono-exonic (*right*) TUs. Representation as in (A).

Figure S5

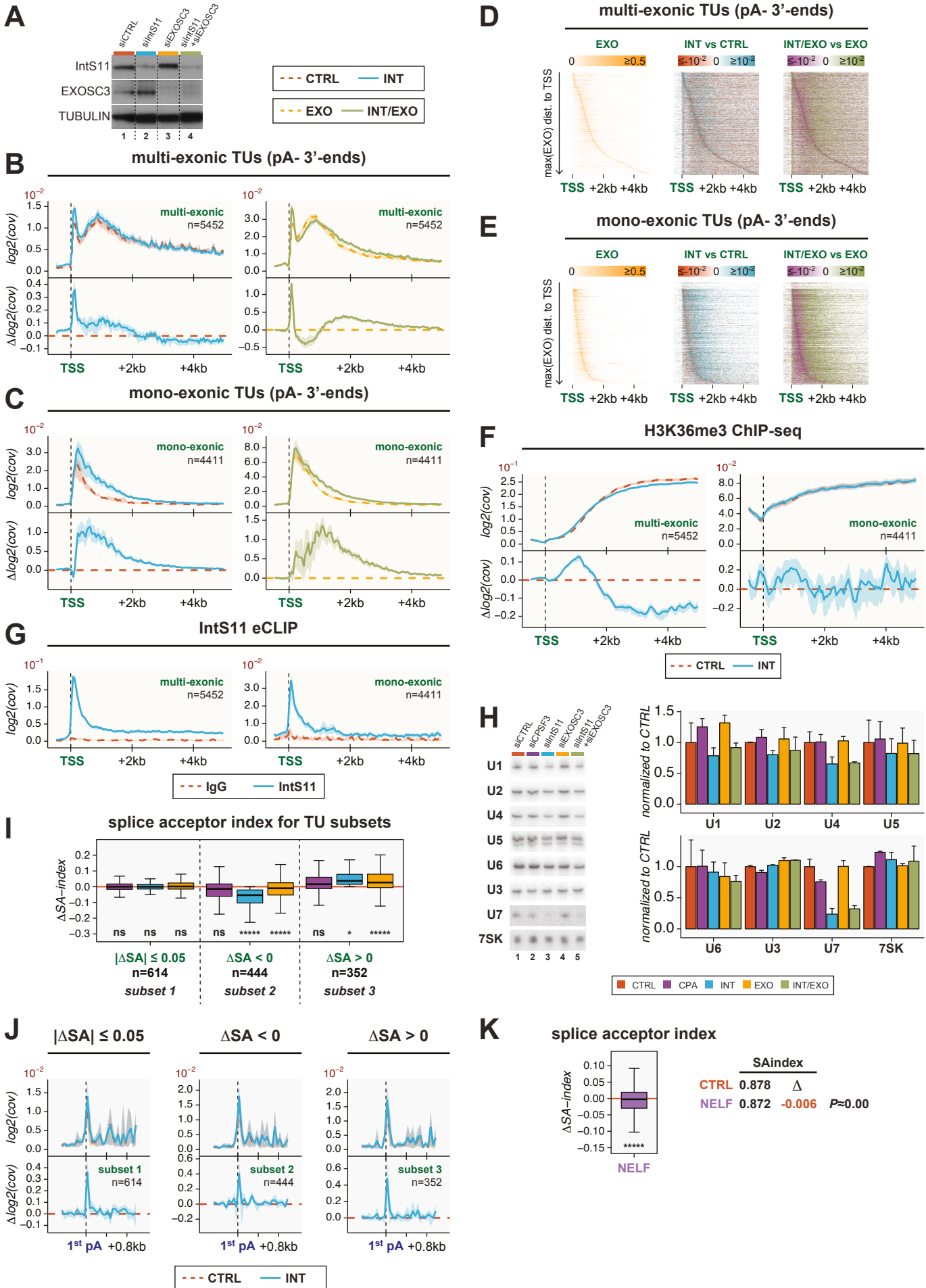


Figure S5, related to Figure 4 and 5. INT depletion exposes non-productive RNAPII transcription events

(A) Western blotting analysis demonstrating efficient depletion of indicated proteins. Tubulin was probed as loading control.

(B)-(C) Aggregate plots as in Figure 4B-C, but of pA⁻ 3'-end-seq data.

(D)-(E) Heatmaps as in Figure 4D-E, but of pA⁻ 3'-end-seq data.

(F) Aggregate plot of H3K36me3 ChIP-seq data (Beckedorff et al., 2020), showing mean coverage for INT (*cyan*) and CTRL (*red*) samples as in Figure 2E-F.

(G) Aggregate plots of Ints11 eCLIP data (Barra et al., 2020), showing mean coverage of (*left*) multi- and (*right*) mono-exonic TUs. Representation as in Figure 2C.

(H) Northern blotting analysis of mature snRNA levels in the indicated depletion samples (*left*). 7SK RNA was used as a loading control. Quantification of Northern blot signals (*right*). Error-bars represent the standard deviation (n=2).

(I) Box plot displaying the changes in splice acceptor indices (SA) for CTRL (*red*), CPA (*purple*), INT (*cyan*) and EXO (*orange*) depletions for the subset of multi-exonic TUs with $|\Delta SA| \leq 0.05$, $\Delta SA < 0$ or $\Delta SA > 0$ for INT depletion samples, respectively. Box limits represent the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range.

(J) Aggregate plots of pA⁺ 3'-end-seq data aligned at the first encountered putative pA signal (1st pA) of multi-exonic TUs with $|\Delta SA| \leq 0.05$ for INT depletion (*cyan*) and CTRL (*red*) samples. Mean values and 95% confidence intervals of log2-transformed coverage for 50-bp bins are displayed.

(K) as 5F, but displaying the changes in splice acceptor indices (ΔSA) for NELF-E depletion relative to CTRL (Stadelmayer et al., 2014).

For the box plots in (I) and (K) the ΔSA distributions were subjected to Wilcoxon tests of significance. *p < 0.05, **p < 0.01, ***p < 0.005, ****p < 0.001, *****p < 2.2e-16, ns, not significant.

Figure S6

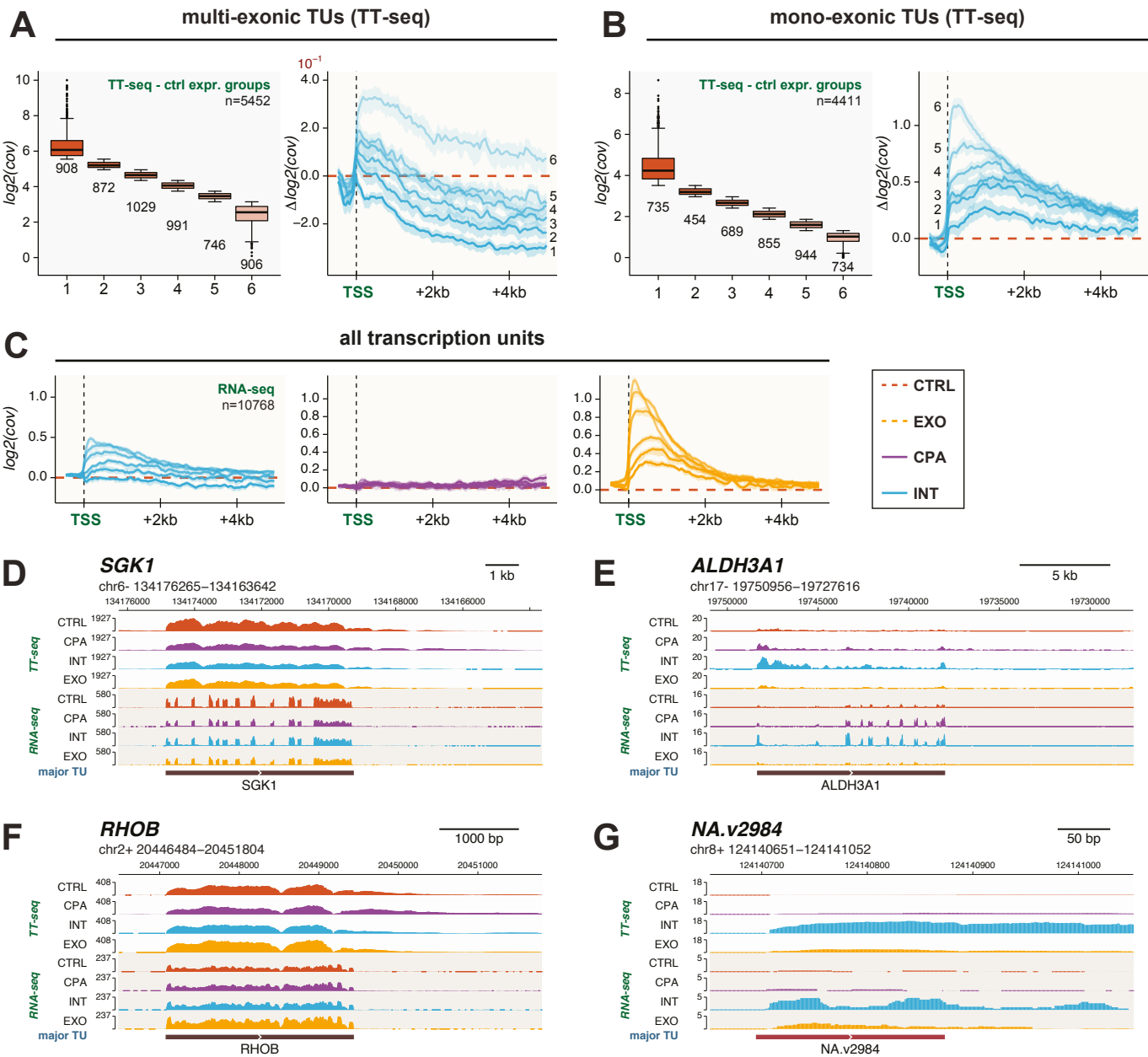


Figure S6, related to Figure 6. INT activity correlates inversely with transcriptional output

(A) Box plot displaying expression levels for six expression-based multi-exonic TU groups (1-6) based on expression in CTRL (*left*). Box limits represent the first and third quartiles, the band inside the box is the median. The ends of the whiskers extend the box by 1.5 times the interquartile range. Outliers are shown in black. The number of TUs in each group is indicated. Aggregate plots (*right*) of the log2-transformed TT-seq Δ coverage of INT (*cyan*) depletion vs. CTRL aligned at the major TSS. Mean values of log2-transformed Δ coverage for 50-bp bins are displayed. The line color intensity (high to low) corresponds to the gene group 1 to 6, respectively.

(B) As (A), but for six expression-based mono-exonic TU groups (1-6) based on expression in CTRL (*left*).

(C) Aggregate plots as in Figure 6B, but for RNA-seq data.

(D)-(G) Genome browser views of (D) *SGK1* (multi-exonic, group 1), (E) *ALDH3A1* (multi-exonic, group 6), (F) *RHOB* (mono-exonic, group 1) and (G) *NA.v2984* (mono-exonic, group 6). Representation as in Figure 2B.

Figure S7

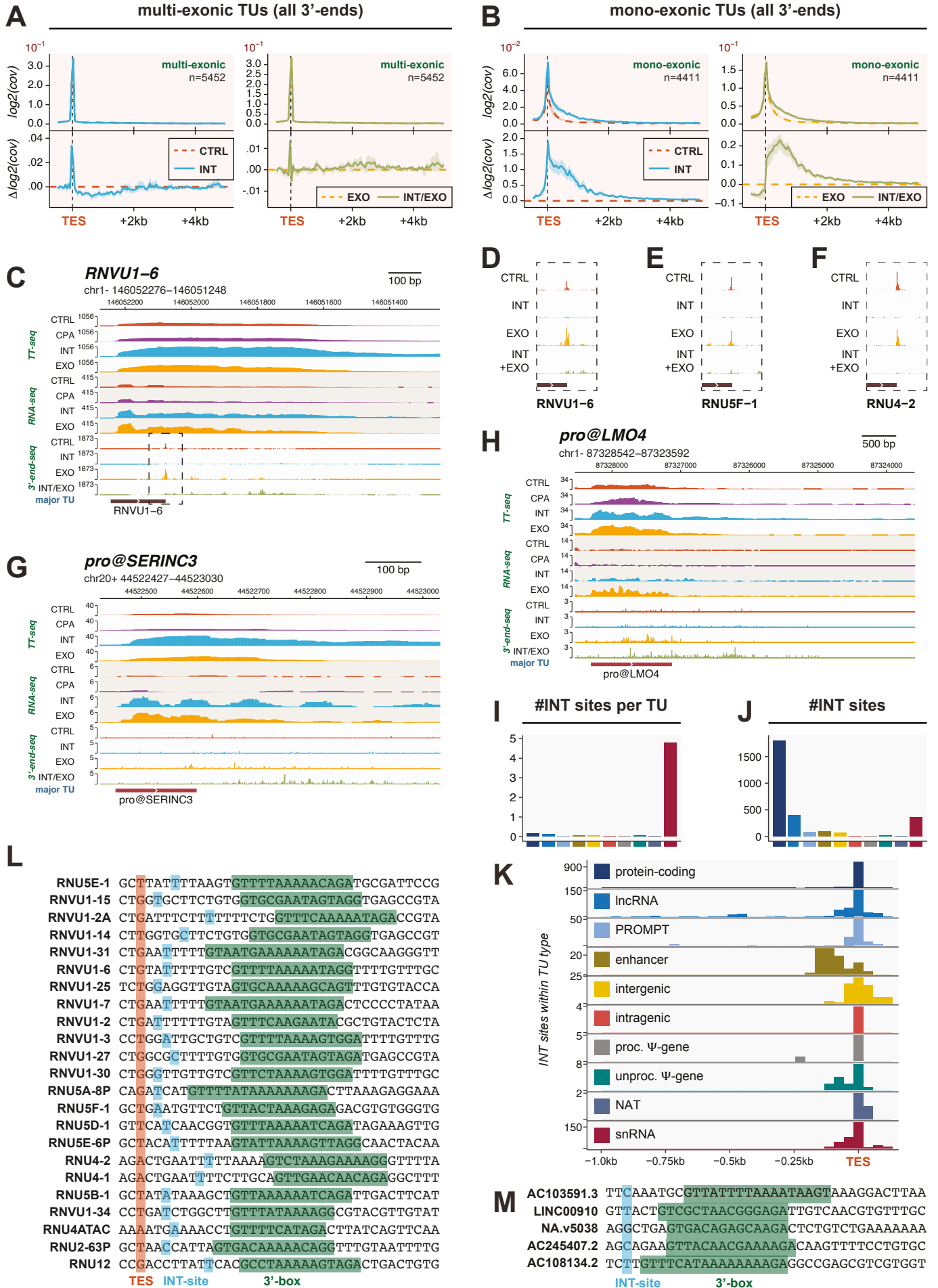


Figure S7, related to Figure 7. snRNA TUs exploit INT-sensitive transcription for their stable RNA production

(A)-(B) Aggregate plots as in Figure 7A-B, but for multi-exonic and all mono-exonic TUs, respectively.

(C) Genome browser view of *RNVU1-6* (snRNA) and downstream region. Mean TT-seq and RNA-seq coverage (of two replicates) from control (CTRL, *red*), INT (*cyan*) and EXO (*yellow*) depleted HeLa cells are shown. Additionally, mean 3'-end-seq data (of three replicates) from control (CTRL, *red*), INT (*cyan*), EXO (*yellow*) and INT/EXO (*olive green*) depleted HeLa cells are shown. Rectangle with dashed outline indicate region that is enlarged in (D).

(D) Enlarged plots of the 3'-end-seq data at the 3'-end of *RNVU1-6*.

(E)-(F) as in (D), but for the snRNA TUs *RNU5F-1* and *RNU4-2*.

(G)-(H) As (C), but for the PROMPT TUs *pro@SERINC3* and *pro@LMO4*.

(I)-(K) As Figure 7C-E, but for the 'top10' TU types as defined in Figure S1.

(L) Sequences from the indicated snRNA TUs aligned to TES (shaded red; positions -2 to +34 are shown). TUs with INT-sensitive site -5 to +10nt from TES are shown. INT-sensitive 3'-end sites (INT-site) are shaded cyan. Putative 3'-boxes (based on Hernandez, 1985) are shaded green. Motif emerging from here: INT-sensitive site $-N_{2-9}GTttN_{0-3}AAaADN_{1-2}AGR$ (where upper- and lowercase letters indicate strict and less strict adherence to motif, N is any nucleotide, D=A,G or T, R=A or G). Note the most significant INT-sensitive site is often, but not always the most abundant 3'-end in the region.

(M) As (L), but for selected non-snRNA TUs. Sequences are aligned to the INT-sensitive 3'-end site. *AC103591.3* (lncRNA; also known as *pro@DNAJB4*), *LINC00910* (lncRNA), *NA.v5038* (intergenic), *AC245407.2* (lncRNA) and *AC108134.2* (lncRNA).