Supplemental information for 'Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms'

Supplemental note

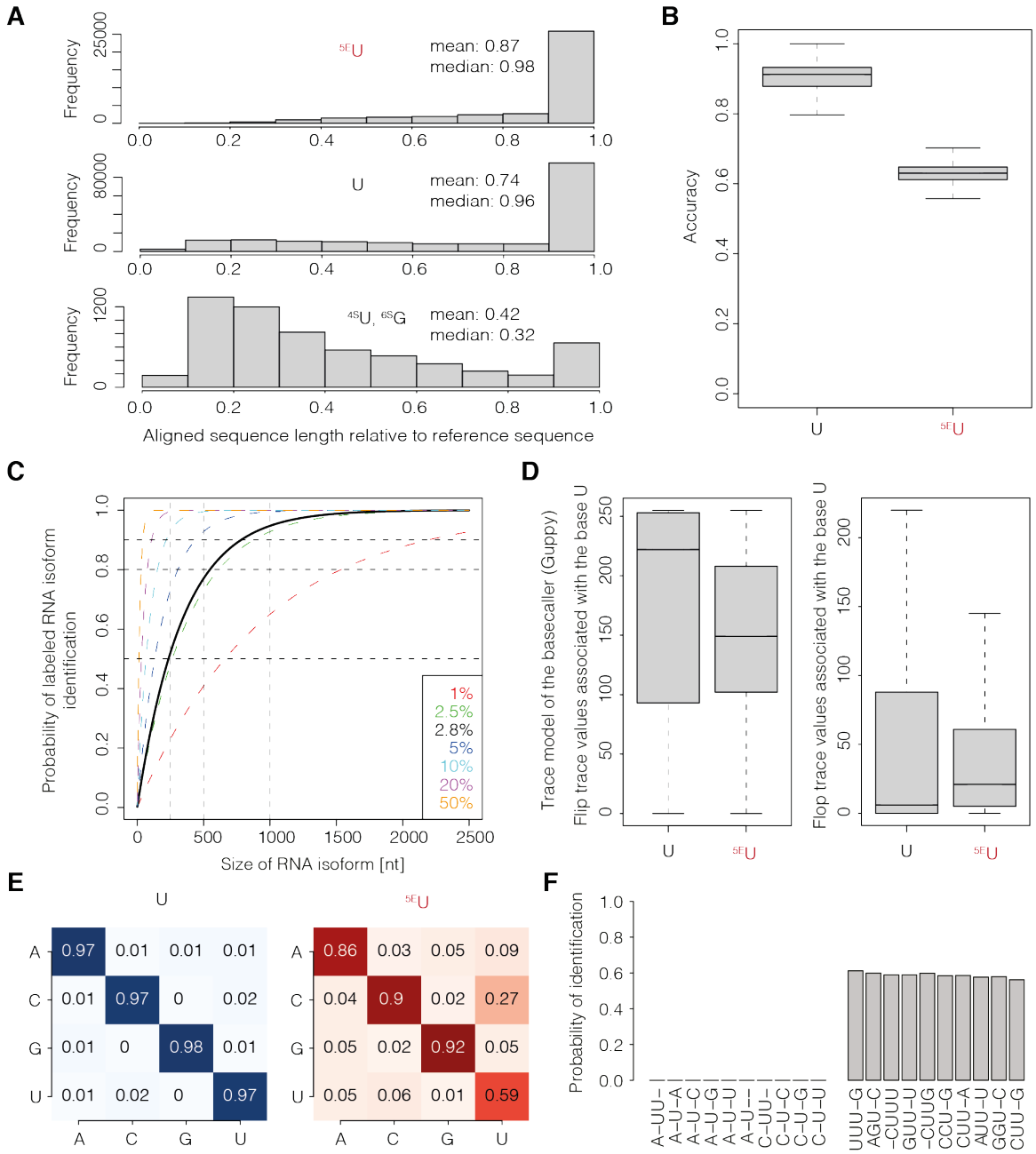
Supplemental Note 1. General and method-specific considerations.

5-Ethynyluridine (^{5E}U) uptake and incorporation efficiency might vary between cell types and organisms. Hence, ^{5E}U-labeling conditions have to be assessed for different model organisms and cell lines prior to metabolic rate (stability) estimation. ^{5E}U labeling duration can be adjusted to the mean RNA half-life of the respective model organisms or cell line to ensure a low combined variation across the entire range of RNA half-lives. In the present study, a labeling duration of 60 min was chosen as an optimal labeling time because this time point exhibits the least propagation of estimation error, measured as the coefficient of variation (see Supplemental Figure 3D). Specifically, the coefficient of variation was simulated for a mean RNA half-life of 80 min given reasonable negative binomially distributed values (mean and dispersion parameters derived from experimental count data, Samples 4-9, Supplemental Tables 1-2) for ^{5E}U-containing and non-^{5E}U-containing RNA molecules (used for half-life calculation) with a sample size of n=10,000 along different labeling durations (5-300 min). The simulation (Supplemental Figure 3D) shows that along the entire range of possible labeling durations, the interval of 60-120 min labeling is optimal for half-life calculation. Based on the simulation, a labeling duration of 60 min was selected in the present study to achieve a high resolution of new synthesis (time, [min]) upon external stimuli such as the heat shock response. Furthermore, the simulation highlights that estimation improvement through increased sequencing depth is superior to the measurement of two different time-points at lower depth.

For a half-life estimate to be considered reliable, the transcript had to be supported by at least 5 reads in each replicate. Given 6 biological replicates for the control (5EU 60 min) and 5 replicates for the heat shock response (5EU 60 min HS, **Supplemental Tables 1-2**, **Methods**), this leads to a minimum of 30 (5EU 60 min) or 25 (5EU 60 min HS) reads per loci on which the estimation is based. This threshold was applied with the exception of Figure 3, where we reduced the threshold from 5 to 2 to increase the number of observations.

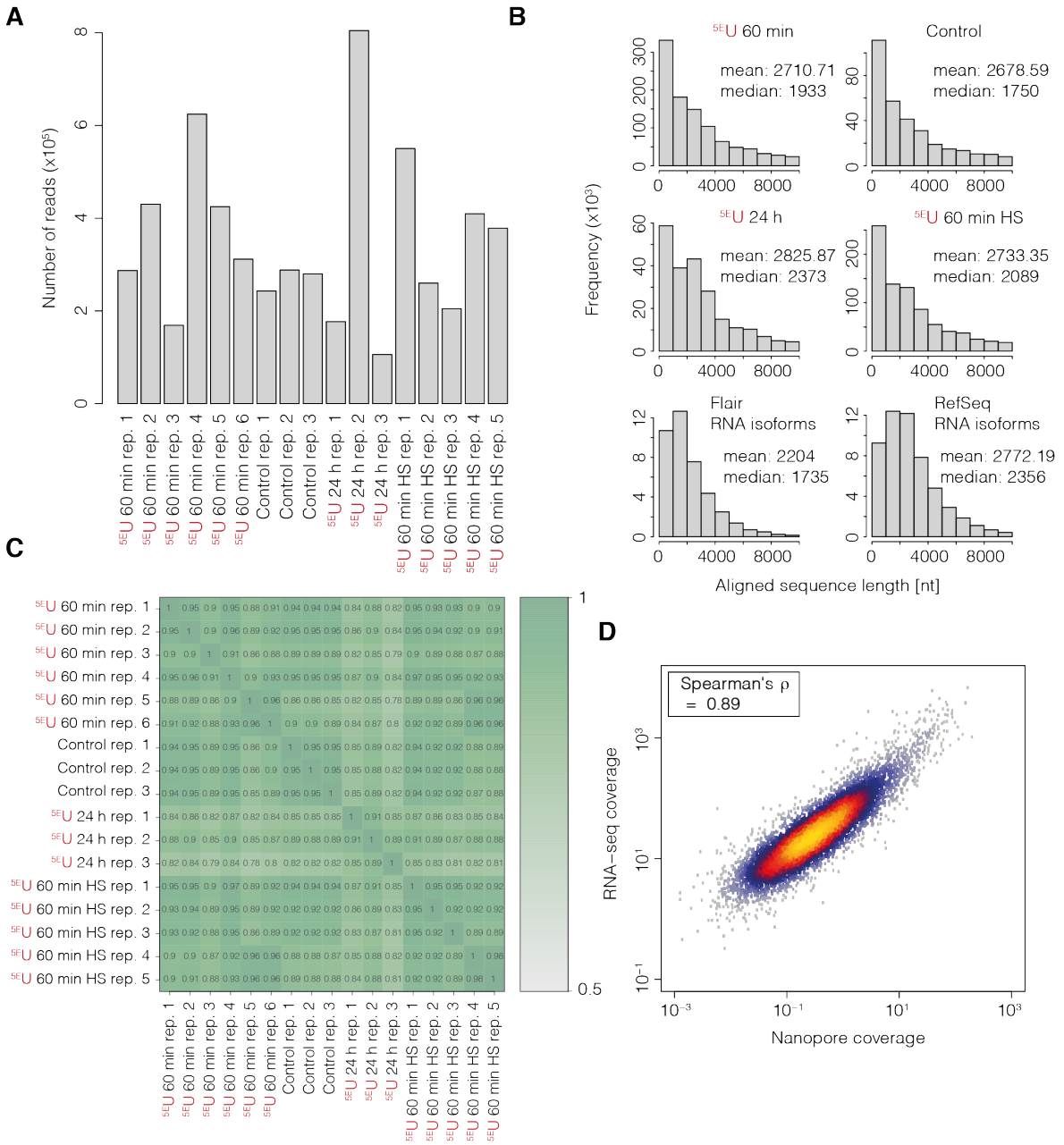
It is worth noting that ^{5E}U is incorporated *in vivo* with an efficiency (~2-3%) (Jao and Salic 2008) that is unlikely disruptive to the physiology of the cell at short (up to few hours) exposure times. ^{5E}U has been shown to be non-toxic to cells by propidium iodide/Annexin staining and it does not affect the global transcriptome of the cell (Invitrogen. Click-iT® Nascent RNA Capture Kit. http://www.invitrogen.jp/mp/pdf/clickit_nascent/mp10365.pdf).

Supplemental figures



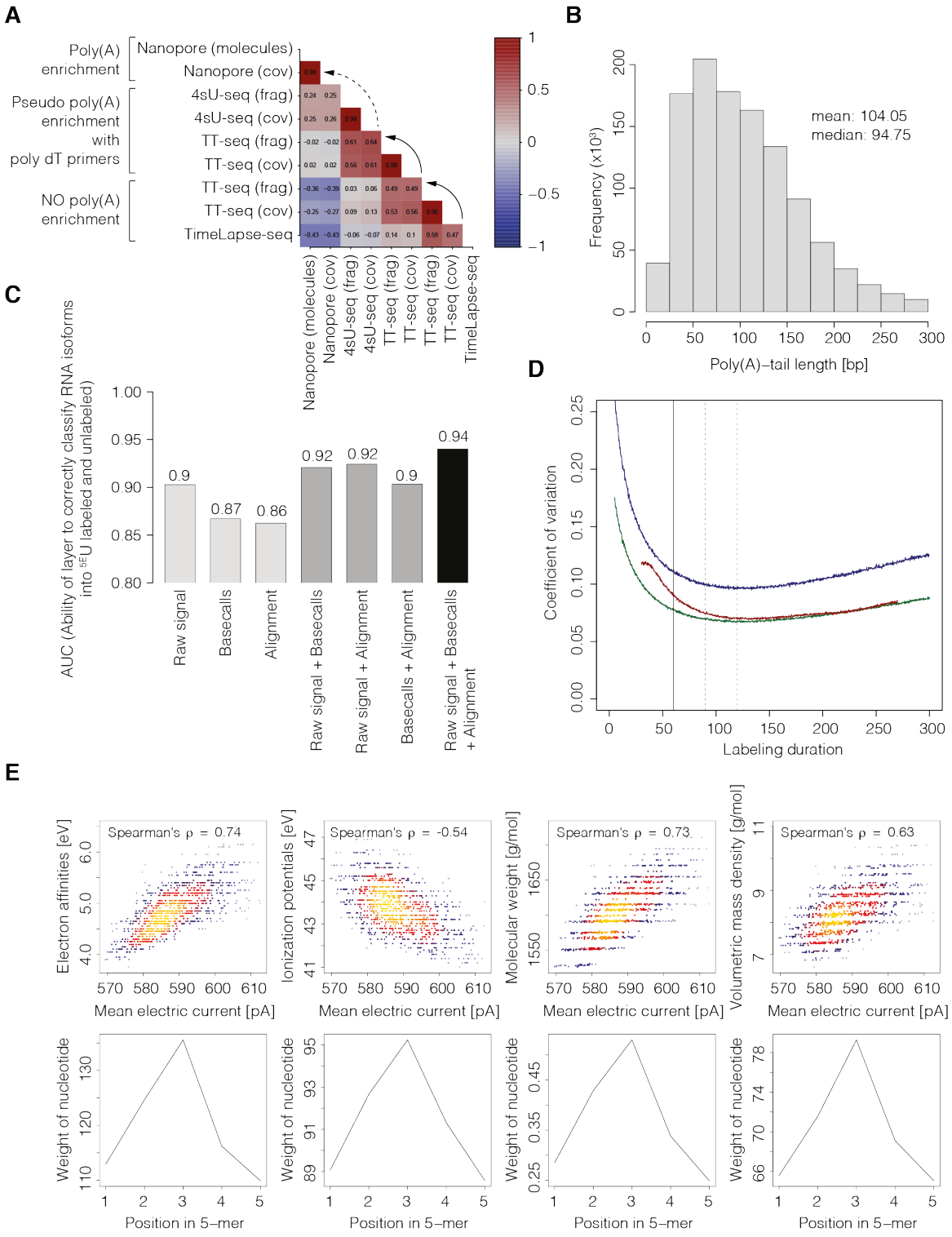
Supplemental Figure 1. Full-length and accuracy assessment of synthetic RNAs and probability of ^{5E}U-labeled RNA isoform identification. (A) Histograms show the relative length of aligned reads of synthetic RNAs (see Supplemental Table 3 for sequences) containing ^{5E}U instead of U (^{5E}U, 37,709 molecules), synthetic control RNAs (U, 179,096 molecules) and synthetic RNAs containing ^{4S}U or ^{6S}G instead of U or G (^{4S}U, ^{6S}G, 6,007 molecules) relative to their original sequence length. (B) Box plot shows the read based accuracy (edit distance)

of aligned synthetic RNAs containing ^{5E}U instead of U (^{5E}U, 37,709 molecules) and synthetic control RNAs (U, 179,096 molecules) based upon the reference sequence (Supplemental Table 3). (C) Plot shows the probability of ^{5E}U-labeled RNA isoform identification dependent on alignment length [nt] per RNA isoform and labeling efficiency (Methods). (D) Box plot shows trace values (flip-, flop-bases) derived from the base-caller of called U or ^{5E}U instances in synthetic RNAs containing ^{5E}U instead of U (^{5E}U, 37,709 molecules) and synthetic control RNAs (U, 95,345 molecules) associated with the base U in the reference sequence alignment. (E) Confusion matrices of read base-calls versus reference bases for synthetic control RNAs (U, 179,096 molecules) and synthetic RNAs containing ^{5E}U instead of U (^{5E}U, 37,709 molecules). (F) Bar plot shows the probability of ^{5E}U causing a mismatch in the aligned base-call (probability of identification) across the 10 least and most dominant U-containing 5-mers.



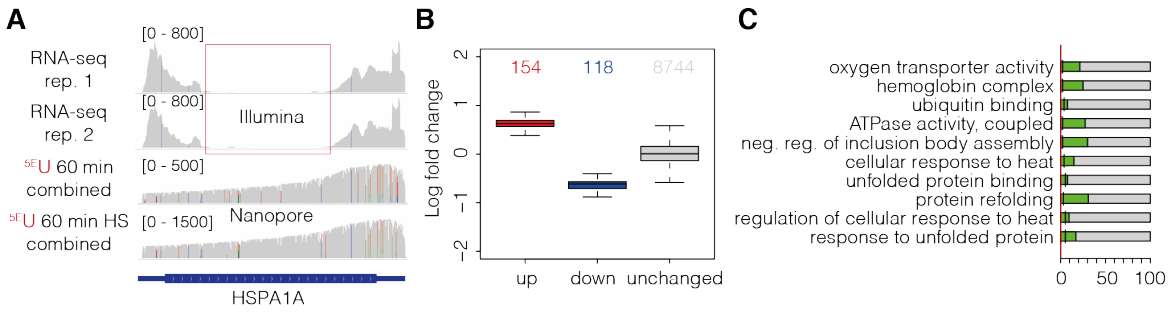
Supplemental Figure 2. Quality metrics and reproducibility assessment of direct RNA long-read nanopore sequencing of RNA isoforms in human K562 cells. (A) Bar plot shows the obtained number of reads for all samples. Biological replicates, n: ^{5E}U-labeling for 60 min (n=6), Control (n=3), ^{5E}U-labeling for 24 h (n=3), ^{5E}U-labeling for 60 min heat shock (HS) (n=5). (B) Histograms show the length distribution of aligned reads of human RNA isoforms. From top to bottom: shown are all replicates of ^{5E}U 60 min (1,304,273 molecules), 3 Control (419,645 molecules), ^{5E}U 24 h (262,872 molecules), or ^{5E}U 60 min HS (1,026,504 molecules) as well as RNA isoforms determined by the Flair algorithm (n=41,090) and RNA isoforms annotated in RefSeq GRCh38 (n=54,286). (C) Heatmap showing pairwise Spearman

correlation coefficients (green gradient of 0.5 to 1) of read coverage of 24,835 RefSeq GRCh38 annotated genes of all replicates across all samples. The green color gradient indicates Spearman correlation coefficients. Coefficients shown are rounded to the second decimal place. (D) Scatter plot with color-coded density of read coverage across biological replicates ^{5E}U 60 min (n=6) and Control (n=3), both read-out by Nanopore sequencing compared to read coverage of poly(A)-selected RNA-seq (n=2) read-out by Illumina sequencing (Methods). Shown are 6,458 RefSeq GRCh38 annotated genes with at least 2 reads across all direct RNA nanopore sequencing samples. Correlation is calculated as Spearman's rank correlation coefficient (0.89) rounded to the second decimal.



Supplemental Figure 3. Assessment of stability estimates, poly(A)-tail lengths, neural network layer combinations, error propagation and biophysical base properties on electric current readout. (A) Comparison of nano-ID derived stability estimates with those of previous methods. TT-seq and 4sU-seq (Pseudo poly(A) enrichment with poly dT primers) (Schwalb et al. 2016), TT-seq (NO poly(A) enrichment) (Gressel et al. 2019) and TimeLapse-

seq (NO poly(A) enrichment) (Schofield et al. 2018). Shown is the color-coded spearman correlation coefficient. (B) Histogram of poly(A)-tail length estimates of 1,304,273 RNA isoforms (poly(A) tail lenth, mean: 104 nt, median: 95 nt). (C) Bar plot showing the ability of individual layers (raw signal, base-calls and alignment) and layer combinations (raw signal + base-calls, raw signal + alignment, base-calls + alignment and raw signal + base-calls + alignment) to correctly classify RNA isoforms into ^{5E}U-containing (labeled) and non-^{5E}U-containing (unlabeled) measured as area under the curve (AUC). (D) Plot (blue curve) shows the coefficient of variation (standard deviation divided by the mean, y-axis) along different labeling durations (5-300 min, x-axis) of half-life estimates of an RNA species with a half-life of 80 min given typical estimation errors (simulation). The additional curves compare two measurement scenarios of either sequencing twice as deep (green curve, 'deeper-sequencing-scenario') or amending half-life calculation with an additional dataset of the same sequencing depth but with a labeling duration of 30 min longer than the original labeling duration (red curve, 'two-time-point scenario'). (E) Top panel: scatter plots with color-coded density of mean electric current of 5-mers in the nanopore against the calculated electron affinity (Russo et al. 2000), ionization potential (Roca-Sanjuan et al. 2006), molecular weight of ssRNA (A = 329.2 g/mol, U = 306.2 g/mol, C = 305.2 g/mol, G = 345.2 g/mol) and volumetric mass density (A = 1.6 g/mol, U = 1.32 g/mol, C = 1.55 g/mol, G = 2.2 g/mol) of the respective 5-mer (sum). Bottom panel: plots showing the weight of individual positions in the 5-mers determined via a simple linear regression.



Supplemental Figure 4. Nano-ID captures the response to heat shock in human K562 cells. (A) Genome browser view of the human HSPA1A gene locus (Chr6:31,813,514-31,819,942). Long-read nanopore sequencing (bottom panels) resolves mappability problems (red box) of short read sequencing such as Illumina-based RNA-seq (top panels). (B) Box plot shows upregulated (red), downregulated (blue) and unchanged RefSeq-annotated genes (grey) in human K562 cells upon 60 min of heat shock at 42 °C (HS). A minimum fold change of 1.25 and a maximum p-value of 0.1 was set for calling a significant expression change. (C) Gene Ontology (GO) analysis (Ashburner et al. 2000) of significantly overrepresented categories linked to upregulated RefSeq-annotated genes upon heat shock for human K562 cells. The red line depicts proportion of upregulated RefSeq-annotated genes in the whole population. The number of upregulated RefSeq-annotated genes in the resp. GO category is given relative to the GO category size (green bar).

Supplemental tables

Supplemental Table 1. Information on experimental conditions used in this study. Abbreviations used: ^{5E}U labeling for 60 minutes (^{5E}U 60 min) or 24 h (^{5E}U 24 h), heat shock (HS) and direct RNA long-read nanopore sequencing SQK-RNA001 (Nanopore) or SQK-RNA002 (Nanopore*). See Methods for experimental details. For HS treatments, temperature was monitored by thermometer and it took 5 min until the cell suspension reached 42 °C.

No.	Assay	Cell type	Condition name	Replicate no.	Treatment
1	Nanopore	-	Synthetic control	1	IVT
2	Nanopore	-	^{5Br}U, ^{4S}U, ^{6S}G	1	^{5Br}U, ^{4S}U, ^{6S}G, IVT
3	Nanopore	-	^{5E}U, ^{5I}U	1	^{5E}U, ^{5I}U, IVT
4	Nanopore	K562	^{5E}U 60 min	1	^{5E}U, 60 min
5	Nanopore	K562	^{5E}U 60 min	2	^{5E}U, 60 min
6	Nanopore	K562	^{5E}U 60 min	3	^{5E}U, 60 min
7	Nanopore	K562	^{5E}U 60 min	4	^{5E}U, 60 min
8	Nanopore*	K562	^{5E}U 60 min	5	^{5E}U, 60 min
9	Nanopore*	K562	^{5E}U 60 min	6	^{5E}U, 60 min
10	Nanopore	K562	Control	1	-
11	Nanopore	K562	Control	2	-
12	Nanopore	K562	Control	3	-
13	Nanopore	K562	^{5E}U 24 h	1	^{5E}U, 24 h (8 h intervals)
14	Nanopore	K562	^{5E}U 24 h	2	^{5E}U, 24 h (8 h intervals)
15	Nanopore	K562	^{5E}U 24 h	3	^{5E}U, 24 h (8 h intervals)
16	Nanopore	K562	^{5E}U 60 min HS	1	^{5E}U, 60 min; 42 °C, 65 min.
17	Nanopore	K562	^{5E}U 60 min HS	2	^{5E}U, 60 min; 42 °C, 65 min.
18	Nanopore	K562	^{5E}U 60 min HS	3	^{5E}U, 60 min; 42 °C, 65 min.
19	Nanopore*	K562	^{5E}U 60 min HS	4	^{5E}U, 60 min; 42 °C, 65 min.

20	Nanopore*	K562	^{5E}U 60 min HS	5	^{5E}U, 60 min; 42 °C, 65 min.
21	RNA-seq	K562	Ctrl	1	-
22	RNA-seq	K562	Ctrl	2	-

137

138

Supplemental Table 2. Sequencing statistics of 20 direct RNA long-read nanopore sequencing libraries generated in this study. All libraries were sequenced on a MinION Mk1B (MIN-101B) for 48 h or shorter if reads sequenced per second stagnated. Numbers refer to experimental conditions listed in Supplemental Table 1.

No.	Condition name	Read counts			Sequencing time
		Measured	Base-called	Aligned	
1	Synthetic control	800,475	800,475 (100%)	179,096 (22%)	20 h
2	^{5Br}U, ^{4S}U, ^{6S}G	42,131	42,073 (100%)	11,253 (27%)	4 h
3	^{5E}U, ^{5I}U	138,119	138,119 (100%)	50,688 (37%)	21 h
4	^{5E}U 60 min I	287,528	287,528 (100%)	120,627 (42%)	48 h
5	^{5E}U 60 min II	430,352	430,352 (100%)	167,767 (39%)	48 h
6	^{5E}U 60 min III	169,197	169,197 (100%)	38,412 (23%)	48 h
7	^{5E}U 60 min IV	624,550	624,550 (100%)	490,017 (78%)	48 h
8	^{5E}U 60 min V	425,037	425,037 (100%)	267,218 (63%)	48 h
9	^{5E}U 60 min VI	311,653	311,653 (100%)	220,246 (71%)	48 h
10	Control I	243,032	243,032 (100%)	131,525 (54%)	48 h
11	Control II	288,320	288,320 (100%)	153,993 (54%)	48 h
12	Control III	280,010	280,010 (100%)	134,127 (48%)	48 h
13	^{5E}U 24 h I	177,088	177,088 (100%)	37,821 (21%)	48 h
14	^{5E}U 24 h II	804,756	804,756 (100%)	205,522 (26%)	48 h
15	^{5E}U 24 h III	106,173	106,173 (100%)	19,529 (18%)	48 h
16	^{5E}U 60 min HS I	550,323	550,323 (100%)	251,401 (46%)	48 h
17	^{5E}U 60 min HS II	260,283	260,283 (100%)	115,148 (44%)	48 h
18	^{5E}U 60 min HS III	204,950	204,950 (100%)	94,819 (46%)	48 h
19	^{5E}U 60 min HS IV	409,767	409,767 (100%)	296,011 (72%)	48 h
20	^{5E}U 60 min HS V	378,622	378,622 (100%)	269,125 (71%)	48 h

Supplemental Table 3. List of synthetic RNAs used in this study. Synthetic RNAs are derived from selected synthetic sequences of the ERCC Spike-in Mix (Methods). Synthetic RNAs are polyadenylated after IVT.

No.	Name	Length (nt)	Derived from	%GC	%U
1	RNA spike-in 2	982	Synthetic, ERCC-00043	34	30
	5'- GGGTGCTTTAACAAGAGGAAATTGTGTTTTTGCCAATTTAAGACCTAATTTAATAGTTAAACCATTAACCTTAGT TGTTCCAAGGCATAATATAGAGAGTGAGATACAGGATGAGCTATTTCAGGGAGTTATTCAGTATGCAGTTGCCA AGGCAGTTGCTGATTTAGATTTAGATGAAGATTTAAAGGTTGTTGTCTCTGTTAATGTCCCAGAGGTTCCAATAA CCAATTTAAATAAAAGAAAACTCTTCCAATACTTCTATGCCTCAGCAAAGTTAGCTATAAACAGAGCTTTAAAT GAATATCCTTCAAAAGAGAAGGTAAAGAAAGAGAAATATAGAGCTTTGCATCCATTAGTTGGATTTAGGGATGT TAGATTGGAGTATCCTCCATATCTACAAATTGCTTTGGATGTCCCAACTATGGAGAATTTGGAATTTTTGTTACA AACAATTCCAAATAGCGACCACATCATCTTAGAGGCTGGAACACCACTAATTAAAAAGTTTGGTTTAGAGGTTA TTGAAATAATGAGAGAATATTTTGATGGCTTTATTGTTGCTGATTTAAAAACCTTAGACACTGGAAGGGTTGAG GTAAGATTGGCATTTGAAGCAACAGCTAATGCAGTGGCAATAAGTGGAGTAGCACCAAAATCAACAATAATTA AAGCTATCCACGAATGTCAAAAATGTGGTTTAATCAGCTATTTGGATATGATGAACGTCTCTGAACCTCAAAAA TTATATGATTCATTAAAATTAAAGCCAGATGTTGTTATCTTGCATAGAGGGATTGATGAGGAGACATTTGGAATT AAAAAGGAATGGAAATTTAAGGAAAACTGCTTATTAGCAATTGCTGGAGGAGTTGGTGTGGAGAATGTTGAAG AGCTTTTAAAAGAATATCAAATATTAATCGTTGGTAGAGCAATTACAAAATCAAAAGACCCAGGAAGAGTAATT AGGATTTTATAAACAAGATGG-3'				
2	RNA spike-in 4	1011	Synthetic, ERCC-00136	43	26
	5'- GGGTTTCGACGTTTTGAAGGAGGGTTTTAAGTAATGATCGAGATTGAAAAACCAAAAATCGAAACGGTTGAAAT CAGCGACGATGCCGAATTTGGTAAGTTTGTCGTAGAGCCACTTGAGCGTGGATATGGTACAACTCTGGGTAACT CCTTACGTCGTATCCTCTTATCCTCACTCCCTGGTGCCGCTGTAACATCAATCCAGATAGATGGTGTACTGCACG AATTCTCGACAATTGAAGGCGTTGTGGAAGATGTTACAACGATTATCTTACACATTAAAAAGCTTGCATTGAAA ATCTACTCTGATGAAGAGAAGACGCTAGAAATTGATGTACAGGGTGAAGGAACTGTAACGGCAGCTGATATTA CACACGATAGTGATGTAGAGATCTTAAATCCTGATCTTCATATCGCGACTCTTGGTGAGAATGCGAGTTTCCGAG TTCGCCTTACTGCTCAAAGAGGACGTGGGTATACGCCTGCTGACGCAAACAAGAGAGGCGATCAGCCAATCGGC GTGATTCCGATCGATTCTATCTATACGCCAGTTTCCCGTGTATCTTATCAGGTAGAGAACACTCGTGTAGGCCAA GTTGCAAACTATGATAAACTTACACTTGATGTTTGGACTGATGGAAGCACTGGACCGAAAGAAGCAATTGCGCT TGGTTCAAAGATTTTAACTGAACACCTTAATATATTCGCTGGTTTAACTGACGAAGCTCAACATGCTGAAATCAT GGTTGAAGAAGAAGAAGATCAAAAAGAGAAAGTTCTTGAAATGACAATTGAAGAATTGGATCTTTCTGTTCGTT CTTACAACTGCTTAAAGCGTGCGGGTATTAACACGGTTCAAGAGCTTGCGAACAAGACGGAAGAAGATATGAT GAAAGTTCGAAATCTAGGACGCAAATCACTTGAAGAAGTGAAAGCGAGACTAGAAGAACTTGGACTCGGACTT CGCAAAGACGATTGACTAGTTTCCCTTGTGAACTAGGATTTTCCCGGGTAC-3'				
3	RNA spike-in 5	1012	Synthetic, ERCC-00145	46	26
	5'- GGGACTGTCCTTTCATCCATAAGCGGAGAAAGAGGGAATGACATTGTTCTTACACGGCACAAGCAGACAAAATC AACATGGTCATTTAGAAATCGGAGGTGTGGATGCTCTCTATTTAGCGGAGAAATATGGTACACCTCTTTACGTAT ATGATGTGGCTTTAATACGTGAGCGTGCTAAAAGCTTTAAGCAGGCGTTTATTTCTGCAGGGCTGAAAGCACAG GTGGCATATGCGAGCAAAGCATTCTCATCAGTCGCAATGATTCAGCTCGCTGAGGAAGAGGGACTTTCTTTAGA TGTCGTATCCGGAGGAGAGCTATATACGGCTGTTGCAGCAGGCTTTCCGGCAGAACGCATCCACTTTCATGGAA ACAATAAGAGCAGGGAAGAACTGCGGATGGCGCTTGAGCACCGCATCGGCTGCATTGTGGTGGATAATTTCTAT GAAATCGCGCTTCTTGAAGACCTATGTAAAGAAACGGGTCACTCCATCGATGTTCTTCTTCGGATCACGCCCGG AGTAGAAGCGCATACGCATGACTACATTACAACGGGCCAGGAAGATTCAAAGTTTGGTTTCGATCTTCATAACG GACAAACTGAACGGGCCATTGAACAAGTATTACAATCGGAACACATTCAGCTGCTGGGTGTCCATTGCCATATC GGCTCGCAAATCTTTGATACGGCCGGTTTTGTGTTAGCAGCGGAAAAAATCTTCAAAAAACTAGACGAATGGAG AGATTCATATTCATTTGTATCCAAGGTGCTGAATCTTGGAGGAGGTTTCGGCATTCGTTATACGGAAGATGATGA ACCGCTTCATGCCACTGAATACGTTGAAAAAATTATCGAAGCTGTGAAAGAAAATGCTTCCCGTTACGGTTTTG ACATTCCGGAAATTTGGATCGAACCGGGCCGTTCTCTCGTGGGAGACGCAGGCACAACTCTTTATACGGTTGGC TCTCAAAAAGAAGTGGATAAGCTGTACAATCGTTTCATCATTCGGCGTGCG-3'				
4	RNA spike-in 8	1076	Synthetic, ERCC-00092	52	27

	5'- GGGGATGTCCTTGGACGGGGTGGCGCAGTATTACTGCAAGAGAGCGGACAGATTAGTGTGTTGGAGCCGACAC ATCAAAGGTTCGTCCGGGGACCGATCTGCAGCCTACGGGACATTTATCCGTAAAAGCATGGCGCTGTTTCGTAC TTATCGGAGGCCAGGTATCGTCGCGGCGAGTCTCCCCGACGACGGAGATGGGCGTTACTATCTGGGCCGTCTCG TACTCTGTTACTTGGCACAGATGCGAGCCCTCGTAATGTGCATCAGCTAAGGGCGATATTATAATGCGACGTTTG TACGGATTCGTTACTAACGTGTTGGACGCTAGTGGAATATGTGTCGTTGGTTAGCCTACCCATGGCTTTCGCGGC GACACATGCTTAGACTCTTTCAAAACTTCGGTGAAGTTCACTCAAGCCGCGGAGCGCCGTCGTAATTCACTAGG GATGGCGGTACCCGTGCCCGTCCGATTCGTAGCAACCTGCATCACGATTTTGTCTTCGGGCGACTTATCAGATAC GGTAATGTAAATACCTGGCATTTGGGCACTTCTTGCGTTTAAGCGGGAAAGATCGCGAGGGCCCGCTATTTGCG ATACTTCCCATGTCGGTGCCGTCGCCTCTATGTACTCGGAGACGTTAATGCAGAGGCTAAGGACAATTTACCATG ACTCGGTAATCCGTTCGTCAAGCAGGTAGCTCGAGTCTCCCCACGGACACGTAGTGGGTTTGTAACGATCGATA CCGAGTCTTTTTGTCTAGTAGAACCAACCAACCATTAAGGAGTTCACTAGCACATCTTTGCGACCCGATCGTCCG TGTGTCGCGTAATACTTTTGTTATGACGAGACATACGCTCAAGCCCTGGGTAGCTAGTCGCGGAGGCACGTTAC CGCGCACAACCCCTATTCGTTTACATGTACATCGCATCTGAGGTAGTACACTTCCGGCGTACGTGAGTATTTGCG CGTAATAAGCGCGTGTTTAGCTGATCCCCTCTCGTATCGAGGTTAAGGCAGATTAGTGCCCAGTAATTGCGTTTT TTTGTCGTTGTCGCAGAACGCGATTTGCTCCGAAAGC-3'				
5	RNA spike-in 9	1034	Synthetic, ERCC-00002	53	25
	5'- GGGCCAGATTACTTCCATTTCCGCCCAAGCTGCTCACAGTATACGGGCGTCGGCATCCAGACCGTCGGCTGATC GTGGTTTTACTAGGCTAGACTAGCGTACGAGCACTATGGTCAGTAATTCCTGGAGGAATAGGTACCAAGAAAAA AACGAACCTTTGGGTTCCAGAGCTGTACGGTCGCACTGAACTCGGATAGGTCTCAGAAAAACGAAATATAGGCT TACGGTAGGTCCGAATGGCACAAAGCTTGTTCCGTTAGCTGGCATAAGATTCCATGCCTAGATGTGATACACGT TTCTGGAAACTGCCTCGTCATGCGACTGTTCCCCGGGGTCAGGGCCGCTGGTATTTGCTGTAAAGAGGGGCGTT GAGTCCGTCCGACTTCACTGCCCCCTTTCAGCCTTTTGGGTCCTGTATCCCAATTCTCAGAGGTCCCGCCGTACG CTGAGGACCACCTGAAACGGGCATCGTCGCTCTTCGTTGTTCGTCGACTTCTAGTGTGGAGACGAATTGCCAGA ATTATTAACTGCGCAGTTAGGGCAGCGTCTGAGGAAGTTTGCTGCGGTTTCGCCTTGACCGCGGGAAGGAGACA TAACGATAGCGACTCTGTCTCAGGGGATCTGCATATGTTTGCAGCATACTTTAGGTGGGCCTTGGCTTCCTTCCG CAGTCAAAACCGCGCAATTATCCCCGTCCTGATTTACTGGACTCGCAACGTGGGTCCATCAGTTGTCCGTATACC AAGACGTCTAAGGGCGGTGTACACCCTTTTGAGCAATGATTGCACAACCTGCGATCACCTTATACAGAATTATC AATCAAGCTCCCCGAGGAGCGGACTTGTAAGGACCGCCGCTTTCGCTCGGGTCTGCGGGTTATAGCTTTTCAGTC TCGACGGGCTAGCACACATCTGGTTGACTAGGCGCATAGTCGCCATTCACAGATTTGCTCGGCAATCAGTACTG GTAGGCGTTAGACCCCGTGACTCGTGGCTGAACGGCCGTACAACTCGACAGCCGGTGCTTGCGTTTTACCC-3'				
6	RNA spike-in 12	947	Synthetic, ERCC-00170	35	31
	5'- GGGGCACAAGTTGCTGAAGTTGCGAGAGGGGCGATAAGTGAGGCAGACAGGCATAATATAAGAGGGGAGAGA ATTAGCGTAGATACTCTTCCAATAGTTGGTGAAGAAAATTTATATGAGGCTGTTAAAGCTGTAGCAACTCTTCCA CGAGTAGGAATTTTAGTTTTAGCTGGCTCTTTAATGGGAGGGAAGATAACTGAAGCAGTTAAAGAATTAAAGGA AAAGACTGGCATTCCCGTGATAAGCTTAAAGATGTTTGGCTCTGTTCCTAAGGTTGCTGATTTGGTTGTTGGAGA CCCATTGCAGGCAGGGGTTTTAGCTGTTATGGCTATTGCTGAAACAGCAAAATTTGATATAAATAAGGTTAAAG GTAGGGTGCTATAAAGATAATTTAATAATTTTTGATGAAACCGAAGCGTTAGCTTTGGGTTATGAAACTCCATG ATTTTCATTTAATTTTTTCCTATTAATTTTCTCCTAAAAAGTTTCTTTAACATAAATAAGGTTAAAGGGAGAGCTC TATGATTGTCTTCAAAAATACAAAGATTATTGATGTATATACTGGAGAGGTTGTTAAAGGAAATGTTGCAGTTG AGAGGGATAAAATATCCTTTGTGGATTTAAATGATGAAATTGATAAGATAATTGAAAAAATAAAGGAGGATGTT AAAGTTATTGACTTAAAAGGAAAATATTTATCTCCAACATTTATAGATGGGCATATACATATAGAATCTTCCCAT CTCATCCCATCAGAGTTTGAGAAATTTGTATTAAAAAGCGGAGTTAGCAAAGTAGTTATAGACCCGCATGAAAT AGCAAATATTGCTGGAAAAGAAGGAATTTTGTTTATGTTGAATGATGCCAAAATTTTAGATGTCTATGTTATGCT TCCTTCCTGTGTTCCAGCTACAAACTTAGAAACAAGTGGAGCTGAGATTACAGCAGA-3'				

147

148 **Supplemental Table 4. List of previously published datasets used in this study.**

No.	Data	Cell line	Used for	Reference	Available at
1	DMS-seq	K562	RNA secondary structure	Rouskin et al., 2014(Rouskin et al. 2014)	NCBI Gene Expression Omnibus (GSE45803)
2	Ribo-seq	K562	Translation	Ingolia et al. 2014(Ingolia et al. 2014)	NCBI Gene Expression Omnibus (GSE60095)
3	eCLIP	K562	RNA binding proteins (RBPs)	Nostrand et al. 2018(Van Nostrand et al. 2018)	ENCODE

149

150 Data accession links

151 1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45803

152 2: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60095

153 3: https://www.encodeproject.org

154

155 **Supplemental Table 5. Layer construction of the neural network.** Total number of
156 parameters: 2,601,913. Trainable parameters: 2,590,493. Non-trainable parameters: 11,420.
157 Output shape: (batch size, units).

Layer (type)	**Output Shape**	**Number of parameters**
Batch normalization layer 1	(None, 4,694)	18,776
Fully connected layer 1 (Dense)	(None, 512)	2,403,840
Batch normalization layer 2	(None, 512)	2,048
Dropout layer 1 (25%)	(None, 512)	0
Fully connected layer 2 (Dense)	(None, 256)	131,328
Batch normalization layer 3	(None, 256)	1,024
Dropout layer 2 (25%)	(None, 256)	0
Fully connected layer 3 (Dense)	(None, 128)	32,896
Batch normalization layer 4	(None, 128)	512
Dropout layer 3 (25%)	(None, 128)	0
Fully connected layer 4 (Dense)	(None, 64)	8,256
Batch normalization layer 5	(None, 64)	256
Dropout layer 4 (25%)	(None, 64)	0
Fully connected layer 5 (Dense)	(None, 32)	2,080
Batch normalization layer 6	(None, 32)	128
Dropout layer 5 (25%)	(None, 32)	0
Fully connected layer 6 (Dense)	(None, 16)	528
Batch normalization layer 7	(None, 16)	64
Dropout layer 6 (25%)	(None, 16)	0
Fully connected layer 7 (Dense)	(None, 8)	136

Batch normalization layer 8	(None, 8)	32
Dropout layer 7 (25%)	(None, 8)	0
Fully connected layer 8 (Dense)	(None, 1)	9

The training was conducted on 294,467 reads. Validation was performed on 126,130 reads in 40 epochs with the R interface to Keras on a TensorFlow backend (Allaire and Chollet 2018), as

$$train.model <- keras\_model\_sequential()$$

$$train.model \,\%>\%$$

$$layer\_batch\_normalization(input\_shape = c(ncol(train\_data)))\,\%>\%$$

$$layer\_dense(units = 512, activation = "relu", input\_shape = c(ncol(train\_data)))\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 256, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 128, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 64, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 32, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 16, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

$$layer\_dropout(rate = 0.25)\,\%>\%$$

$$layer\_dense(units = 8, activation = "relu")\,\%>\%$$

$$layer\_batch\_normalization()\,\%>\%$$

```
layer_dropout(rate = 0.25) %>%
layer_dense(units = 1, activation = "sigmoid")

train.model %>% compile(
optimizer = optimizer_rmsprop(),
loss = 'binary_crossentropy')
```

References

Allaire JJ, Chollet F. 2018. keras: R Interface to 'Keras'.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.

Gressel S, Schwalb B, Cramer P. 2019. The pause-initiation limit restricts transcription activation in human cells. *Nat Commun* **10**: 3603.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**: 1365-1379.

Jao CY, Salic A. 2008. Exploring RNA transcription and turnover in vivo by using click chemistry. *Proc Natl Acad Sci U S A* **105**: 15779-15784.

Roca-Sanjuan D, Rubio M, Merchan M, Serrano-Andres L. 2006. Ab initio determination of the ionization potentials of DNA and RNA nucleobases. *J Chem Phys* **125**: 084302.

Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701-705.

Russo N, Toscano M, Grand A. 2000. Theoretical determination of electron affinity and ionization potential of DNA and RNA bases. *Journal of Computational Chemistry* **21**: 1243-1250.

Schofield JA, Duffy EE, Kiefer L, Sullivan MC, Simon MD. 2018. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat Methods* **15**: 221-225.

Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science* **352**: 1225-1228.

218 Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen J-Y, Cody
219 NAL, Dominguez D et al. 2018. A Large-Scale Binding and Functional Map of Human RNA
220 Binding Proteins. *bioRxiv* doi:10.1101/179648: 179648.

221