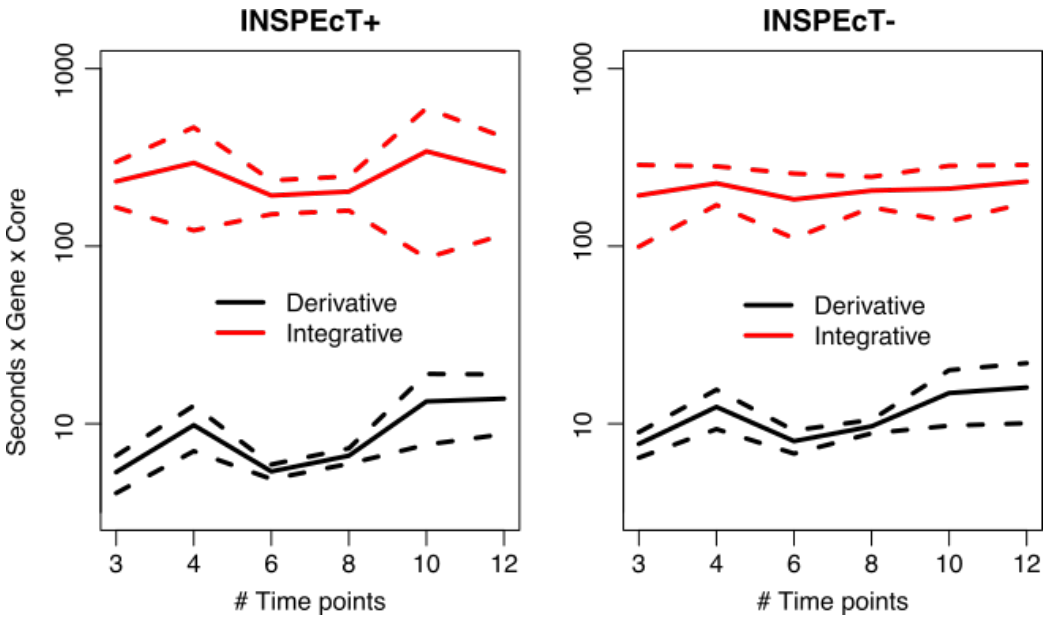
Genome-wide dynamics of RNA synthesis, processing and degradation without RNA metabolic labeling

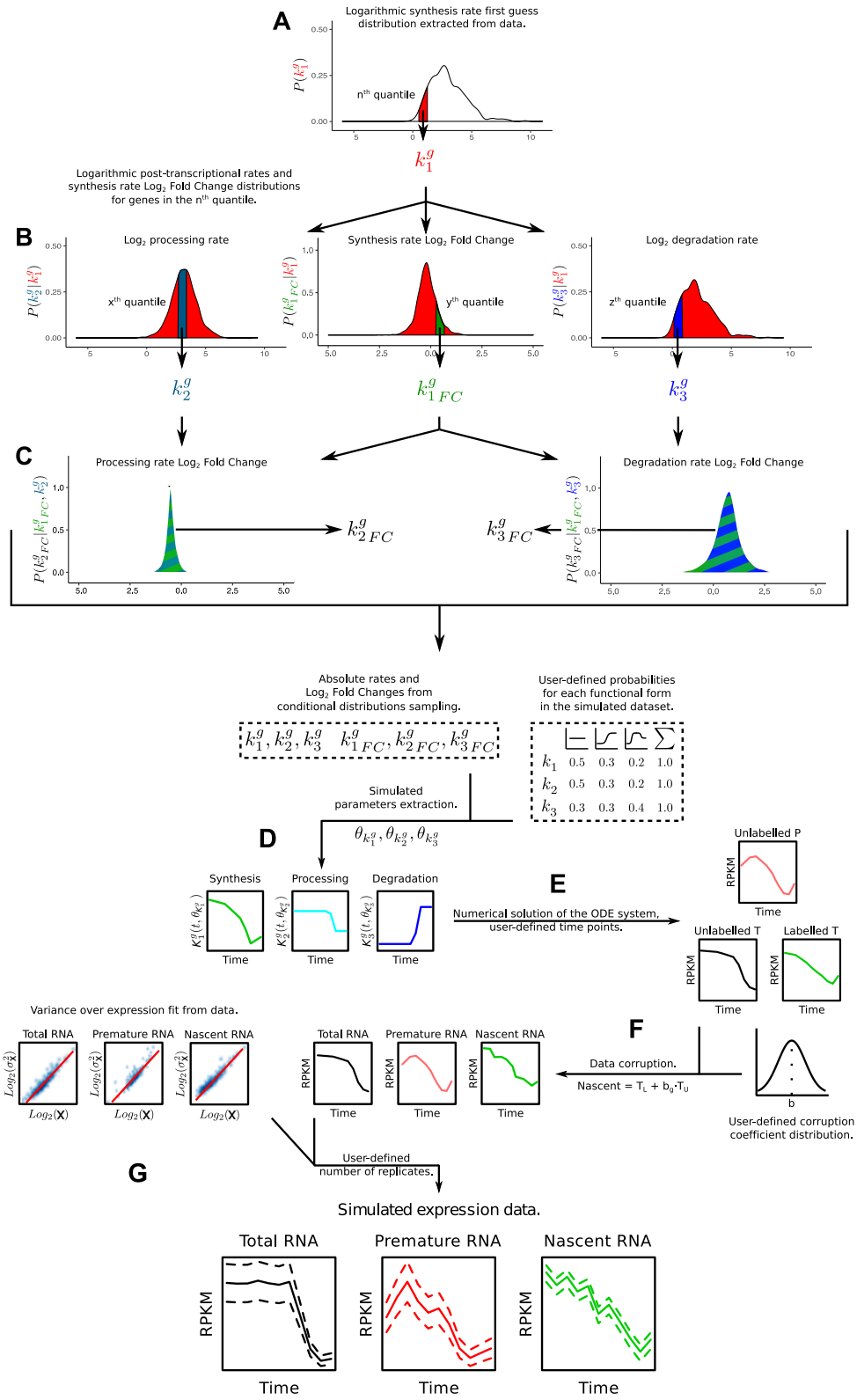
Supplemental Figures

M. Furlan, E. Galeota, N. Del Gaudio,
E. Dassi, M. Caselle, S. de Pretis, M. Pelizzola

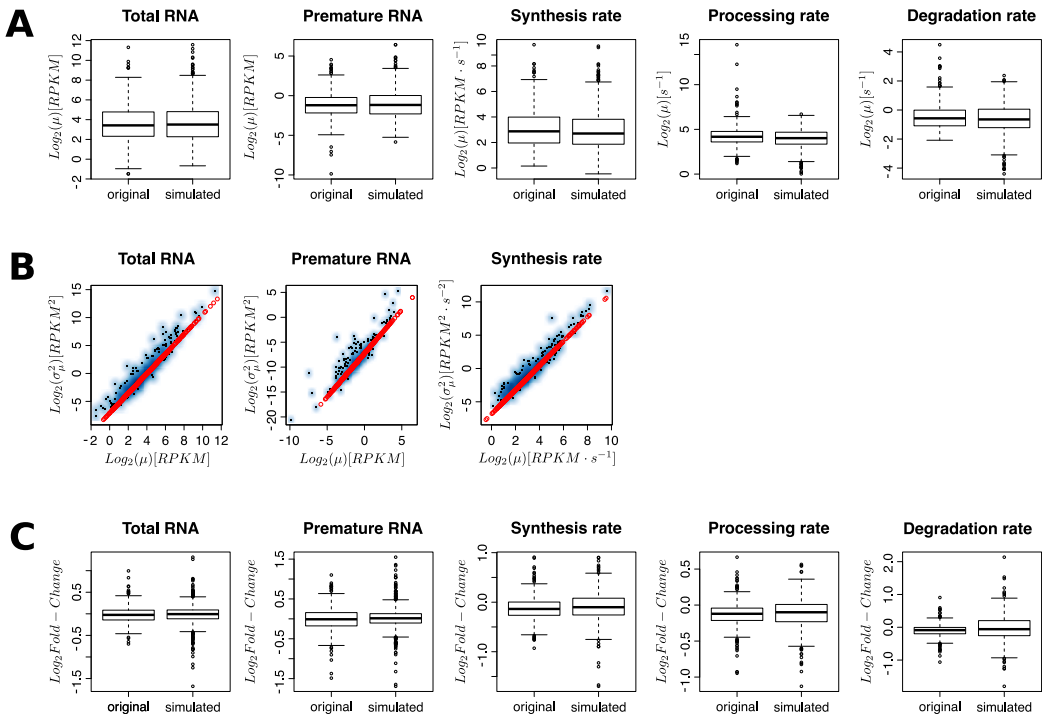
Contents



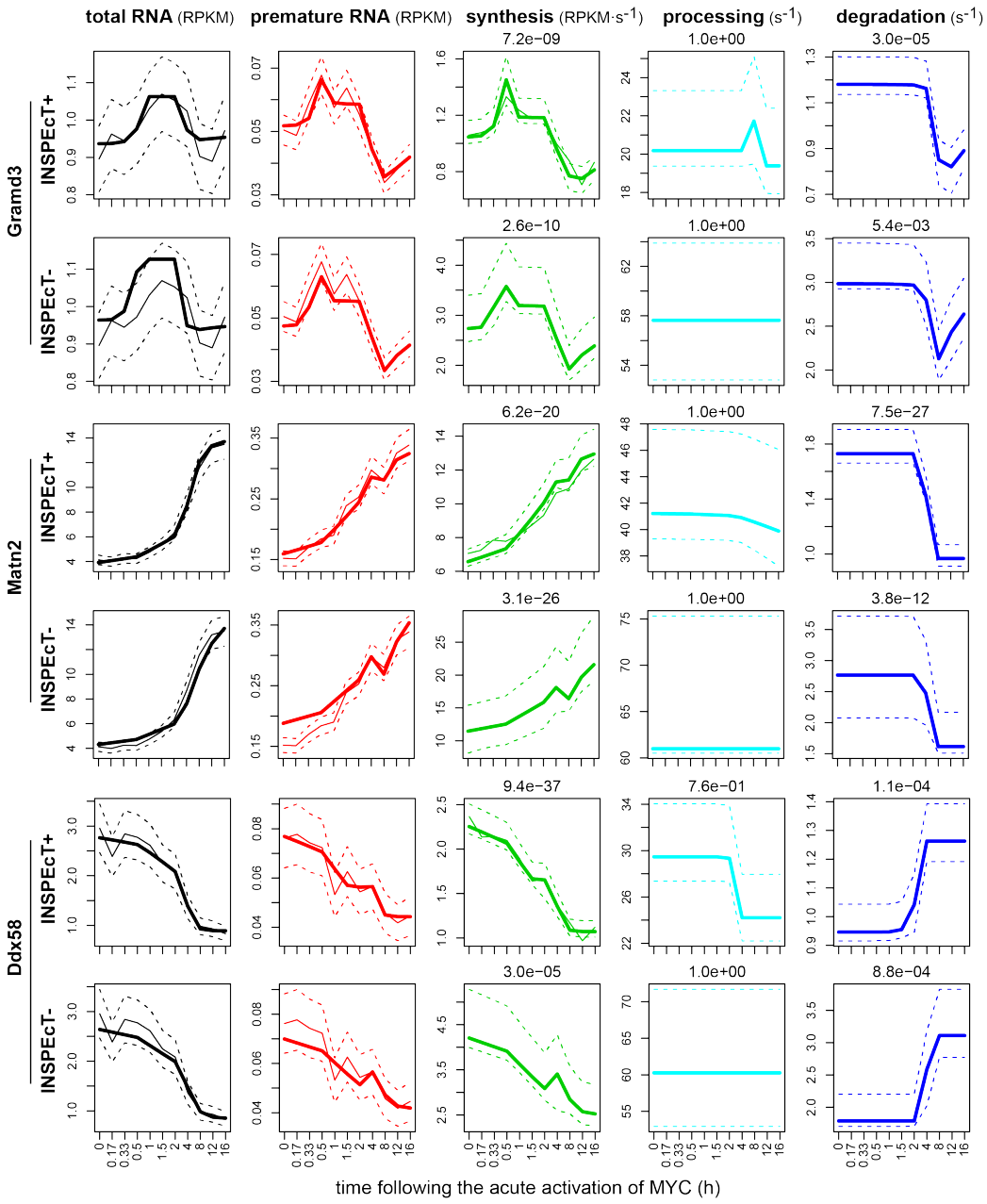
Supplemental Figure 1: **INSPEcT run times per gene.** On the left it is reported the computational times required by INSPEcT+ with derivative and integrative approaches. On the right, the same data are depicted for INSPEcT-. Solid lines represent mean values over ten replicates; dashed lines represent the standard deviation of the mean.


A
Logarithmic synthesis rate first guess distribution extracted from data.
n^th quantile
B
Logarithmic post-transcriptional rates and synthesis rate Log2 Fold Change distributions for genes in the n^th quantile.
Log2 processing rate
x^th quantile
Synthesis rate Log2 Fold Change
y^th quantile
Log2 degradation rate
z^th quantile
C
Processing rate Log2 Fold Change
Degradation rate Log2 Fold Change
Absolute rates and Log2 Fold Changes from conditional distributions sampling.
User-defined probabilities for each functional form in the simulated dataset.
Simulated parameters extraction.
D
Synthesis
Processing
Degradation
Time
E
Numerical solution of the ODE system, user-defined time points.
Unlabelled P
Unlabelled T
Labelled T
RPKM
F
Data corruption.
Nascent = T_L + b_g·T_U
User-defined corruption coefficient distribution.
Variance over expression fit from data.
Total RNA
Premature RNA
Nascent RNA
G
User-defined number of replicates.
Simulated expression data.

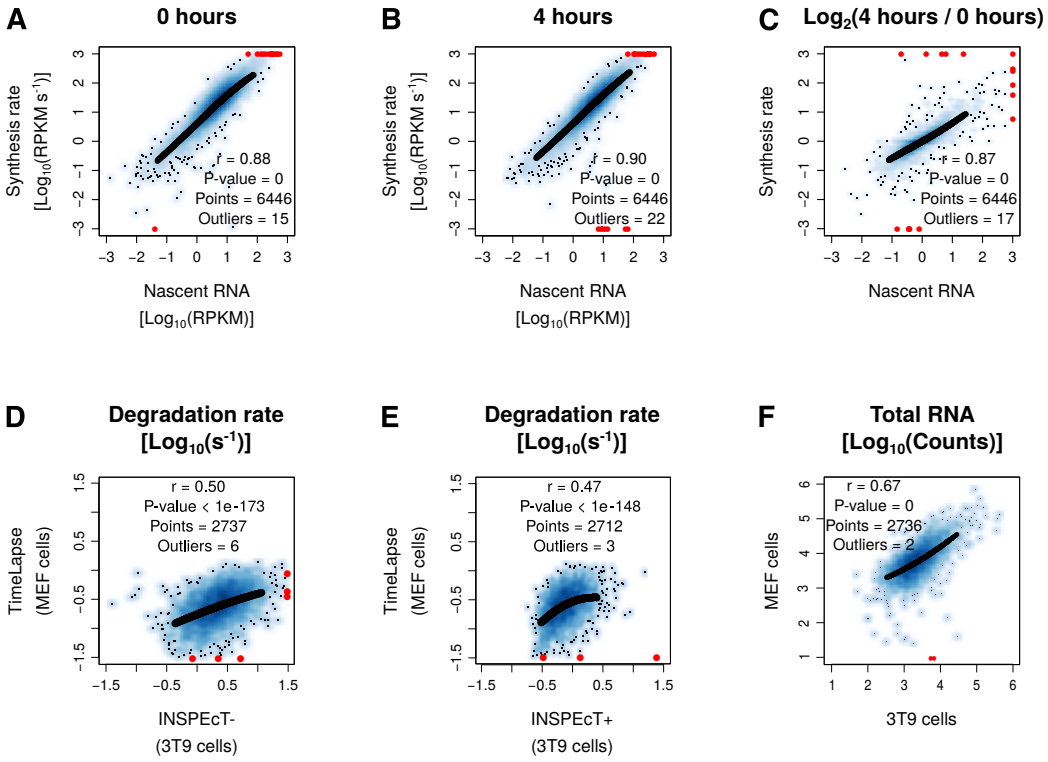
Supplemental Figure 2: **Generation of INSPEcT simulated data.** The procedure is implemented in the makeSimModel and makeSimDataset INSPEcT methods, and requires the following input: (i) an INSPEcT dataset that includes nascent RNA data, (ii) the number of genes and replicates to be simulated, (iii) a set of time points, (iv) a contamination coefficient together with its dispersion parameter, and (v) the probability of each kinetic rate to be constant, sigmoidal or impulsive. (A) The procedure is performed independently for each gene (g). It starts by sampling the distribution of first guess synthesis rates. (B) In order to preserve the correlations among the RNA kinetic rates of the input data, all the genes belonging to the selected quantile are considered. Specifically, the distributions of their processing and degradation rates are determined, together with their synthesis rate fold changes. These distributions are independently sampled to determine these quantities for the gene g (empirical distribution conditioning). (C) The procedure is repeated to sample fold changes of processing and degradation rates from their respective conditional distributions. (D) Rates temporal profiles are generated based on their magnitude and fold change (the six values returned by the initial part of the procedure), according to the probabilities provided by the user for each functional form. (E) Once the simulated kinetic rates are defined, INSPEcT estimates the amount of total and premature RNA within both labeled and unlabeled conditions, by solving the Ordinary Differential Equations (ODE) system at each time point. (F) Labeled RNA is optionally corrupted, according to the equation shown in the figure, to simulate its contamination due to unlabeled transcripts. The gene specific corruption coefficient is extracted from a Gaussian distribution whose first and second moments can be defined by the user. For this study, these were set to match 30% contamination rate (based on Main Fig. 2 results). (G) Finally, to simulate the experimental replicates, we determine the variance expected for each expression datum, based on the power-law global error model implemented in the PLGEM package.



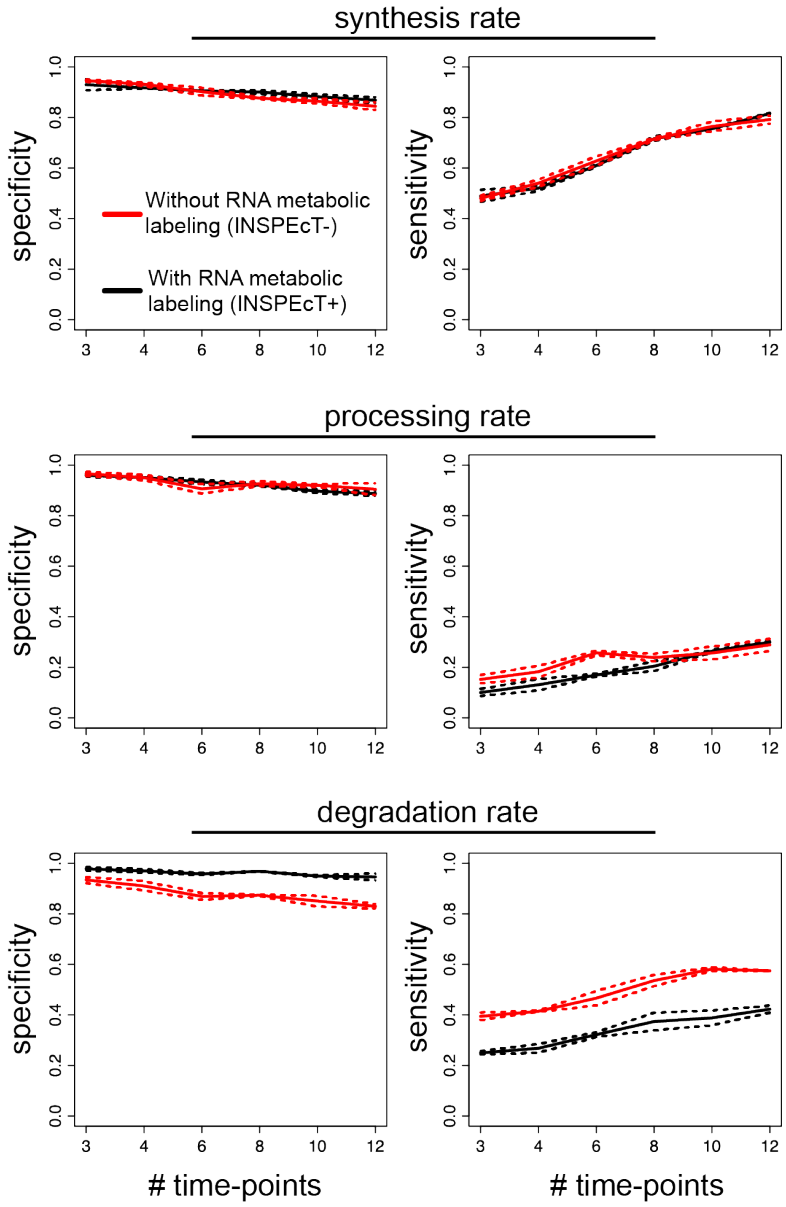
Supplemental Figure 3: **Validation of INSPEcT simulated data.** Absolute levels, variance, and temporal changes of simulated data for 1000 genes (11 time points and 3 replicates) are compared with the original data, a subset of the data from S. de Pretis et al., Genome Research 2017. Simulated data are generated with the makeSimDataset and makeSimModel functions of INSPEcT, as detailed in the supplementary source code. Boxplots in the first row display Log_2 expression data and first guess kinetic rates for original (left) and simulated (right) data. Scatterplots in the second row display the relation between experimental data and their variance in the Log_2-Log_2 space (density scatter plot in blue); overlaid are the linear fits (determined with the PLGEM package) used to assign a variance to the simulated data. Boxplots in the third row display mean Log_2 temporal changes for RNA species concentrations and kinetic rates.



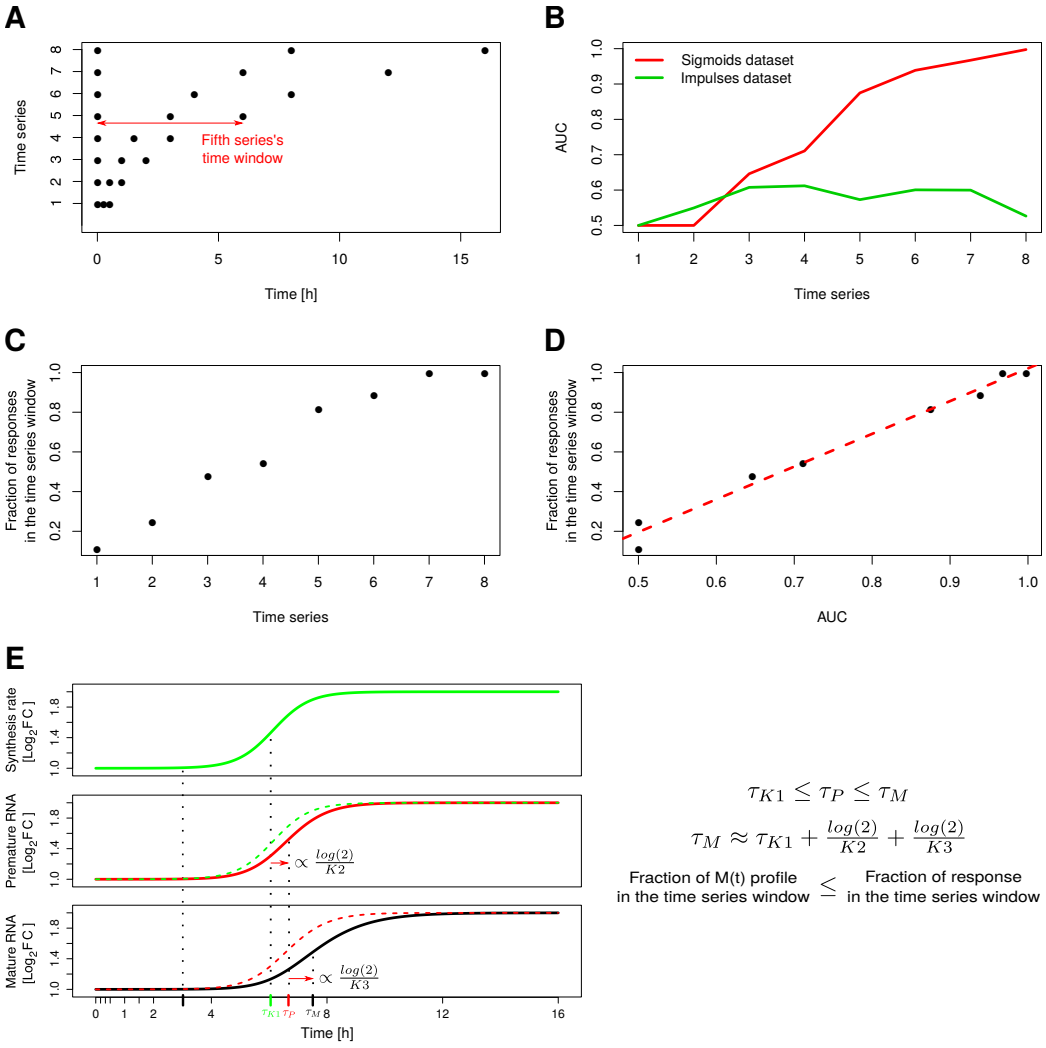
Supplemental Figure 4: **INSPEcT- vs INSPEcT+ RNA dynamics for three genes in 3T9 cells following the acute activation of MYC.** Solid bold lines indicate the model fit; tin solid and dashed lines indicate mean and standard deviation of experimental data for total and premature RNA, respectively; dashed lines indicate 95% confidence intervals for the kinetic rates models. P-values are reported on the top of each kinetic rate indicating the significance of a model in which the rate is varying over time.



Supplemental Figure 5: **Benchmark of synthesis and degradation rates quantified by INSPEcT- against alternative independent quantifications.** (A) Smooth-scatter plot, in the Log_{10}-Log_{10} space, comparing INSPEcT- synthesis rates and nascent RNA expression levels (untreated 3T9 murine fibroblasts). Spearman's correlation coefficient (r) and the associated P-value are reported inside the box, together with the number of genes involved in the analysis (genes transcriptionally regulated), and the number of outliers. Loess smoothing is superimposed. (B) Similar to panel A but for the 4 hours time point. (C) Similar to (A) but, instead of absolute values, we compare the Log_2 Fold Changes of synthesis rates and nascent RNA expression levels at 4 hours against the 0 hours condition. (D) Similar to (A), comparing degradation rates of INSPEcT- (3T9 murine fibroblasts) versus those resolved with TimeLapse (MEF mouse embryonic fibroblast). (E) Similar to (D) but for INSPEcT+ vs TimeLapse. (F) Similar to (D) for total RNA counts in MEF and 3T9 cells.



Supplemental Figure 6: **Specificity and sensitivity in the classification of variable rates with (INSPEcT+) or without (INSPEcT-) RNA metabolic labeling.** For each RNA kinetic rate, specificity and sensitivity are reported at increasing number of time points. Means and standard deviations are determined based on three simulated datasets.

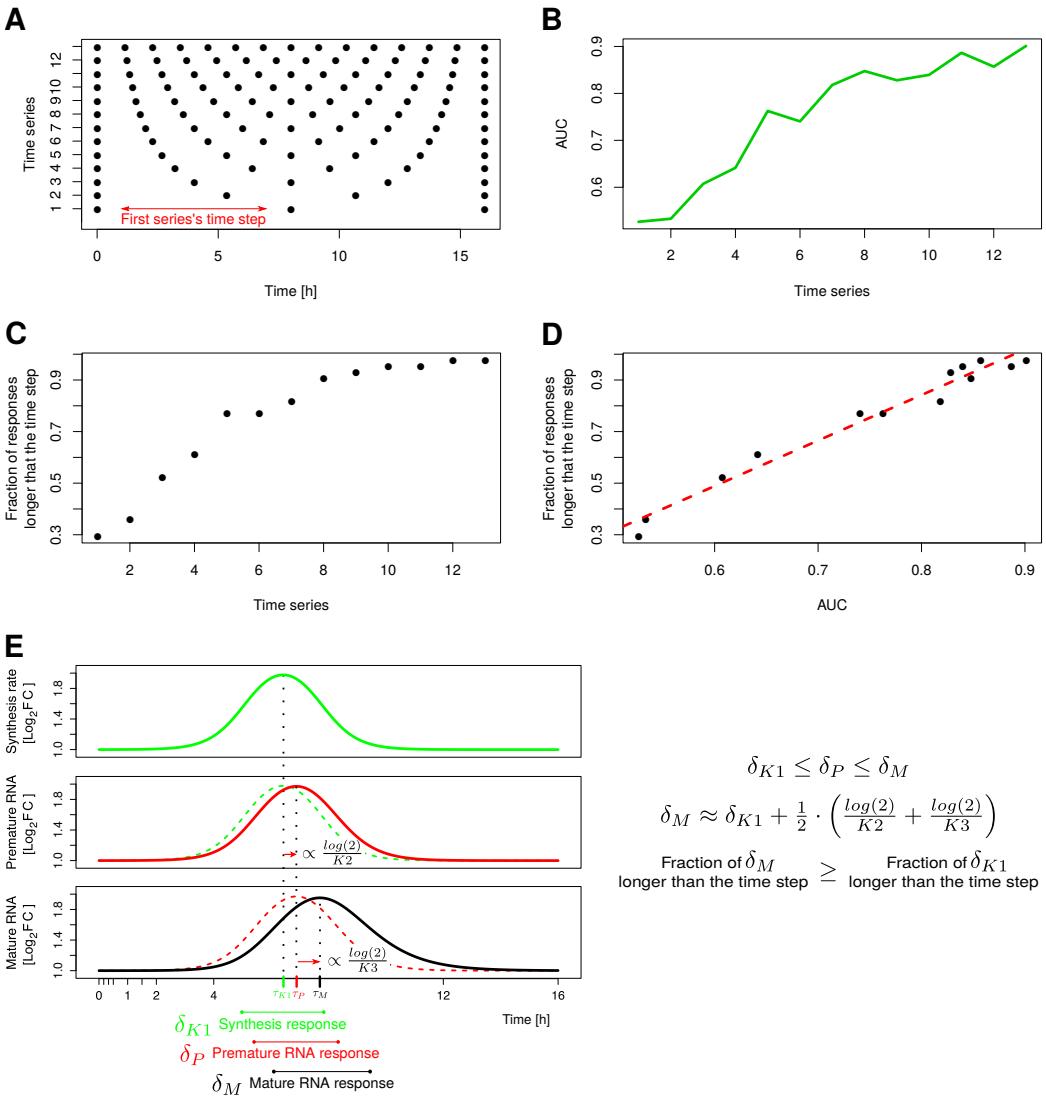


$$\tau_{K1} \leq \tau_P \leq \tau_M$$

$$\tau_M \approx \tau_{K1} + \frac{log(2)}{K2} + \frac{log(2)}{K3}$$

Fraction of M(t) profile in the time series window $\leq$ Fraction of response in the time series window

Supplemental Figure 7: **Impact of time series design on INSPEcT-classification performance on the sigmoidal dataset.** (A) Time series design, we selected three linearly separated time points between 0 and 0.5, 1, 2, 3, 6, 8, 12, 16 hours respectively. (B) Synthesis rate classification performance (AUC score) for different time series on the sigmoidal (red) and impulsive (green) datasets. (C) Fraction of sigmoidal synthesis rate half response times included in each time series. (D) Fraction of sigmoidal synthesis rate half response times against the corresponding synthesis rate's AUC. The red dashed line represents a linear fit of the data obtained with the lm functions from the R stat package. (E) Cartoon showing the relation between synthesis rate (green), premature RNA (red) and mature RNA (black) sigmoidal profiles for constant processing and degradation rates. The figure highlights the shift of the half response times, from synthesis to mature RNA, which is due to the effect of finite post-transcriptional rates. We also reported an approximated equation which links mature RNA half response time with the synthesis rate counterpart and the post-transcriptional rates.

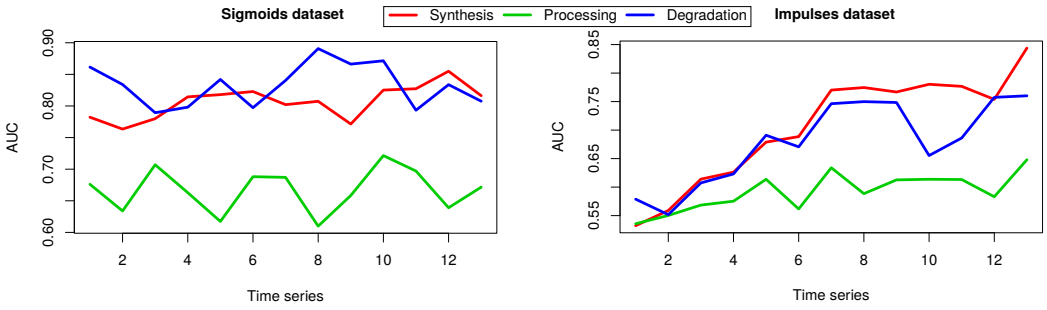


$$\delta_{K1} \leq \delta_P \leq \delta_M$$

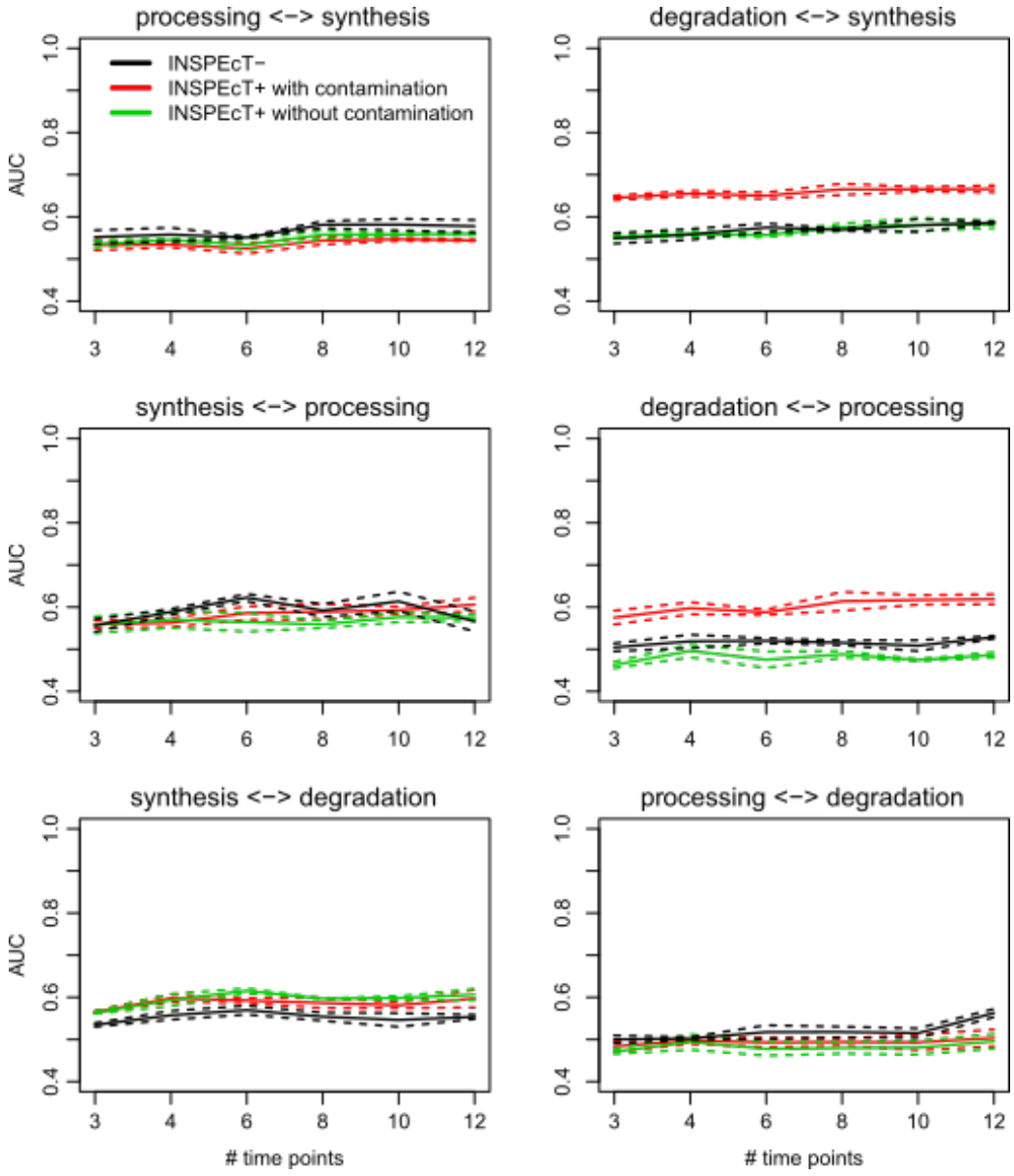
$$\delta_M \approx \delta_{K1} + \frac{1}{2} \cdot \left(\frac{log(2)}{K2} + \frac{log(2)}{K3} \right)$$

Fraction of δ_M longer than the time step $\geq$ Fraction of δ_{K1} longer than the time step

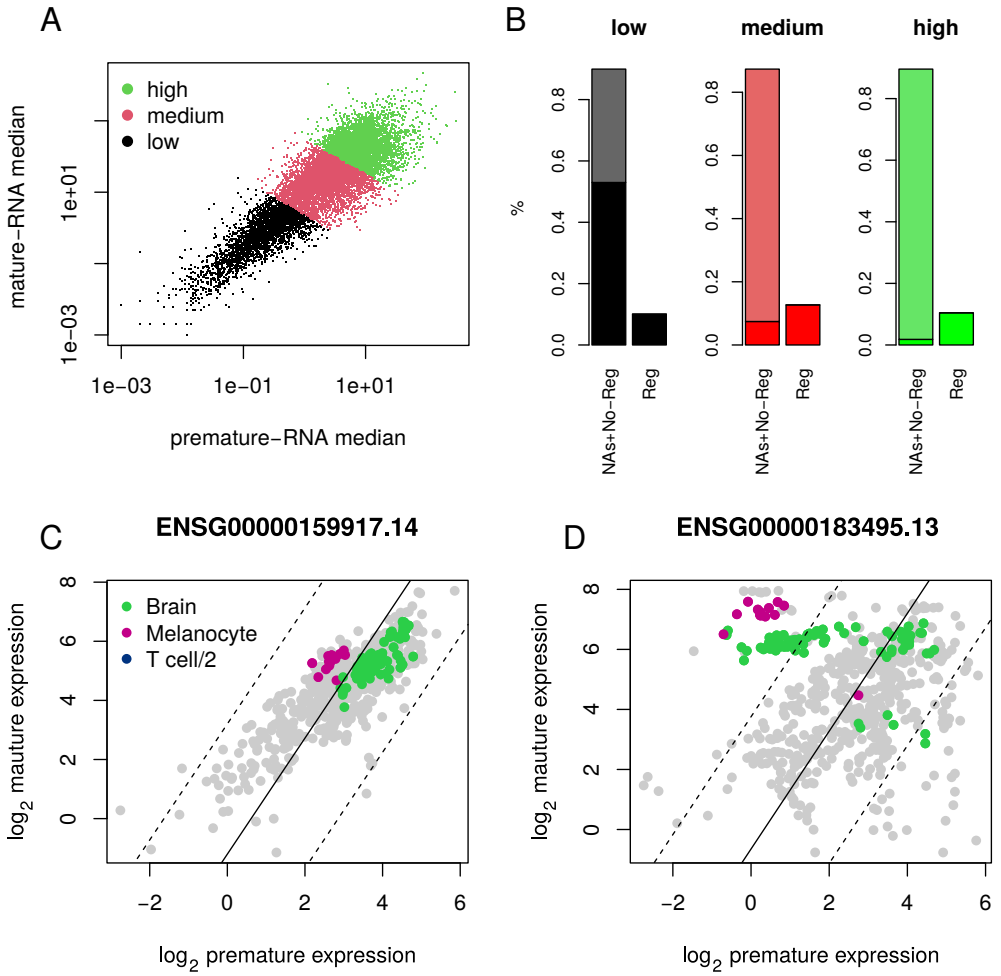
Supplemental Figure 8: **Impact of time series design on INSPEcT-classification performance on the impulsive dataset.** (A) Time series design, we selected a variable number of linearly separated time points, from 3 up to 15, between 0 and 16 hours. (B) Synthesis rate classification performance (AUC score) for different time series on the impulsive datasets. (C) Fraction of synthesis rate response intervals, defined as the difference between the second and the first half response times, larger than the time series' step. (D) Fraction of synthesis rate response intervals larger than the time series' step against the corresponding synthesis rate's AUC. The red dashed line represents a linear fit of the data obtained with the lm functions from the R stat package. (E) Cartoon showing the relation between synthesis rate (green), premature RNA (red) and mature RNA (black) impulsive profiles for constant processing and degradation rates. The figure highlights the expansion of the response intervals, from synthesis to mature RNA, which is due to the effect of finite post-transcriptional rates. We also reported an empirical approximated equation which links mature RNA response intervals with the synthesis rate counterpart and the post-transcriptional rates. (F) Impact of time points position, compared to the transcriptional impulsive response (black histogram), on the synthesis rates' AUC score.



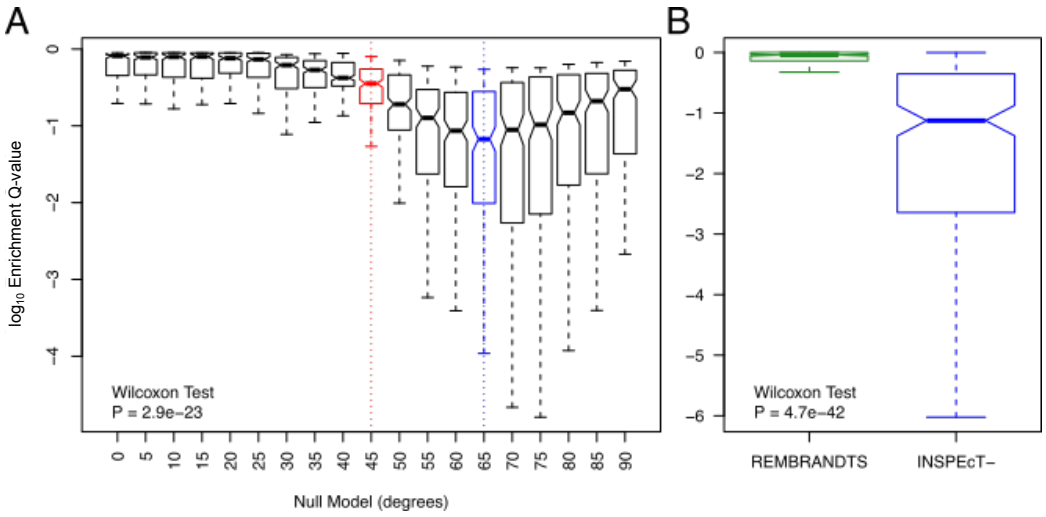
Supplemental Figure 9: **Impact of time series design on INSPEcT-classification performance for post-transcriptionally regulated datasets.** AUC scores for synthesis (red), processing (green) and degradation (blue) rates as a function of the time series' length; sigmoidal modulations on the left and impulsive modulations on the right. Below each box, we reported the pearson correlation coefficient computed comparing these two quantities and the associated P value; a joint test was performed for all the data-points displayed in each box.



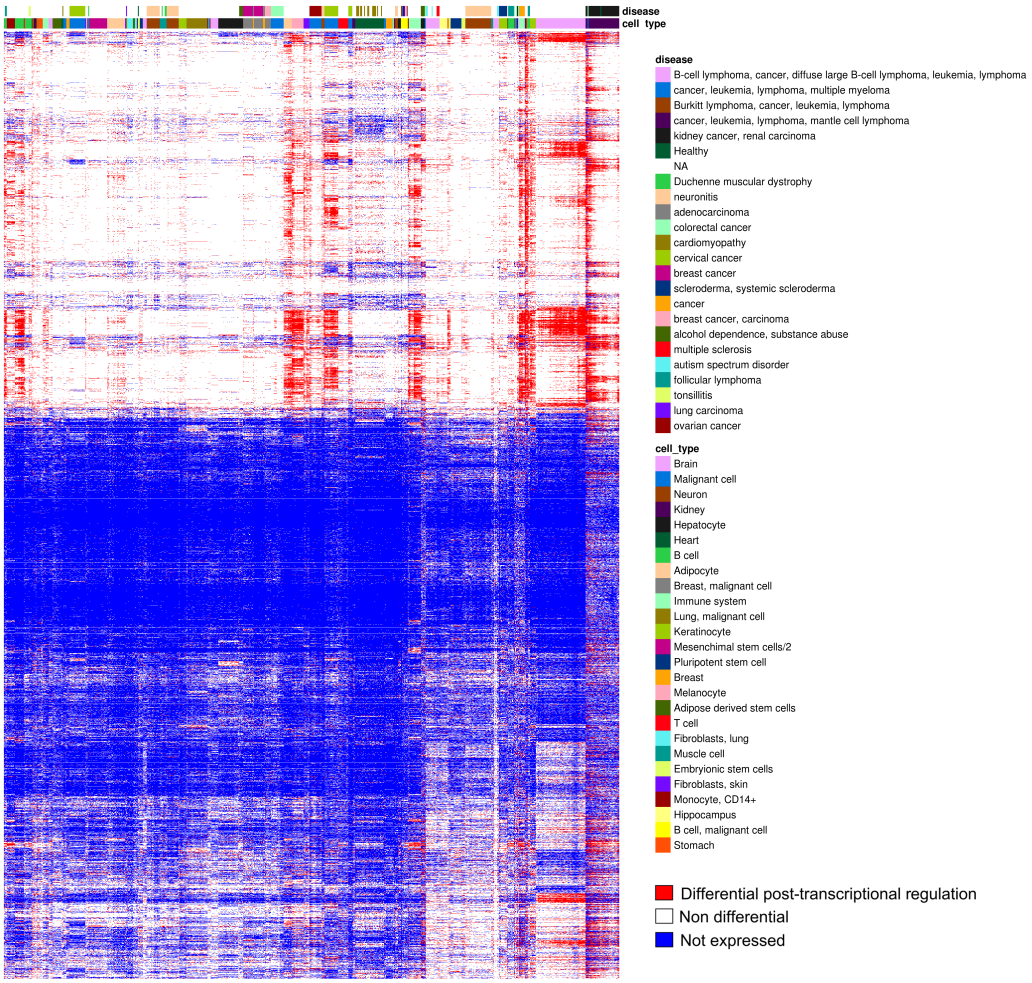
Supplemental Figure 10: **Analysis of the indetermination within INSPEcT- and INSPEcT+.** AUCs obtained by attempting to classify the variability of a given rate (on the left of the arrow in each title) with the estimated variability of another rate (indicated on the right of the arrow in each title). INSPEcT+ (based on simulated data with and without contamination) and INSPEcT– approaches are compared as a function of the number of time points in the simulated dataset. Solid lines are representative of the mean performance over three datasets (1000 simulated genes each); dashed lines show the standard deviation of the mean.



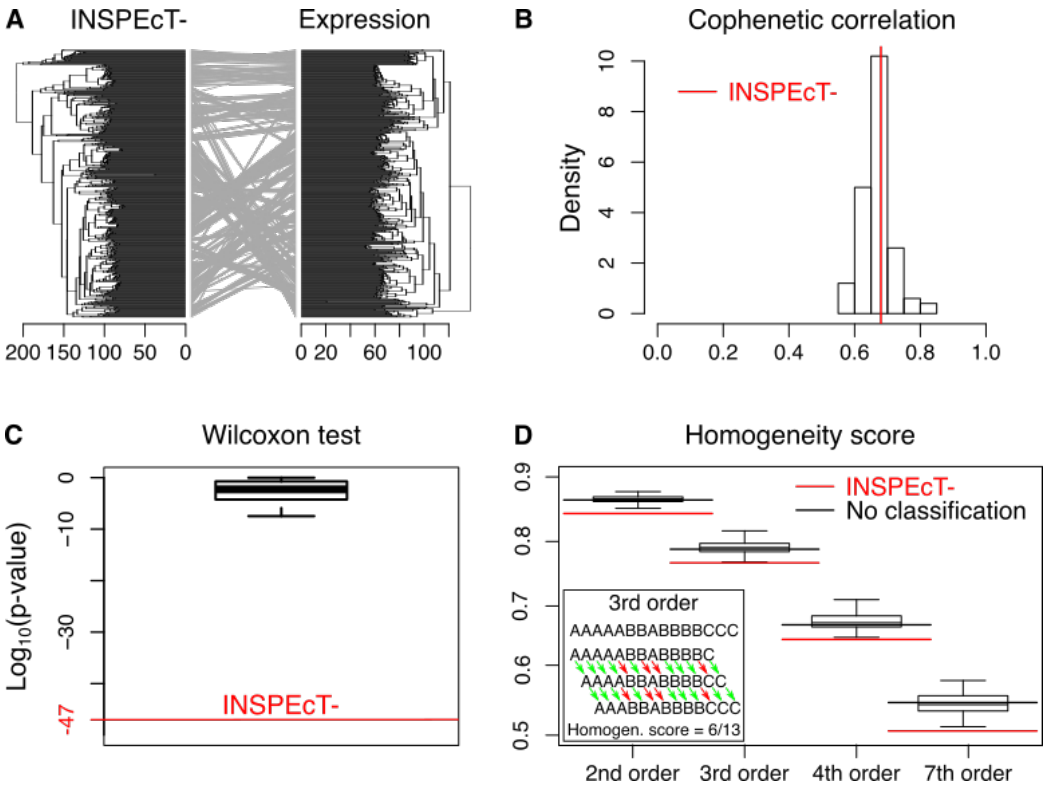
Supplemental Figure 11: **Deviation from mature-vs-premature trend is not biased by expression.** (A) Premature and mature RNA median expression of 16'182 protein-coding genes quantified in 669 samples downloaded from Recount are divided in three regimes of expression ("low", "medium" and "high"). (B) Proportion of regulated genes in the three regimes of expression defined in (A). Per each regime, the first bar sums up the proportion of not expressed (NAs, darkest bar) and not regulated genes (No-Reg, brightest bar), while the second bar represents regulated genes (Reg). (C) Dot-plot of the Log_2 premature and premature expression in the same samples described in (A) of the Ensembl gene ENSG00000159917.14. Samples derived from "Brain", "Melanocyte" and "T cell" annotations, are represented with distinct colors. The NULL model derived from the mature-vs-premature trend in the Recount dataset is reported (solid line) together with the corresponding confidence intervals (dashed lines). (D) Same as (C) but for Ensembl gene ENSG00000183495.13, which spans the same expression levels of the gene 9310 but shows evidence for tissue specific post-transcriptional regulation.



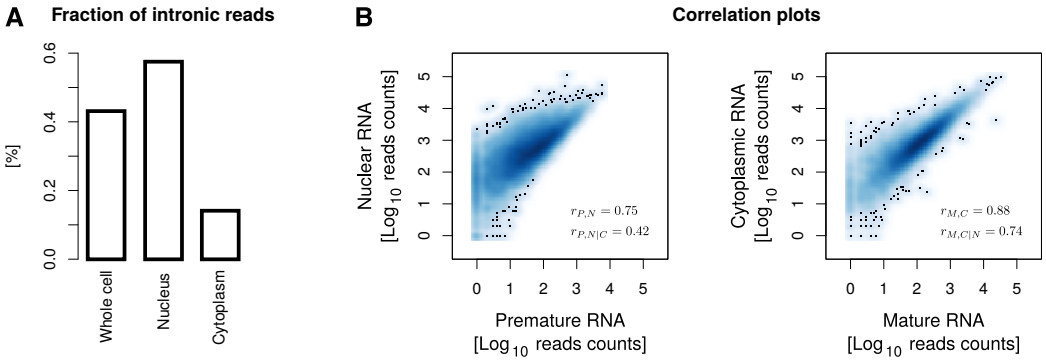
Supplemental Figure 12: **Enrichments of miRNA targets among the post-transcriptionally regulated genes identified by INSPEcT- and REMBRANDTS.** (A) Boxplots of Log_{10} enrichment q-values for 221 classes of miRNA targets annotated in the MIR_Legacy database among the 2'500 most regulated genes identified by INSPEcT- using different NULL models (shown on the X-axis). The analyses were run on a set of 16'182 protein-coding genes quantified in 669 samples downloaded from Recount. The red box-plot corresponds to the lack of NULL model, the blue box-plot corresponds to the NULL model derived with INSPEcT- based on the P-M trend of this gene set. (B) Same as (A), based on the genes identified as regulated by REMBRANDTS (filtered with absolute Log_2 variation of RNA half-lives greater than 1); for INSPEcT- the NULL model from the P-M trend for this geneset is adopted. The analyses were run on a subset of 8'283 genes and 602 samples that could be analyzed by REMBRANDTS.



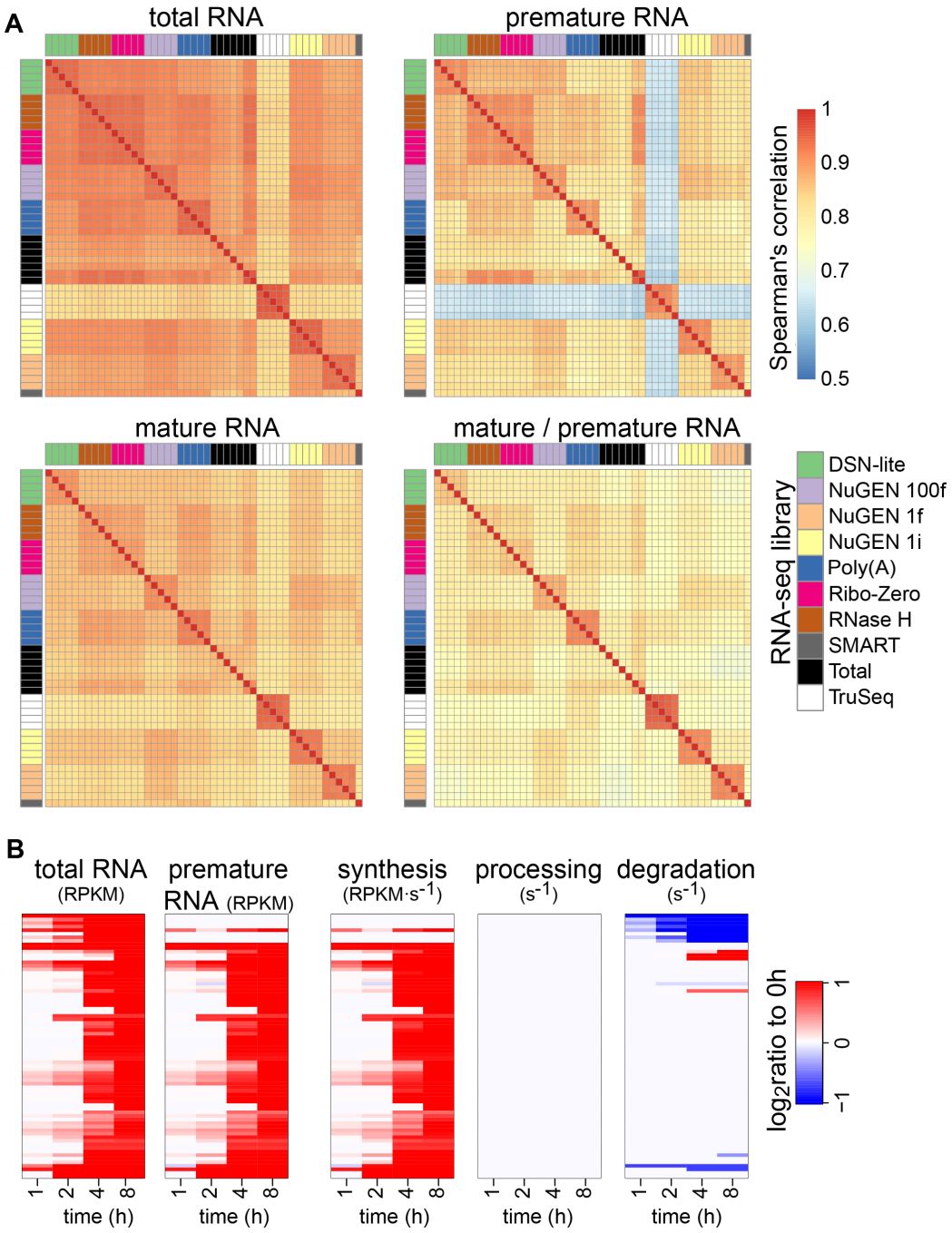
Supplemental Figure 13: **Disease and cell types annotations.** Heatmap displaying the degree of post-transcriptional regulation for each gene (row) in each sample (column). The heatmap combines the three heatmaps displayed in Main Figure 7D, including the legend of disease and cell type annotations.



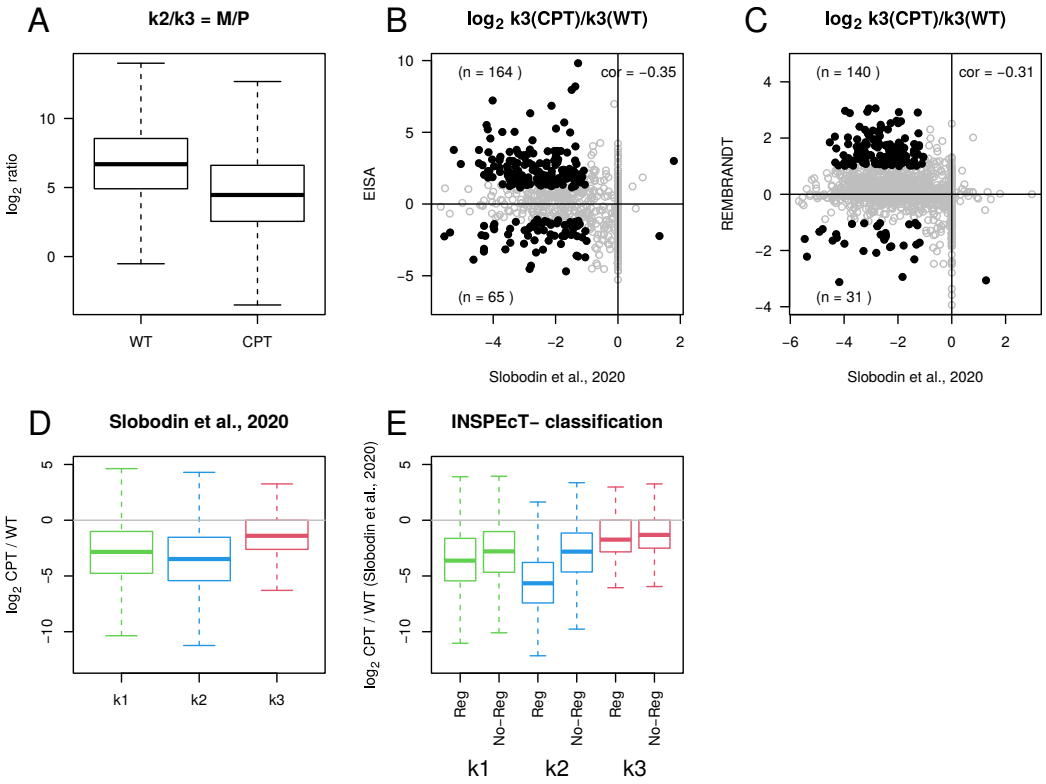
Supplemental Figure 14: **Impact of differential post-transcriptional regulation on the clustering of samples.** (A) Comparison between the hierarchical clustering of samples obtained with (left - INSPEcT-) and without (right - expression) the information about differential post-transcriptional regulations. In the latter, the underlying data matrix reduces to an object containing binary information: expression or lack of expression of each gene in each sample. Gray edges indicate the repositioning of samples in the two different dendrograms. The Cophenetic correlation (0.68) is determined to quantify the similarity between the two dendrograms, denoting ¿30% difference. (B) Cophenetic correlations between the dendrogram produced clustering the expression matrix and 100 trees (histogram) produced clustering matrices with randomly distributed regulated conditions. In red, we reported the score of INSPEcT- classification. (C) In red, we reported the Log_{10} Wilcoxon test p-value from the comparison of Cophenetic correlation coefficients obtained testing INSPEcT-'s dendrogram against the random ones, versus the distribution of pairwise comparisons of random trees. The boxplot shows the distribution of the same quantity estimated for each random configuration. (D) Score proportional to the homogeneity of samples' labels for random classification matrices (boxplots), INSPEcT- (red lines) and expression matrix (black lines). The order parameters (x-axes) indicates the number of neighbours involved in the score computation (see the insert for an explicative cartoon).



Supplemental Figure 15: **Validation of premature and mature RNA expression provided by INSPEcT.** (A) Histogram showing the fraction of intronic reads in whole-cell, nuclear, and cytoplasmic RNA. (B) Smooth-scatter plot, in the $Log_{10} - Log_{10}$ space, comparing nuclear to premature RNA (left panel), and mature to cytoplasmic RNA (right panel). Included in the scatter plot are Spearman's correlation ($r_{P,N}$ and $r_{M,C}$), and Spearman's partial correlation coefficients ($r_{P,N|C}$ and $r_{M,C|N}$), which removes the contribution of the correlation between nuclear and cytoplasmic RNA.



Supplemental Figure 16: **Impact of various RNA-seq library preparation methods on the quantification of RNA species abundance.** Spearman's correlation of gene-level total, premature and mature RNA abundance through alternative RNA-seq library preparations. (B) Temporal changes in total and mature RNA, and in the kinetic rates, following the induction of RAF profiled through RNA-seq of poly(A) transcripts.



Supplemental Figure 17: **Comparison between INSPEcT-, EISA and REMBRANDTS methods for the differential analysis of RNA metabolism from steady state RNA-seq data.** (A) The ratio between processing rate (k2) and degradation rate (k3) estimated by INSPEcT- in WT and CPT-treated MCF7 cells. (B) Dotplot of the differential degradation rates estimated by EISA in CPT-treated vs WT MCF7 cells compared to the ones measured by block of transcription in Slobodin et al., 2020. (C) Same as (B) but for REMBRANDTS. (D) Boxplot of the differential synthesis (k1), processing (k2) and degradation (k3) rates in CPT-treated vs WT MCF7 cells. (E) Boxplots of the differential k1, k2 and k3 rates in CPT-treated vs WT MCF7 cells, divided for genes identified as regulated (Reg) or not by INSPEcT-.

*