Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms

Kerstin C. Maier[†1], Saskia Gressel[1], Patrick Cramer[1*], and Björn Schwalb[1†*]

[1]Max-Planck-Institute for Biophysical Chemistry, Department of Molecular Biology, Am Faßberg 11, 37077 Göttingen, Germany.

[†]These authors contributed equally.

[*]Correspondence to: Patrick Cramer, (patrick.cramer@mpibpc.mpg.de), and Björn Schwalb, (bjoern.schwalb@mpibpc.mpg.de).

Abstract

Eukaryotic genes often generate a variety of RNA isoforms that can lead to functionally distinct protein variants. The synthesis and stability of RNA isoforms is however poorly characterized. The reason for this is that current methods to quantify RNA metabolism use 'short-read' sequencing that cannot detect RNA isoforms. Here we present nanopore sequencing-based Isoform Dynamics (nano-ID), a method that detects newly synthesized RNA isoforms and monitors isoform metabolism. nano-ID combines metabolic RNA labeling, 'long-read' nanopore sequencing of native RNA molecules and machine learning. Application of nano-ID to the heat shock response in human cells reveals that many RNA isoforms change their synthesis rate, stability, and splicing pattern. nano-ID also shows that the metabolism of individual RNA isoforms differs strongly from that estimated for the combined RNA signal at a specific gene locus. And although combined RNA stability correlates with poly(A)-tail length, individual RNA isoforms can deviate significantly. nano-ID enables studies of RNA metabolism on the level of single RNA molecules and isoforms in different cell states and conditions.

25 **Main**

26 In metazoan cells, a single gene locus can give rise to a variety of different RNA molecules that
27 are generally referred to as isoforms. These RNA isoforms can differ in their 5'- and 3'-ends that
28 arise from the use of different transcription start sites and polyadenylation sites, respectively [1-4].
29 In addition, alternative splicing results in RNA isoforms that differ in the composition of their
30 RNA body [5,6]. Different mRNA isoforms can result in functionally different proteins.
31 Vulnerabilities in splicing can lead to non-functional protein products. Diseases have been linked
32 to alternative splicing, which can generate malignant RNA isoforms [7]. Duchenne muscular
33 dystrophy (DMD), for example, can be pinpointed to a single gene encoding the protein
34 dystrophin. The underlying malignant RNA isoform exhibits a different splicing pattern and
35 leads to a non-functional protein, which disrupts muscular cell integrity [8]. Likewise, the three
36 most common types of breast tumors are linked to exon skipping and intron retention [9].

37 RNA isoforms can also differ in their stability. The untranslated region of an RNA
38 isoform can differ in length and contains regulatory elements [10]. The length of the poly(A)-tail at
39 the 3'-end of RNA isoforms can also differ and influence RNA stability [11,12], and this is relevant
40 for disease as well [13]. Finally, introns may be retained in RNAs and can influence stability [14].

41 Little is known however about the synthesis and stability of single RNA isoforms in cells.
42 This is because the systematic characterization of RNA isoforms and their metabolism is
43 technically difficult. In particular, the detection, quantification and estimation of the stability of
44 RNA isoforms is essentially impossible with 'short-read' RNA sequencing methods because
45 reads generally cannot be assigned to RNA isoforms. Also, alternative splicing patterns can be
46 manifold and are difficult to identify using 'short-read' sequencing approaches [15]. Finally,
47 although the length of poly(A)-tails of RNAs can be measured genome-wide [16,17], they can
48 currently not be obtained at the level of individual RNA isoforms.

49 The architecture of RNA isoforms has been addressed so far by 'short-read' RNA
50 sequencing approaches such as DARTS [18], VastDB [19] and MPE-seq [20] to study alternative
51 splicing or TIF-seq [1,3] to elucidate combinations of paired 5'- and 3'-ends of individual RNAs.
52 More recent approaches include 'long-read' sequencing approaches on the PacBio SMRT
53 Sequencing platform [6] or Oxford Nanopore Technologies nanopore sequencing platform [5,21,22].

54 These methods however are not able to study the metabolism of individual RNA isoforms
55 because they lack the ability to assign age to single reads.

56 Methods to measure the synthesis and stability of combined RNA for entire gene loci are
57 available [23-25]. Transient transcriptome sequencing (TT-seq) is a protocol that allows to
58 distinguish newly synthesized from pre-existing RNA in human cells [26]. TT-seq involves a brief
59 exposure of cells to the nucleoside analogue 4-thiouridine (4sU). 4sU is incorporated into RNA
60 during transcription, and the resulting 4sU-labeled RNA can be purified and sequenced to
61 provide a snapshot of immediate transcription activity. This then enables to computationally
62 infer RNA synthesis and stability at the level of the combined RNA signal from a gene locus.

63 Recent methods to assess RNA stability include SLAM-seq [27] and TimeLapse-seq [28].
64 Like TT-seq, SLAM-seq and TimeLapse-seq involve an exposure of cells to 4sU for labeling of
65 newly synthesized RNA. A chemical modification of the incorporated 4sU then allows for the
66 identification of labeled RNA *in silico* without the need for purification. All of these methods,
67 however, have limitations. First, sequencing reads can normally only be assigned to entire gene
68 loci and not to RNA isoforms and thus only allow a combined RNA stability assessment.
69 Second, they require template amplification, which can lead to an imbalance in measured
70 sequences and information loss, e.g. modified RNA bases [29]. Third, labeled RNA purification
71 (TT-seq) and cDNA library preparation (TT-seq, SLAM-seq & TimeLapse-seq) can also
72 introduce biases.

73 Therefore, monitoring RNA metabolism at the level of RNA isoforms requires a method
74 that can detect individual RNA molecules. Recent advances in 'long-read' nanopore sequencing
75 indeed enable the sequencing of single, full-length RNA molecules [5]. Nanopore technology can
76 directly sequence the original native RNA molecule with its modifications, may they be natural
77 or acquired by metabolic RNA labeling. Moreover, the availability of the entire RNA and coding
78 sequence (CDS) within a single read allows to unambiguously and directly determine exon usage
79 [30]. Direct RNA 'long-read' nanopore sequencing also has the potential to detect the position and
80 length of the poly(A)-tail along with each single isoform.

81 Here we developed nanopore sequencing-based Isoform Dynamics (nano-ID), which
82 combines metabolic RNA labeling with native RNA 'long-read' nanopore sequencing for RNA
83 isoform detection. In combination with computational modeling and machine learning this

84 allows for a full characterization of RNA isoforms dynamics. nano-ID can identify and quantify
85 RNA isoforms along with their synthesis rate, stability and poly(A)-tail length in the human
86 myelogenous leukemia cell line K562. We show that this is possible with nano-ID in a
87 quantitative manner in steady state and also during the transcriptional response to heat shock.
88 nano-ID is able to resolve the dynamic metabolism of RNA isoforms upon heat shock and
89 demonstrates the need for individual RNA isoform assessment. Taken together, nano-ID can be
90 used to elucidate a largely unexplored complex layer of gene regulation at the level of single
91 native RNA isoforms and their metabolism.

92

93 **Results**

94 **Experimental design**

95 To monitor the metabolism of RNAs at the level of single isoforms, we sought to combine
96 metabolic RNA labeling with direct, single-molecule RNA nanopore sequencing (**Figure 1a**). By
97 culturing cells in the presence of a nucleoside analogue, cells will take up and incorporate the
98 analogue in nascent RNA during transcription, allowing to distinguish newly synthesized RNA
99 isoforms from pre-existing RNA isoforms *in silico* based on the quantification of analogue-
100 containing subpopulations. This will allow to infer the synthesis rate and stability of single RNA
101 isoforms. In order to dynamically characterize functional and fully processed RNA transcripts,
102 we decided to measure poly-adenylated RNA species. The library preparation kit offered by
103 Oxford Nanopore Technologies for direct RNA sequencing (SQK-RNA001) is specifically
104 optimized for this purpose. A 3' poly(A)-tail specific adapter is ligated to the transcript in a first
105 step. Then a second sequencing adapter equipped with a motor protein is ligated to the first
106 adapter. The preparation of RNA libraries from biological samples for direct RNA nanopore
107 sequencing is established and can be carried out within 2h [31]. Major challenges that we faced
108 were however the search of a suited nucleoside analogue for RNA labeling and the detection of
109 labeled RNA isoforms, provided that the labeling efficiency is known to be limited to about 2-
110 3%, i.e. only two or three out of 100 natural nucleotides are replaced by the analogue [32].

111

112 **5-Ethynyluridine (^{5E}U) can be detected in RNA by nanopore sequencing**

113 To investigate if nucleoside analogues incorporated into RNA are detectable in the nanopore, we
114 used synthetic RNAs derived from the ERCC RNA spike-in mix (Life Technologies). These
115 synthetic RNAs of an approximate length of 1,000 nucleotides were chosen with similar U
116 content (**Supplementary Table 3**). RNAs were transcribed *in vitro* using either the standard
117 bases A, U, C, G as a control, or with one of the natural bases exchanged for a nucleoside
118 analogue (**Figure 1b**, **Methods**). Subsequently, we subjected these synthetic RNAs to direct
119 RNA nanopore sequencing (**Supplementary Figure 1a-b**). We compared the nucleoside
120 analogues 5-Ethynyluridine (^{5E}U), 5-bromouridine (^{5Br}U), 5-iodouridine (^{5I}U), 4-thiouridine (^{4s}U)
121 and 6-thioguanine (^{6s}G). To this end we used the base-called and mapped direct RNA sequencing
122 results to calculate how probable the identification would be on the level of single nucleotides. In
123 particular, we compared the error rate in single nucleotide base-calls of nucleoside analogues to
124 that of natural U or G (**Figure 1c, Methods**).

125 The thiol-based analogues, ^{4s}U and ^{6s}G, showed lower incorporation efficiencies during
126 *in vitro* transcription (IVT) and led to blockages during nanopore sequencing. ^{5E}U and ^{5I}U could
127 be detected to a similar extent by nanopore sequencing, whereas ^{5Br}U was less easily recognized
128 (**Figure 1c**). Since ^{5E}U is not toxic to cells [32-34], we used ^{5E}U for a more detailed analysis.
129 Approximately 50% of all U positions in ^{5E}U-containing synthetic RNAs are consistently
130 miscalled by the standard base-calling algorithm and can thus be discerned from U (**Figure 1d,**
131 **Supplementary Figure 1b**). This is clearly visible in the raw data. Aberrations caused by
132 stretches of RNA containing ^{5E}U are distinguishable from stretches of RNA containing the
133 naturally occurring U in the nanopore (**Figure 1d**). Taken together, ^{5E}U-based RNA labeling is
134 well suited for nanopore sequencing.

135

136 **Detection and sequencing of newly synthesized RNA isoforms**

137 We next investigated whether it is possible to use metabolic RNA labeling with ^{5E}U in human
138 cells to detect single RNA molecules by nanopore sequencing. Calculations on the direct RNA
139 nanopore sequencing results of the ^{5E}U-containing synthetic RNAs showed that RNAs are
140 recognizable as ^{5E}U containing with a probability of 0.9 when a minimum length of 500
141 nucleotides is reached (**Supplementary Figure 1c-d**). This covers the vast majority (93%) of all
142 mature RNAs in the human organism (UCSC RefSeq GRCh38).

143 We then established direct RNA nanopore sequencing in the human myelogenous
144 leukemia cell line K562. We cultured K562 cells in the presence of ^{5E}U for 60 minutes (^{5E}U 60
145 min) in 4 biological replicates (**Methods**). For comparison, we created 3 biological replicates
146 exposed to ^{5E}U labeling for 24 h (^{5E}U 24 h) and 3 biological replicates that were not labeled
147 (Control). After standard base-calling, we could map reads to support 13,110 RefSeq annotated
148 transcription units (RefSeq-TUs, **Methods**), 8,098 of these were supported in all conditions and
149 1,726 were supported in all samples.

150 All combined samples were then used to perform a full-length alternative RNA isoform
151 analysis by means of the FLAIR algorithm [22]. This allows defining instances of unique exon-
152 intron architecture with unique start and end sites in human K562 cells. Raw human direct RNA
153 nanopore reads were corrected with the use of short-read sequencing data (RNA-seq) to increase
154 splice site accuracy. We could detect 33,199 distinct RNA isoforms with an average of 3
155 isoforms per gene. This shows that direct RNA nanopore sequencing uncovers individual RNA
156 isoforms in human K562 cells (**Figure 2**) with high reproducibility (**Supplementary Figure 2**).

157

158 **A neural network identifies newly synthesized RNA isoforms**

159 The next step was to derive a computational method that could classify each sequenced RNA
160 molecule into one of two groups, newly synthesized (^{5E}U-labeled) or pre-existing (unlabeled)
161 RNA. To this end, the nucleoside analogue ^{5E}U had to be detected in RNA molecules. This
162 would allow the quantification of RNA isoforms generated during the ^{5E}U labeling pulse. Due to
163 the high error rate of nanopore sequencing, a single ^{5E}U base-call is inappropriate as an indicator.
164 We rather used the raw signal of the entire RNA nanopore read, including the base-calls and the
165 alignment, to discriminate labeled from unlabeled RNAs. This discrimination was implemented
166 as a classifying neural network. We developed a custom multi-layered data collection scheme to
167 train a neural network for the classification of human RNA isoforms under the assumption that
168 the ^{5E}U 24 h samples solely contain labeled reads and the fact that the Control samples solely
169 contain unlabeled reads (**Figure 3a, Methods**).

170 We then trained a neural network (**Methods**) on the ^{5E}U 24 h versus Control samples
171 with an accuracy of 0.87 and a false discovery rate (FDR) of 0.025 (5-fold cross-validated). A
172 ROC analysis (1 – specificity versus sensitivity) for all reads of the test set showed an area under

173 the curve (AUC) of 0.94. For reads with an alignment length larger than 500 nt and 1,000 nt the
174 AUC improved to 0.96 (**Figure 3b, Supplementary Figure 3a, b**). Subsequently we used the
175 trained neural network to classify reads of the ^{5E}U 60 min samples into ^{5E}U-labeled and
176 unlabeled. Taken together, ^{5E}U containing RNA isoforms are computationally detectable with
177 high accuracy (**Figure 3c**). For validation purposes, we used another machine learning approach.
178 We trained a random forest on the same data, which yielded similar results (**Supplementary**
179 **Figure 3c, d**). Thus, we were able to determine for each single RNA molecule if it has been
180 produced during ^{5E}U labeling or before, with a low false discovery rate (**Figure 3c**).

181

182 **nano-ID provides the stability and poly(A) tail length of RNA isoforms**

183 The ability to distinguish newly synthesized and pre-existing RNA molecules allowed us to
184 derive estimates for the stability of RNA isoforms. For each single direct RNA nanopore read we
185 were able to assign the RNA isoform it reflects. Additionally, we were able to assess the stability
186 of RNA for single RNA isoforms by applying a first-order kinetic model (**Methods,**
187 **Supplementary Figure 3e-f**) to derive estimates for RNA isoform specific synthesis and
188 stability. This can be done based on the number of reads classified as ^{5E}U-labeled and unlabeled
189 by the neural network. Taken together, nano-ID has the capability to infer synthesis and stability
190 of individual RNA isoforms in different cell states and conditions, and thus to monitor their
191 dynamic metabolism.

192 Moreover, we developed an algorithm to determine poly(A)-tail lengths for each RNA
193 isoform (**Figure 4**). This is possible by estimating the dwell time of the poly(A)-tail in the
194 nanopore by factoring in the measurement frequency in kHz and the speed of RNA translocation
195 through the nanopore (**Methods**). Sequencing adaptor ligation in the direct RNA nanopore
196 sequencing library preparation guarantees full-length poly(A)-tails because ligation of the
197 adapter would not be successful otherwise. The resulting poly(A)-tail length distribution is in
198 line with the current literature [16] (**Figure 4a**) and reveals a pattern that likely corresponds to the
199 26 nucleotide footprint of the poly(A) binding protein (**Supplementary Figure 4a**) [35]. The direct
200 RNA nanopore sequencing kit contains the so-called RNA calibration strand (RCS). The RCS is
201 a synthetic RNA with a poly(A)-tail of exactly 30 adenines. Using the RCS of the direct RNA
202 nanopore sequencing kit, we could assess the accuracy of the poly(A)-tail length estimates

203 (coefficient of variation 0.63). Our algorithm derives this length for the added RCS
204 subpopulation (**Figure 4b**). Taken together, nano-ID reveals the synthesis, stability, and poly(A)
205 tail length for individual RNA isoforms in human cells.

206

207 **nano-ID monitors RNA isoform dynamics during heat shock**

208 To demonstrate the advantages of nano-ID, we subjected human K562 cells to heat shock (42
209 °C) for 60 min in the presence of ^{5E}U (^{5E}U 60 min HS) (**Figure 5a**). The heat shock response
210 provides a well-established model system [36-41] (**Supplementary Figure 5**). We first asked
211 whether RNA isoforms do retain more introns after heat shock as this was shown in the mouse
212 system [42]. Indeed, we observed widespread intron retention which significantly increased upon
213 heat shock (**Figure 5b**). Although intron retention generally influences the stability of an RNA, it
214 does not explain changes in RNA isoform stability upon heat shock (**Figure 5c**). This finding is
215 consistent with the idea that specific RNA elements occurring only in specific RNA isoforms
216 influence RNA stability.

217 We next asked if RNA isoform synthesis is altered by heat shock and observed
218 significant differential RNA isoform synthesis for 285 isoforms (fold change > 1.25 and p-value
219 < 0.1). 187 RNA isoforms were significantly upregulated, while 98 were downregulated (**Figure**
220 **5d**). RNA isoforms that changed their synthesis during heat shock were also observed to alter
221 their stability (**Figure 5e-f**). In particular, RNA isoforms that were upregulated in their synthesis
222 during heat shock also showed a lower stability, and the other way around, resembling typical
223 stress response behavior [24]. The destabilization of upregulated RNA isoforms is likely to ensure
224 their rapid removal toward the end of the stress response. Similarly, downregulated RNA
225 isoforms are stabilized, perhaps to preserve them for translation at later stages.

226

227 **nano-ID reveals the biogenesis of RNA isoforms**

228 Although standard native RNA isoform sequencing can reveal isoforms present in a sample after
229 perturbation, it cannot distinguish whether these isoforms were derived by synthesis, stability,
230 splicing, or any combination of these. nano-ID however is able to disentangle these parameters.
231 For example, although we observe a general increase in intron retention upon heat shock, we find

232 exceptions at the level of RNA isoforms. This can be clearly seen at the human C1orf63 gene
233 locus (**Supplementary Figure 5g**). Here, the majority of reads, that retain the entire 3rd intron,
234 were newly synthesized in the control samples. It is however unclear if this intron will be
235 retained throughout the existence of these RNA molecules. Investigation of the same gene locus
236 upon heat shock showed that the vast majority of reads were pre-existing RNAs. This indicates
237 that this RNA is not transcribed anymore upon heat shock and allows for the conclusion that
238 intron retention is not altered, rather, less introns are seen retained when only old RNA is
239 detected. Taken together, this shows that nano-ID is able to resolve the dynamic behavior of
240 RNA isoforms upon stimuli that could not be seen otherwise. It demonstrates the need for
241 individual RNA isoform detection and classification into newly synthesized and pre-existing
242 molecules. By providing information on the age of RNA molecules, nano-ID enables an analysis
243 of the biogenesis of RNA isoforms.

244

245 **The metabolism of individual RNA isoforms differs from combined RNA estimates**

246 To demonstrate the importance of individual RNA isoform assessment, we first derived
247 estimates for the half-lives of combined RNAs that stem from entire gene loci under steady state
248 conditions (**Methods, Supplementary Figure 3e-f**). We found a robust correlation of combined
249 RNA stability with poly(A)-tail length (Spearman's rank correlation coefficient 0.48) (**Figure**
250 **5g**). We now asked whether changes in RNA stability would also be reflected in changes in
251 poly(A)-tail length upon heat shock, and this was not the case (**Figure 5h**). Instead, we found
252 genes that showed the opposite behavior to the overall correlation as demonstrated for the human
253 HSPB1 locus (**Figure 6a-b**). Here, destabilization of combined RNAs is accompanied by
254 lengthening of the poly(A)-tail. This view changes dramatically when considering individual
255 RNA isoforms (**Figure 6c**). For those three RNA isoforms at the human HSPB1 gene locus for
256 which stability estimates were supported by all 3 biological replicates (**Methods**) we found that
257 poly(A)-tails were generally longer. RNA stability however was decreased for 2 out of the 3
258 RNA isoforms and increased for the third. This clearly indicates the need for detailed individual
259 RNA isoform assessment as individual RNA isoforms can lead to functionally distinct protein
260 variants. Thus, it is crucial to also study the behavior of individual RNA isoforms instead of
261 breaking it down to the combined view of the entire gene locus.

262 As a second example, we picked RNA isoforms at the human TAGLN2 gene locus
263 (**Figure 6d**). We could identify 7 different RNA isoforms and reliably calculate RNA stability
264 for 6 RNA isoforms. Two of them were stabilized upon heat shock, 4 of them were destabilized.
265 All 4 destabilized RNA isoforms include the second to last exon, which might cause this change
266 in stability. RNA isoform 7 is an exception to this observation as it is stabilized upon heat shock.
267 It, however, also contains a 3' UTR that is 42 bases shorter than all the other RNA isoforms. We
268 asked whether there is differential behavior of individual RNA isoforms genome-wide or if RNA
269 isoforms generally reflect the changes in stability of the combined RNA from their respective
270 gene loci. To that end, we compared RNA stability estimates of individual RNA isoforms to
271 those from combined RNAs and found that the dynamics of individual RNA isoform during heat
272 shock varies globally (**Figure 6e**, **Supplementary Figure 6**). Taken together, this shows that
273 conclusions can be misleading when combined RNAs are used and how much can be learned on
274 the level of single RNA isoforms by using nano-ID.

275

276 **Discussion**

277 Here we develop nano-ID, a method that allows for dynamic characterization of functional and
278 fully processed RNA isoforms on the level of single native RNA molecules. nano-ID combines
279 metabolic RNA labeling with native RNA nanopore sequencing to enable RNA isoform
280 identification, estimation of its stability, and a measurement of its poly(A)-tail length from a
281 single sample. nano-ID is able to visualize changes in RNA isoform synthesis and stability and
282 reveals a hidden layer of gene regulation. nano-ID thus allows to study transcriptional regulation
283 in unprecedented detail and can prevent misleading conclusions that would be obtained when
284 only combined RNAs from an entire gene locus are considered, as is done by RNA-seq, 4sU-seq
285 or TT-seq.

286 nano-ID has many advantages over other sequencing-based transcriptomic strategies as it
287 allows to sequence the original native RNA molecule. In particular, there is no need for
288 fragmentation of RNA prior to sequencing and hence no ambiguity in assigning reads to RNA
289 isoforms. nano-ID also does not require template amplification and thus omits copying errors and
290 sequence-dependent biases. It comes without a lengthy library protocol and eliminates
291 sequencing by synthesis and therefore prevents loss of information on epigenetic modifications
292 and artificially introduced RNA base analogues. It is PCR-free and shows neither sequence bias

293 nor read duplication events. Taken together, it overcomes drawbacks and limitations of state-of-
294 the-art approaches and increases the gathered information vastly.
295 Generally, nanopore sequencing has still limitations in throughput and accuracy. These
296 drawbacks, however, are outweighed by the information obtained on the sequencing substrates.
297 The longer the sequenced molecules are, the less problematic is the lack in accuracy in
298 identifying their origin or classifying it into newly-synthesized or pre-existing. On top of that,
299 there are strategies to improve splice site calling with already existing high accuracy 'short-read'
300 sequencing data to reduce sequencing errors or to assess the likelihood of real nucleotide
301 variants. We can however show that our algorithmic strategies are already sufficient to address
302 metabolic rate estimation in a reliable manner. Technical improvements in nanopore sequencing
303 or their computational processing will strongly improve the accuracy of individual read
304 sequences and thus detectability of ^{5E}U. The task at hand will be the development of a novel
305 base-calling algorithm for direct RNA nanopore sequencing with extended base alphabet (A, C,
306 G, U & ^{5E}U). Furthermore, increased throughput will foster statistical precision of metabolic rate
307 estimation and will also allow to elucidate low abundant or transient processes.
308 Nanopore-based transcriptomic studies will allow us to monitor the formation of
309 transcripts, post-transcriptional processing, export and translation at the level of single RNA
310 isoforms. nano-ID is in principle also transferable to single cell methodologies, to catch
311 heterogeneity of the RNA population in any state of the cell. This however requires sequencing
312 library preparation with lower input amounts. The use of ^{5E}U is widely established for *in vivo*
313 applications in the field such as fluorescence microscopy. We thus envision that nano-ID is in
314 principle applicable to many types of organisms, cells and conditions.
315
316

317 **Methods**

318

319 **Labeling and direct RNA nanopore sequencing of synthetic RNAs.** Labeled synthetic RNAs
320 and synthetic control RNAs are derived from selected RNAs of the ERCC RNA Spike-in Mix
321 (Ambion) as described in [26]. Characteristics of selected RNAs of the ERCC RNA Spike-in Mix
322 are listed in (**Supplementary Table 3**). Briefly, selected spike-in sequences were cloned into a
323 pUC19 cloning vector and verified by Sanger sequencing. For IVT template generation, the
324 plasmid (3 µg) was linearized using EcoRV-HF (blunt end cut) digestion mix containing
325 CutSmart buffer and EcoRV-HF enzyme. The digestion mix was incubated at 37 °C for 1 h and
326 the reaction was terminated adding 1/20 volume of 0.5 M EDTA. Subsequently, DNA was
327 precipitated in 1/10 volume of 3 M sodium acetate pH 5.2, and 2 volumes of 100 % ethanol at -
328 20 °C for 15 min. DNA was collected by centrifugation at 4 °C and 16,000 x g for 15 min. The
329 pellet was washed twice using 75 % ethanol. DNA was air-dried and resuspended in 5 µL of
330 H_2O at a concentration of 0.1-1.0 µg/µL (quantified by NanoDrop). Synthetic RNAs were *in*
331 *vitro* transcribed using the MEGAscript T7 kit (Ambion). *In vitro* transcription (IVT) of
332 synthetic control RNAs was performed following the manufacturer's instruction. For IVT of
333 labeled synthetic RNAs, 100 % of UTP (resp. GTP) was substituted with either 5-ethynyl-UTP
334 (^{5E}U, Jena Bioscience), 5-bromo-UTP (^{5Br}U, Sigma), 5-iodo-UTP (^{5I}U, TriLink BioTechnologies
335 LLC), 4-thio-UTP (^{4S}U, Jena Bioscience) or 6-thio-GTP (^{6S}G, Sigma). Note that, for performing
336 a successful IVT with 4-thio-UTP and 6-thio-GTP, only a reduction to 80% substitution gave
337 successful yield. IVT reactions were incubated at 37 °C. After 4 h, reaction volume was filled up
338 with H_2O to 40 µL, then 2 µL of TURBO DNase was added and incubated at 37 °C for
339 additional 15 min. Synthetic RNAs were purified with RNAClean XP beads (Beckman Coulter)
340 following the manufacturer's instructions. The final synthetic RNA pool contained equal mass of
341 all respective synthetic RNAs in a given library (**Supplementary Table 1**). RNA was quantified
342 using Qubit (Invitrogen). RNA quality was assessed with the TapeStation System (Agilent)
343 Synthetic RNA pools were poly(A)-tailed using the *E. coli* Poly(A) Polymerase (NEB). The
344 reaction was incubated for 5 min and stopped with 0.1 M EDTA. Spike-ins were then purified
345 with phenol:chloroform:isoamyl alcohol and precipitated. Poly(A)-tailed synthetic RNA pools
346 were subsequently subjected to direct RNA nanopore sequencing library preparation (SQK-
347 RNA001, Oxford Nanopore Technologies) following manufacturer's protocol. All libraries were

348 sequenced on a MinION Mk1B (MIN-101B) for 20 h, unless reads sequenced per second
349 stagnated dramatically.

350

351 **Culturing of human K562 cells.** Human K562 erythroleukemia cells were obtained from
352 DSMZ (Cat. # ACC-10). K562 cells were cultured antibiotic-free in accordance with the DSMZ
353 Cell Culture standards in RPMI 1640 medium (Thermo Fisher Scientific) containing 10 % heat
354 inactivated fetal bovine serum (FBS) (Thermo Fisher Scientific), and 1x GlutaMAX supplement
355 (Thermo Fisher Scientific) at 37 °C in a humidified 5 % CO_2 incubator. Cells used in this study
356 display the phenotypic properties, including morphology and proliferation rate, that have been
357 described in literature. Cells were verified to be free of mycoplasma contamination using Plasmo
358 Test Mycoplasma Detection Kit (InvivoGen). Biological replicates were cultured independently.

359

360 **^{5E}U labeling and direct RNA nanopore sequencing of human K562 cells.** K562 cells were
361 kept at low passage numbers (<6) and at optimal densities (3x10^5 - 8x10^5) during all
362 experimental setups. Per biological replicate, K562 cells were diluted 24 h before the experiment
363 was performed (**Supplementary Table 1**). Per ^{5E}U 60 min sample (4 replicates), cells were
364 incubated at 37 °C, 5 % CO_2 for 1 h after a final concentration of 500 µM 5-Ethynyluridine (^{5E}U,
365 Jena Bioscience) was added. Per ^{5E}U 24 h sample (3 replicates), cells were incubated at 37 °C,
366 5% CO_2 for 24 h. ^{5E}U was added 3 times during the 24h incubation, i.e. every 8 hours (0h, 8h,
367 16h) at a final concentration of 500 µM. Control samples were not labeled (3 replicates). Per ^{5E}U
368 60 min HS (heat shock) sample (3 replicates), cells were incubated at 42 °C for 5 min (until cell
369 suspension reached 42 °C), and then ^{5E}U was added at a final concentration of 500 µM. Further,
370 heat shock treatments were performed in a water bath (LAUDA, Aqualine AL12) at 42 °C. for 1
371 h. Temperature was monitored by thermometer. To avoid transcriptional changes by freshly
372 added growth medium, fresh growth medium was added ~24 h prior to heat shock treatments [43].
373 Exactly after the labeling duration, cells were centrifuged at 37 °C and 1,500 x g for 2 min. Total
374 RNA was extracted from K562 cells using QIAzol (Quiagen) according to manufacturer's
375 instructions. Poly(A) RNA was purified from 1 mg of total RNA using the µMACS mRNA
376 Isolation Kit (Milteny Biotec) following the manufacturer's protocol. The quality of poly(A)
377 RNA selection was assessed using the TapeStation System (Agilent). Poly(A) selected RNAs

378 were subsequently subjected to direct RNA nanopore sequencing library preparation (SQK-
379 RNA001, Oxford Nanopore Technologies) following manufacturer's protocol with 1000 ng
380 input. All libraries were sequenced on a MinION Mk1B (MIN-101B) for 48 h, unless reads
381 sequenced per second stagnated dramatically.

382

383 **RNA-seq.** Two biological replicates of K562 cells were diluted 24 h before the experiment was
384 performed. Per replicate, 3.6 x 10^7 cells in growth medium were labeled at a final concentration
385 of 500 µM 4-thio-uracil (4sU, Sigma-Aldrich), and incubated at 37 °C, 5 % CO_2 for 5 min.
386 Exactly after 5 min of labeling, cells were harvested at 37 °C and 1,500 x g for 2 min. Total
387 RNA was extracted from K562 cells using QIAzol according to manufacturer's instructions
388 except for the addition of 150 ng RNA spike-in mix [26] together with QIAzol. To isolate polyA
389 RNA from 75 µg of total RNA, two subsequent rounds of purification by Dynabeads
390 Oligo $(dT)_{25}$ (invitrogen) were performed. Purification based on manufacturer's instructions was
391 performed twice, using 1 mg of Dynabeads Oligo $(dT)_{25}$ beads for the first round and 0.5 mg for
392 the second round of purification. The quality of polyadenylated RNA selection was assessed
393 using RNA ScreenTape on a TapeStation (Agilent). Sequencing libraries were prepared using the
394 NuGEN Ovation Universal RNA-seq kit according to manufacturer's instructions. Fragments
395 were amplified by 10 cycles of PCR, and sequenced on an Illumina NextSeq 550 in paired-end
396 mode with 75 bp read length.

397

398 **Direct RNA nanopore sequencing data preprocessing of synthetic RNAs**. Direct RNA
399 nanopore sequencing reads were obtained for each of the samples (**Supplementary Table 1**).
400 FAST5 files were base-called using Albacore 2.3.1 (Oxford Nanopore Technologies) with the
401 following parameters: read_fast5_basecaller.py -f FLO-MIN106 -k SQK-RNA001. Direct RNA
402 nanopore sequencing reads were mapped with GraphMap 0.5.2 [44] to the synthetic RNA reference
403 sequence with the following parameters: graphmap align --evalue 1e-10. Further data processing
404 was carried out using the R/Bioconductor environment.

405

Direct RNA nanopore sequencing data preprocessing of human K562 cells. Direct RNA nanopore sequencing reads were obtained for each of the samples (**Supplementary Table 1**). FAST5 files were base-called using Albacore 2.3.1 (Oxford Nanopore Technologies) with the following parameters: read_fast5_basecaller.py -f FLO-MIN106 -k SQK-RNA001. Direct RNA nanopore sequencing reads were mapped with Minimap2 2.10 [45] to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium) with the following parameters: minimap2 -ax splice -k14 --secondary=no. Samtools [46] was used to quality filter SAM files, whereby alignments with MAPQ smaller than 20 (-q 20) were skipped. Further data processing was carried out using the R/Bioconductor environment and custom python scripts.

Probability of ^{5E}U-labeled RNA isoform identification based on synthetic RNAs. The following parameters were collected on the direct RNA nanopore sequencing data of synthetic RNAs and used to calculate the probability of identification of a ^{5E}U-labeled RNA isoform as labeled. Detectability d - the number of ^{5E}U caused mismatches in the ^{5E}U-labeled sample. Background b - the number of U caused mismatches in the unlabeled control sample. Given these parameters, the probability of identification p can be calculated as the probability of a U-based mismatch being caused by a ^{5E}U in the transcript as

$$p = 0.25 \cdot 0.028 \cdot \big(d \cdot (1 - b)\big)$$

with *0.25* - the empirical probability of U content, and labeling efficiency *0.028* - the empirical probability of a U being replaced by a ^{5E}U in the labeling process [32]. This then allows to calculate the probability of labeled RNA identification p^{RNA} as

$$p^{RNA} = 1 - (1 - p)^{\#bases}$$

, the probability, that an RNA contains at least 1 detectable ^{5E}U.

Definition of transcription units based on the UCSC RefSeq genome assembly GRCh38 (RefSeq-TUs). For each annotated gene, transcription units were defined as the union of all existing inherent transcript isoforms (UCSC RefSeq GRCh38).

434 **Definition of isoform-independent exonic and intronic regions (constitutive exons and**
435 **introns).** Isoform-independent exonic and intronic regions were determined using a model for
436 constitutive exons [47] and constitutive introns respectively based on UCSC RefSeq annotation
437 (GRCh38).

438

439 **Isoform determination for human K562 cells.** The FLAIR (Full-Length Alternative Isoform
440 analysis of RNA) algorithm [22] was used for the correction and isoform definition of raw human
441 K562 direct RNA nanopore reads. Corrected and collapsed isoforms were obtained by adding
442 short-read data (RNA-seq) to help increase splice site accuracy of the nanopore read splice
443 junctions (https://github.com/BrooksLabUCSC/FLAIR).

444

445 **Parameter collection for neural network training and classification.** For each read in each
446 human K562 sample (^{5E}U 60 min, Control, ^{5E}U 24 h & ^{5E}U 60 min HS) we obtained ~1500
447 parameters from three different layers: Raw signal (ionic current), base-call event probabilities
448 and alignment derived mismatch properties. As raw signal, 1193 parameters were gathered
449 consisting of the raw ionic current measurements gathered for each possible 5-mer of nucleotides
450 as well as the raw ionic current measurements gathered for each possible 3-mer centered in a 5-
451 mer. The latter parameters were collected for U-containing and non-U-containing instances. In
452 addition to that, raw ionic current measurements were gathered for 5-mers with all possible
453 nucleotides in their center position also for U-containing and non-U-containing instances, as well
454 as 5-mers exclusively leading or lagging U content. All collected raw signal parameters were z-
455 score normalized on all non-U-containing instances given the mean values of the pore model on
456 which the original base-calling algorithm is based provided by Oxford Nanopore Technologies.
457 As base-call event probabilities, 120 parameters were gathered including 'model state', 'move',
458 'weights', 'p model state', the probability that 'model state' gave rise to the observation of the
459 event, the most probable 'model state', the probability that 'p model state' gave rise to the
460 observation of the event and the probabilities that events may be associated with the certain base
461 from the event probabilities table provided by the base-calling algorithm. As alignment derived
462 mismatch properties, 135 parameters were gathered including length of the reads, nucleotide

occurrences, number of nucleotide transitions, number of inserts and deletions on a single nucleotide basis as well as on a 5-mer basis for U-containing and non-U-containing instances.

Neural network training, validation and classification of human RNA isoforms into ^{5E}U-labeled and unlabeled. Neural network was trained on the ^{5E}U 24 h versus Control samples under the assumption that ^{5E}U 24 h sample solely contains labeled reads and the fact that the Control sample solely contains unlabeled reads. The trained neural network consists of a batch normalization layer and 3 dense layers with decreasing output shape (**Supplementary Figure 3a**). 2 dropout layers (with 25% dropout) in between regularize the attempted classification. Training was conducted on 404.201 reads, validation was performed on 173.240 reads in 40 epochs with the R interface to Keras on a TensorFlow backend [48], as

$$model <- keras_model_sequential()$$

$$model \%>\%$$

$$layer_batch_normalization(input_shape = 1448) \%>\%$$

$$layer_dense(units = 64, activation = "relu", input_shape = 1448) \%>\%$$

$$layer_dropout(rate = 0.25) \%>\%$$

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

$$layer_dropout(rate = 0.25) \%>\%$$

$$layer_dense(units = 1, activation = "sigmoid")$$

$$model \%>\% compile($$

$$optimizer = optimizer_rmsprop(),$$

$$loss = 'binary_crossentropy')$$

The neural network was 5-fold cross-validated with an accuracy of 0.87 and a false discovery rate (FDR) of 0.025 and used to classify reads of the ^{5E}U 60 min and ^{5E}U 60 min HS samples into ^{5E}U-labeled and unlabeled. A ROC analysis (1 – specificity vs sensitivity) for all reads of the test set showed an area under the curve (AUC) of 0.94. For reads with an alignment length larger than 500 nt and 1000 nt the AUC improved to 0.96. Note that, limiting the neural network

classification to reads produced in the first few hours of sequencing, i.e. reads with a generally higher accuracy, improves the AUC to 0.98.

Random forest training, validation and classification of human RNA isoforms into ^{5E}U-labeled and unlabeled. For validation purposes, a random forest [49] was trained on the ^{5E}U 24 h versus Control samples on the same data as the neural network above. The random forest was 5-fold cross-validated with an accuracy of 0.85 and a false discovery rate (FDR) of 0.32 and used to classify reads of the ^{5E}U 60 min sample into ^{5E}U-labeled and unlabeled.

Poly(A)-tail length determination. Poly(A)-tail length is estimated by identifying the dwell time of the poly(A)-tail in the nanopore. For each direct RNA nanopore sequencing read, the raw signal readout of the nanopore in pico-Ampere [pA] was extracted from the FAST5 file. Every data point above the 99.99% quantile or below the 0.001% quantile was set to the respective cut-off value for reasons of robustness (**Supplementary Figure 5c, upper panel**). Subsequently kmeans clustering was used to define two trend lines at 1/3 and 2/3 the distance between the two cluster centers. The two trend lines were then used to squish the raw data by taking the parallel minimum or maximum (**Supplementary Figure 5c, lower panel**). A loss score of a piecewise linear function of two consecutive segments of the trend lines is then used to identify segments along the squished data points (**Supplementary Figure 5c, middle panel**). The length of the third identified segment r_j is used to calculate the length of the poly(A)-tail l_j of read j in sample i as

$$l_j = \underset{j}{\text{median}}(s_j) \cdot \frac{r_j}{hertz_i} + 5$$

with the sequencing read speed s_j of read j in [nt/s] and the frequency $hertz_i$ in [Hz] used in measuring sample i and 5 additional adenines that are concealed in the flanking 5-mers.

Intron retention ratio. For each RefSeq-TU (UCSC RefSeq GRCh38) the intron retention ratio for the ^{5E}U 60 min and ^{5E}U 60 min HS samples were calculated using the above defined model of constitutive exons and introns by calculating the ratio of length normalized coverages of the

521 maximum value for all respective introns and the average of all respective exons. This yielded
522 358 gene loci with at least 5% intron retention in either of the samples.

523

524 **RNA stability (degradation rate $\boldsymbol{\lambda_{ij}}$, half-life $\boldsymbol{hl_{ij}}$) and synthesis rate $\boldsymbol{\mu_{ij}}$ estimation of**
525 **human RNA isoforms.** Each neural network classified direct RNA nanopore sequencing read of
526 the ^{5E}U 60 min and ^{5E}U 60 min HS samples was assigned to a FLAIR defined human isoform (or
527 RefSeq-TU) either as ^{5E}U-labeled L_{ij} and unlabeled $T_{ij} - L_{ij}$. The resulting counts were
528 subsequently converted into synthesis rates μ_{ij} and degradation rates λ_{ij} for isoform i in sample
529 j assuming first-order kinetics as in [24] using the following equations:

530 $$\lambda_{ij} = -\alpha_j - \frac{1}{\mathrm{t}} \cdot log\left(1 - L_{ij}/T_{ij}\right)$$

531 $$\mu_{ij} = T_{ij}\left(\alpha_j + \lambda_{ij}\right)$$

532 where t is the labeling duration in minutes and α is the growth rate (dilution rate, i.e. the
533 reduction of concentration due to the increase of cell volume during growth) defined as

534 $$\alpha_j = \frac{\log(2)}{CCL_j}$$

535 with cell cycle length CCL_j [min]. The half-life hl_{ij} for isoform i in sample j can thus be
536 calculated as

537 $$hl_{ij} = \frac{\log(2)}{\lambda_{ij}}$$

538 in minutes [min].

539

540 **RNA-seq data preprocessing and antisense bias correction**. Paired-end 75 base reads with
541 additional 6 base reads of barcodes were obtained for each of the samples (**Supplementary**
542 **Table 1**). Reads were demultiplexed and mapped with STAR 2.3.0 [50] to the hg20/hg38
543 (GRCh38) genome assembly (Human Genome Reference Consortium). Samtools [46] was used to
544 quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped and
545 only proper pairs (-f2) were selected. Further data processing was carried out using the

546 R/Bioconductor environment. We used a spike-in (RNAs) normalization strategy essentially as
547 described [26] to allow observation of antisense bias ratio c_j (ratio of spurious reads originating
548 from the opposite strand introduced by the reverse transcription reaction). Antisense bias ratios
549 were calculated for each sample *j* according to

550 $$c_j = \underset{i}{\text{median}}\left(\frac{k_{ij}^{antisense}}{k_{ij}^{sense}}\right)$$

551 for all available spike-ins *i*. Read counts (k_{ij}) for spike-ins were calculated using HTSeq [51]. The
552 number of transcribed bases (tb_j) for all samples was calculated as the sum of the coverage of
553 evident (sequenced) fragment parts (read pairs only) for all fragments in addition to the sum of
554 the coverage of non-evident fragment parts for fragments with an inner mate interval not entirely
555 overlapping a Refseq annotated intron (UCSC RefSeq GRCh38). The number of transcribed
556 bases (tb_j) or read counts (k_j) for all features (RefSeq-TUs) were corrected for antisense bias c_j as
557 follows using the parameter calculated as described above. The real number of read counts or
558 coverage s_{ij} for transcribed unit *i* in sample *j* was calculated as

559 $$s_{ij} = \frac{S_{ij} - c_j A_{ij}}{1 - c_j^2}$$

560 where S_{ij} and A_{ij} are the observed numbers of read counts or coverage on the sense and antisense
561 strand. RPKs were calculated upon antisense bias corrected read counts (k_j) falling into the
562 region of a RefSeq-TU divided by its length in kilobases. Coverages were calculated upon
563 antisense bias corrected number of transcribed bases (tb_j) falling into the region of a RefSeq-TU
564 divided by its length in bases.
565

566 **References**

567 1 Chen, Y. *et al.* Principles for RNA metabolism and alternative transcription initiation
568 within closely spaced promoters. *Nat Genet* **48**, 984-994, doi:10.1038/ng.3616 (2016).
569 2 Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation
570 regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320,
571 doi:10.1038/ng.3142 (2014).
572 3 Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity
573 revealed by isoform profiling. *Nature* **497**, 127-131, doi:10.1038/nature12121 (2013).
574 4 Turner, R. E., Pattison, A. D. & Beilharz, T. H. Alternative polyadenylation in the
575 regulation and dysregulation of gene expression. *Semin Cell Dev Biol* **75**, 61-69,
576 doi:10.1016/j.semcdb.2017.08.056 (2018).
577 5 Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores.
578 *Nat Methods*, doi:10.1038/nmeth.4577 (2018).
579 6 Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read
580 sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol*
581 **33**, 736-742, doi:10.1038/nbt.3242 (2015).
582 7 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease.
583 *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).
584 8 Long, C. *et al.* Correction of diverse muscular dystrophy mutations in human engineered
585 heart muscle by single-site genome editing. *Sci Adv* **4**, eaap9004,
586 doi:10.1126/sciadv.aap9004 (2018).
587 9 Eswaran, J. *et al.* RNA sequencing of cancer reveals novel splicing alterations. *Sci Rep* **3**,
588 1689, doi:10.1038/srep01689 (2013).
589 10 Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171-194,
590 doi:10.1146/annurev-genet-120116-024704 (2017).
591 11 Falcone, C. & Mazzoni, C. RNA stability and metabolism in regulated cell death, aging
592 and diseases. *FEMS Yeast Res* **18**, doi:10.1093/femsyr/foy050 (2018).
593 12 Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136**, 763-
594 776, doi:10.1016/j.cell.2009.01.019 (2009).
595 13 Yamaguchi, T. *et al.* The CCR4-NOT deadenylase complex controls Atg7-dependent cell
596 death and heart function. *Sci Signal* **11**, doi:10.1126/scisignal.aan3638 (2018).

597 14 Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene
598 expression programs. *Hum Genet* **136**, 1043-1057, doi:10.1007/s00439-017-1791-x
599 (2017).
600 15 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat*
601 *Methods* **10**, 1177-1184, doi:10.1038/nmeth.2714 (2013).
602 16 Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: genome-wide determination of
603 poly(A) tail length and 3' end modifications. *Mol Cell* **53**, 1044-1052,
604 doi:10.1016/j.molcel.2014.02.007 (2014).
605 17 Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail
606 profiling reveals an embryonic switch in translational control. *Nature* **508**, 66-71,
607 doi:10.1038/nature13007 (2014).
608 18 Zhang, Z. *et al.* Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat*
609 *Methods* **16**, 307-310, doi:10.1038/s41592-019-0351-9 (2019).
610 19 Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals
611 new regulatory programs and genes that simultaneously express multiple major isoforms.
612 *Genome Res* **27**, 1759-1768, doi:10.1101/gr.220962.117 (2017).
613 20 Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M. & Pleiss, J. A. Detection of splice isoforms
614 and rare intermediates using multiplexed primer extension sequencing. *Nat Methods* **16**,
615 55-58, doi:10.1038/s41592-018-0258-x (2019).
616 21 Clark, M. *et al.* Long-read sequencing reveals the splicing profile of the calcium channel
617 gene CACNA1C in human brain. *bioRxiv*, 260562, doi:10.1101/260562 (2018).
618 22 Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation
619 in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv*
620 (2018).
621 23 Dolken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic
622 parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959-1972,
623 doi:10.1261/rna.1136108 (2008).
624 24 Miller, C. *et al.* Dynamic transcriptome analysis measures rates of mRNA synthesis and
625 decay in yeast. *Mol Syst Biol* **7**, 458, doi:10.1038/msb.2010.112 (2011).

626 25 Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and
627 degradation dynamics in mammalian cells. *Nat Biotechnol* **29**, 436-442,
628 doi:10.1038/nbt.1861 (2011).
629 26 Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225-
630 1228, doi:10.1126/science.aad9841 (2016).
631 27 Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat*
632 *Methods* **14**, 1198-1204, doi:10.1038/nmeth.4435 (2017).
633 28 Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-
634 seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat*
635 *Methods* **15**, 221-225, doi:10.1038/nmeth.4582 (2018).
636 29 Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353,
637 doi:10.1038/nature24286 (2017).
638 30 Clark, M. W., T.; Garcia-Bea, A.; Kleinman, J.; Hyde, T.; Weinberger, D.; Haerty, W.;
639 Tunbridge, E. Long-read sequencing reveals the splicing profile of the calcium channel
640 gene CACNA1C in human brain. *bioRxiv* **10.1101/260562**,
641 doi:https://doi.org/10.1101/260562 (2018).
642 31 Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores.
643 *bioRxiv*, doi:10.1101/068809 (2016).
644 32 Jao, C. Y. & Salic, A. Exploring RNA transcription and turnover in vivo by using click
645 chemistry. *Proc Natl Acad Sci U S A* **105**, 15779-15784, doi:10.1073/pnas.0808480105
646 (2008).
647 33 Abe, K. *et al.* Analysis of interferon-beta mRNA stability control after poly(I:C)
648 stimulation using RNA metabolic labeling by ethynyluridine. *Biochem Biophys Res*
649 *Commun* **428**, 44-49, doi:10.1016/j.bbrc.2012.09.144 (2012).
650 34 Bharmal, H., Clarke, S., Lemire, A., Agnew, B. & Kumar, K. Capture and Analysis of
651 Newly Synthesized RNA A Novel Enabling Technology to Study High Resolution Gene
652 Expression. *Journal of Biomolecular Techniques : JBT* **21**, S43-S43 (2010).
653 35 Rissland, O. S. The organization and regulation of mRNA-protein complexes. *Wiley*
654 *Interdiscip Rev RNA* **8**, doi:10.1002/wrna.1369 (2017).
655 36 Theodorakis, N. G., Zand, D. J., Kotzbauer, P. T., Williams, G. T. & Morimoto, R. I.
656 Hemin-induced transcriptional activation of the HSP70 gene during erythroid maturation

657 in K562 cells is due to a heat shock factor-mediated stress response. *Mol Cell Biol* **9**,
658 3166-3173 (1989).
659 37 Sistonen, L., Sarge, K. D., Phillips, B., Abravaya, K. & Morimoto, R. I. Activation of
660 heat shock factor 2 during hemin-induced differentiation of human erythroleukemia cells.
661 *Mol Cell Biol* **12**, 4104-4111 (1992).
662 38 Mathew, A., Mathur, S. K. & Morimoto, R. I. Heat shock response and protein
663 degradation: regulation of HSF2 by the ubiquitin-proteasome pathway. *Mol Cell Biol* **18**,
664 5091-5098 (1998).
665 39 Niskanen, E. A. *et al.* Global SUMOylation on active chromatin is an acute heat stress
666 response restricting transcription. *Genome Biol* **16**, 153, doi:10.1186/s13059-015-0717-y
667 (2015).
668 40 Vihervaara, A. *et al.* Transcriptional response to stress in the dynamic chromatin
669 environment of cycling and mitotic cells. *Proc Natl Acad Sci U S A* **110**, E3388-3397,
670 doi:10.1073/pnas.1305275110 (2013).
671 41 Vihervaara, A. *et al.* Transcriptional response to stress is pre-wired by promoter and
672 enhancer architecture. *Nat Commun* **8**, 255, doi:10.1038/s41467-017-00151-0 (2017).
673 42 Shalgi, R., Hurt, J. A., Lindquist, S. & Burge, C. B. Widespread inhibition of
674 posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell*
675 *Rep* **7**, 1362-1370, doi:10.1016/j.celrep.2014.04.044 (2014).
676 43 Mahat, D. B. & Lis, J. T. Use of conditioned media is critical for studies of regulation in
677 response to rapid heat shock. *Cell Stress Chaperones* **22**, 155-162, doi:10.1007/s12192-
678 016-0737-x (2017).
679 44 Sovic, I. *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap.
680 *Nat Commun* **7**, 11307, doi:10.1038/ncomms11307 (2016).
681 45 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-
682 3100, doi:10.1093/bioinformatics/bty191 (2018).
683 46 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
684 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
685 47 Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods
686 for normalization and differential expression in mRNA-Seq experiments. *BMC*
687 *Bioinformatics* **11**, 94, doi:10.1186/1471-2105-11-94 (2010).

48 keras: R Interface to 'Keras' (2018).

49 Classification and Regression by randomForest (2002).

50 Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* **51**, 11 14 11-19, doi:10.1002/0471250953.bi1114s51 (2015).

51 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

52 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

Acknowledgments

We would like to thank Johannes Söding, Christian Roth and Dmitry Tegunov (Max Planck Institute for Biophysical Chemistry) for advice on machine learning aspects. We would also like to thank Nikolaos Papadopoulos and Noah Wulff Mottelson (Max Planck Institute for Biophysical Chemistry) for help with python. In addition, we would like to thank Michael Lidschreiber (Max Planck Institute for Biophysical Chemistry) and Brian Lee (UCSC Genomics Institute) for help with the UCSC Genome Browser Track Hub. Moreover, we would like to thank Anna Sawicka and Kristina Žumer (Max Planck Institute for Biophysical Chemistry) for sharing the pUC19 spike-in plasmids. PC was supported by the Deutsche Forschungsgemeinschaft (SFB860, SPP1935), the European Research Council Advanced Investigator Grant TRANSREGULON (grant agreement No 693023), and the Volkswagen Foundation.

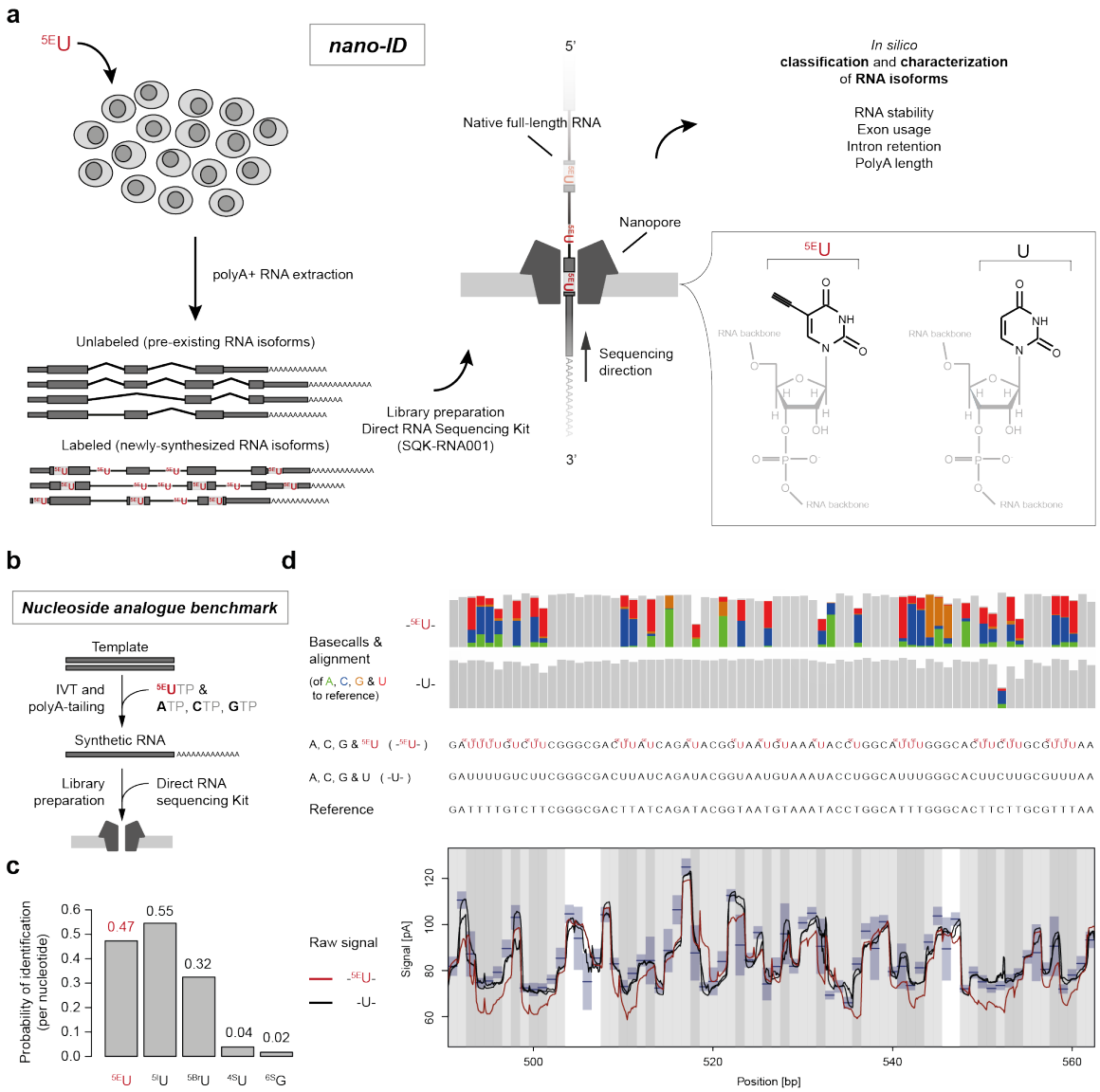
Competing interests

The authors declare that no competing interests exist.

Authors' contributions

KM, BS and SG carried out experiments. BS designed and carried out all bioinformatics analysis. BS conceptualized, designed and supervised research. BS and PC prepared the manuscript, with input from all authors.

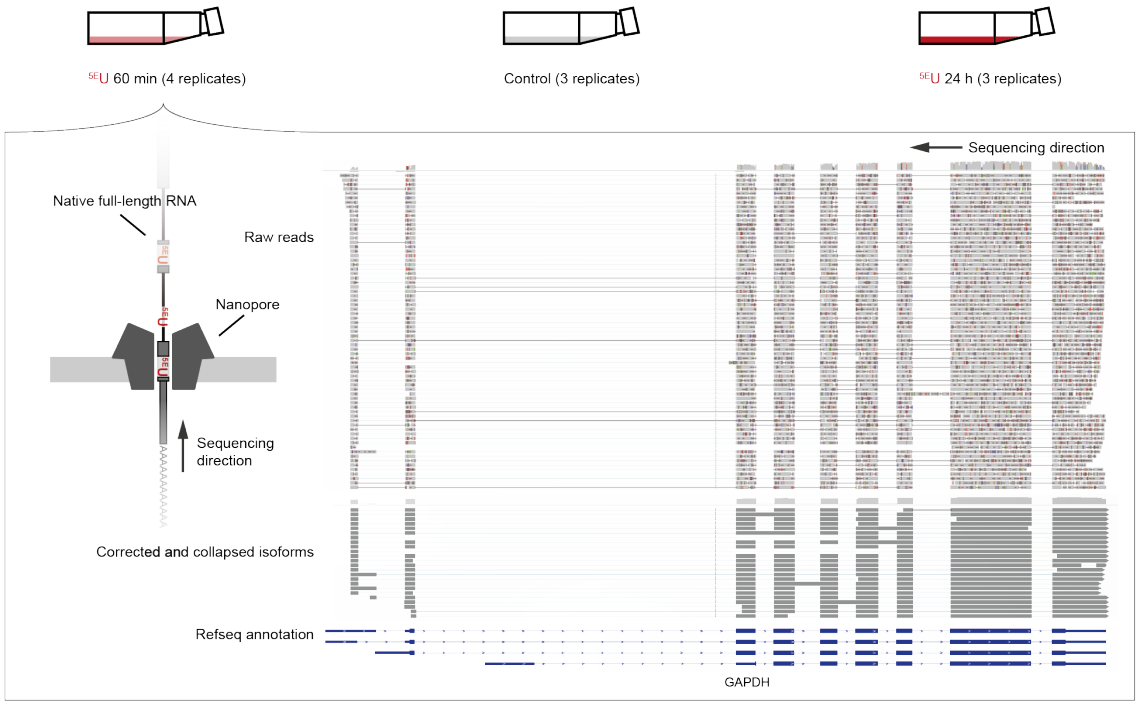
720 **Figures**



721

722 **Figure 1. Nanopore sequencing-based Isoform Dynamics (nano-ID) combines metabolic**
723 **RNA labeling with 'long-read' nanopore sequencing of native RNA molecules.** (a)
724 Experimental schematic of ^{5E}U-labeled RNA isoforms subjected to direct RNA 'long-read'
725 nanopore sequencing. Metabolic labeling of human K562 cells with the nucleoside analogue 5-
726 Ethynyluridine (^{5E}U) *in vivo*. Newly-synthesized RNA isoforms will incorporate ^{5E}U instead of
727 standard uridine (U) residues. This allows to distinguish the newly synthesized RNA isoforms
728 (Labeled) from pre-existing RNA isoforms (Unlabeled) *in silico* after sequencing the native full-
729 length molecules on an array of nanopores [5]. ^{5E}U containing RNA isoforms are computationally

730 traceable and thus allow classification. Identification and quantification of RNA isoforms
731 subsequently enable assessment of RNA stability, exon usage, intron retention and polyA-tail
732 length. (b) Experimental schematic to derive synthetic RNAs for nucleoside analogue
733 benchmark. RNAs were *in vitro* transcribed using either the standard bases A, U, C, G as a
734 control, or one of the natural bases was exchanged for a nucleoside analogue (shown for ^{5E}U). (c)
735 Barplot showing the probability of nucleoside analogue identification compared to natural
736 UTP/GTP based on base-miscalls (**Methods**) of all tested nucleoside analogues (^{5E}U, 5-
737 bromouridine (^{5Br}U), 5-iodouridine (^{5I}U), 4-thiouridine (^{4S}U) and 6-thioguanine (^{6S}G)). (d) Upper
738 panel: Base miscalls (colored vertical bars) of the standard base-calling algorithm for synthetic
739 RNAs containing ^{5E}U instead of U (-^{5E}U-, 3.563 molecules) and synthetic control RNAs (-U-,
740 15.840 molecules) in comparison to the original sequence (Reference) of an exemplary region on
741 synthetic RNA 'Spike-in 8' (**Methods, Supplementary Table 3**). Middle panel: Synthetic RNA
742 sequences with (-^{5E}U-) and without ^{5E}U (-U-) depicted above the reference sequence (Reference).
743 Lower panel: Alignment of the raw signal readout of the nanopore in pico-Ampere [pA] to the
744 reference sequence. Synthetic control RNAs (-U-) are shown in black. ^{5E}U containing synthetic
745 RNAs are shown in red (-^{5E}U-). ^{5E}U containing synthetic RNAs show a clear deviation from the
746 expected signal level in blue. Blue boxes indicate mean and standard deviation of the pore model
747 on which the original base-calling algorithm is based.
748
749

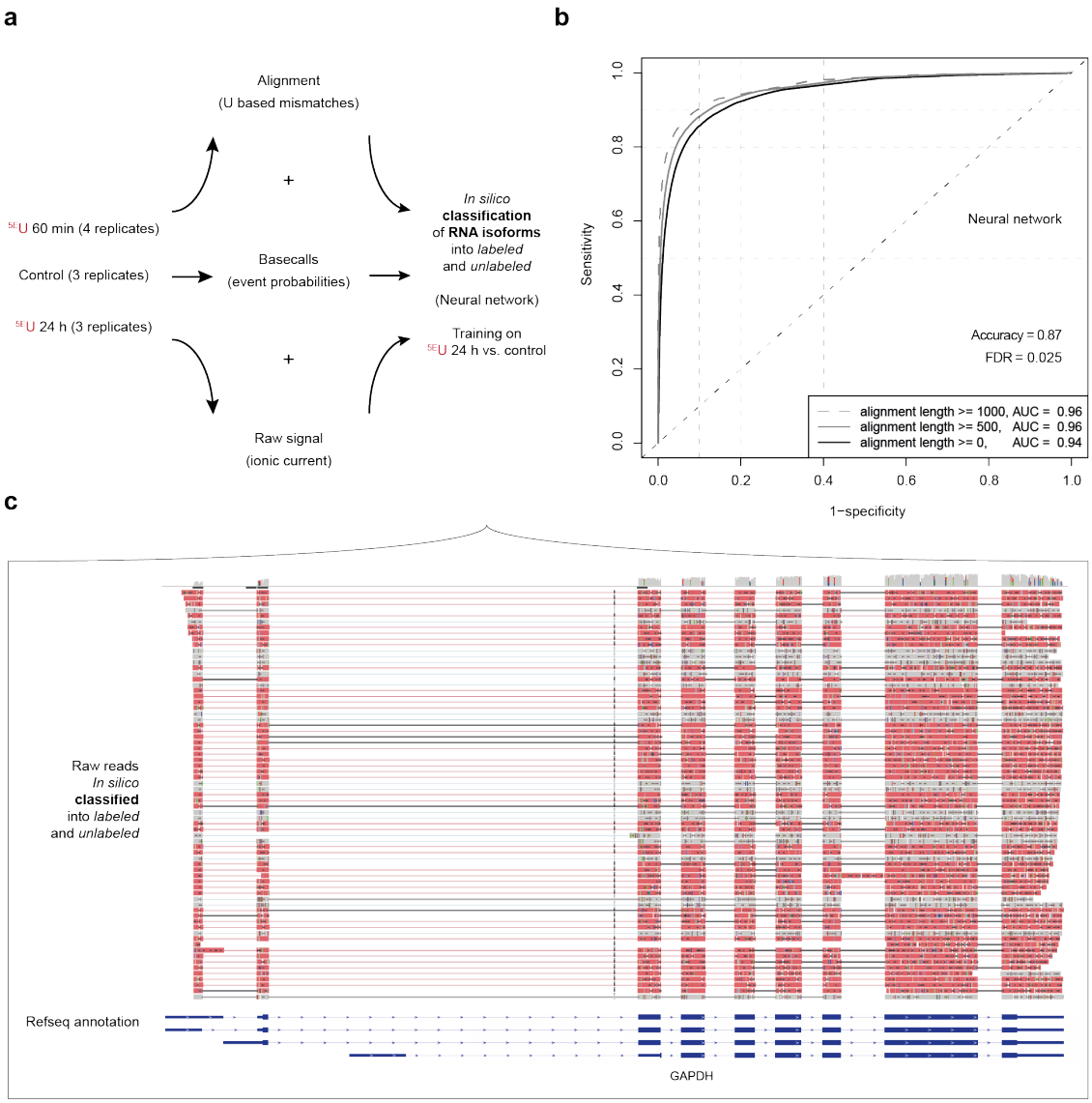


750

751 **Figure 2. Direct RNA 'long-read' nanopore sequencing of ^{5E}U-labeled RNA isoforms in**
752 **human K562 cells.** Upper panel: Illustration of the experimental set-up. Human K562 cells were
753 cultured in the presence of the nucleoside analogue ^{5E}U for 60 minutes (^{5E}U 60 min, 4 replicates)
754 and 24 h (^{5E}U 24 h, 3 replicates). Control samples were not labeled (Control, 3 replicates). Lower
755 panel: Genome browser view of direct RNA 'long-read' nanopore sequencing results of the
756 human GAPDH gene locus on chromosome 12 (~8 kbp, chr12: 6,532,405-6,540,375) visualized
757 with the Integrative Genomics Viewer (IGV, version 2.4.10; human hg38) [52]. From top to
758 bottom: raw nanopore sequencing reads (light grey, shown are typical aligned raw reads below
759 the accumulated coverage of all measured reads), corrected and collapsed isoforms (dark grey)
760 determined with the FLAIR algorithm [22] based on raw reads and RefSeq GRCh38 annotation
761 (blue).

762

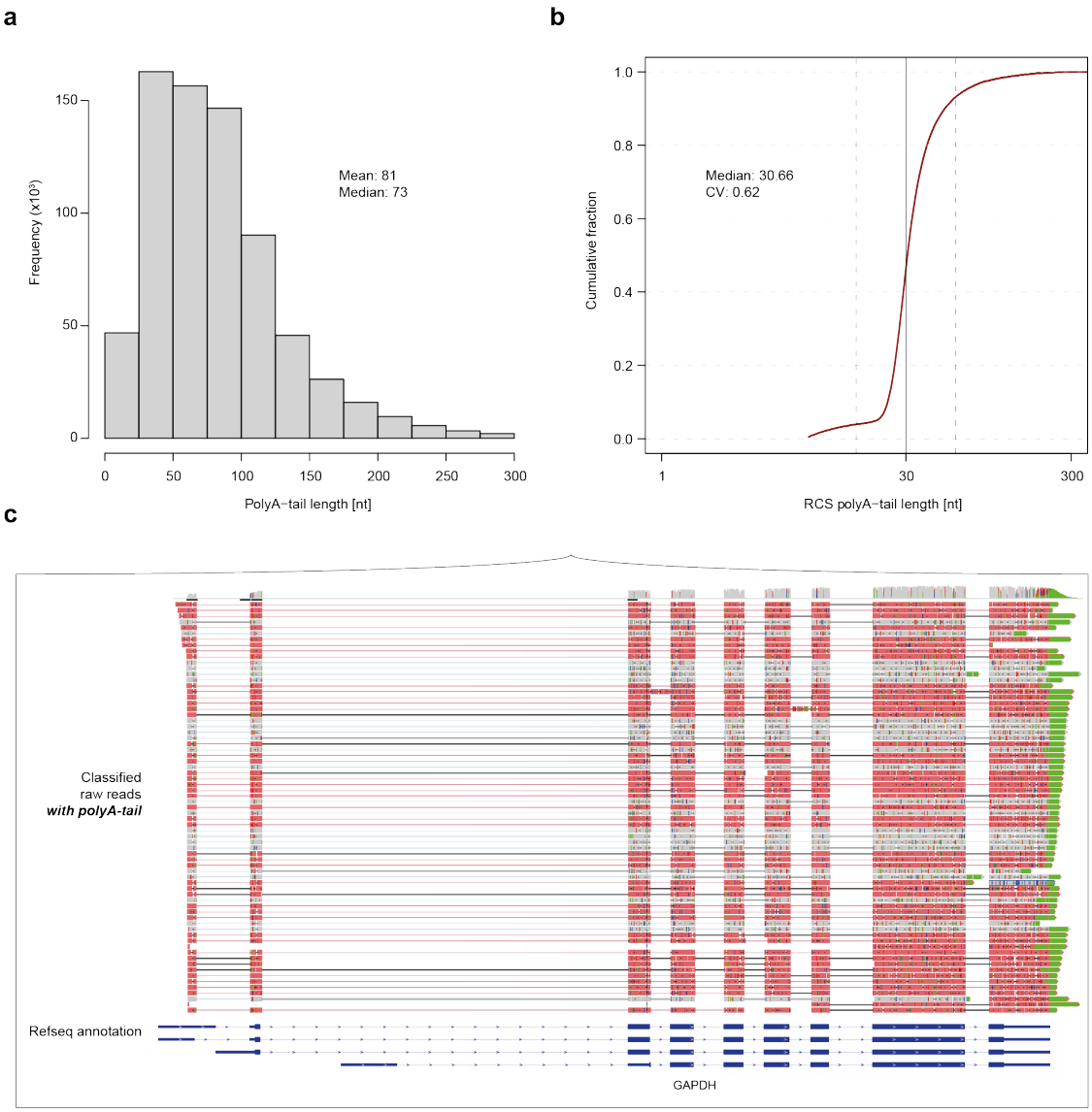
763



764

765 **Figure 3. Neural network based classification of human RNA isoforms into ^{5E}U-labeled and**
766 **unlabeled.** (a) Multi-layered data collection scheme. Parameter collection of human K562
767 samples (^{5E}U 60 min, Control & ^{5E}U 24 h) was realized on three different layers: Raw signal
768 (ionic current), base-call event probabilities and alignment derived U based mismatch properties
769 (**Methods**). Neural network was trained on the ^{5E}U 24 h versus Control samples with an
770 accuracy of 0.87 and a false discovery rate (FDR) of 0.025 and used to classify reads of the ^{5E}U
771 60 min samples into ^{5E}U-labeled and unlabeled. (b) ROC analysis of 5-fold cross-validated
772 neural network training. Plot shows ROC curves (1 – specificity versus sensitivity) for all reads
773 of the test set (black, alignment length >=0 nt, AUC = 0.94), for reads with an alignment length

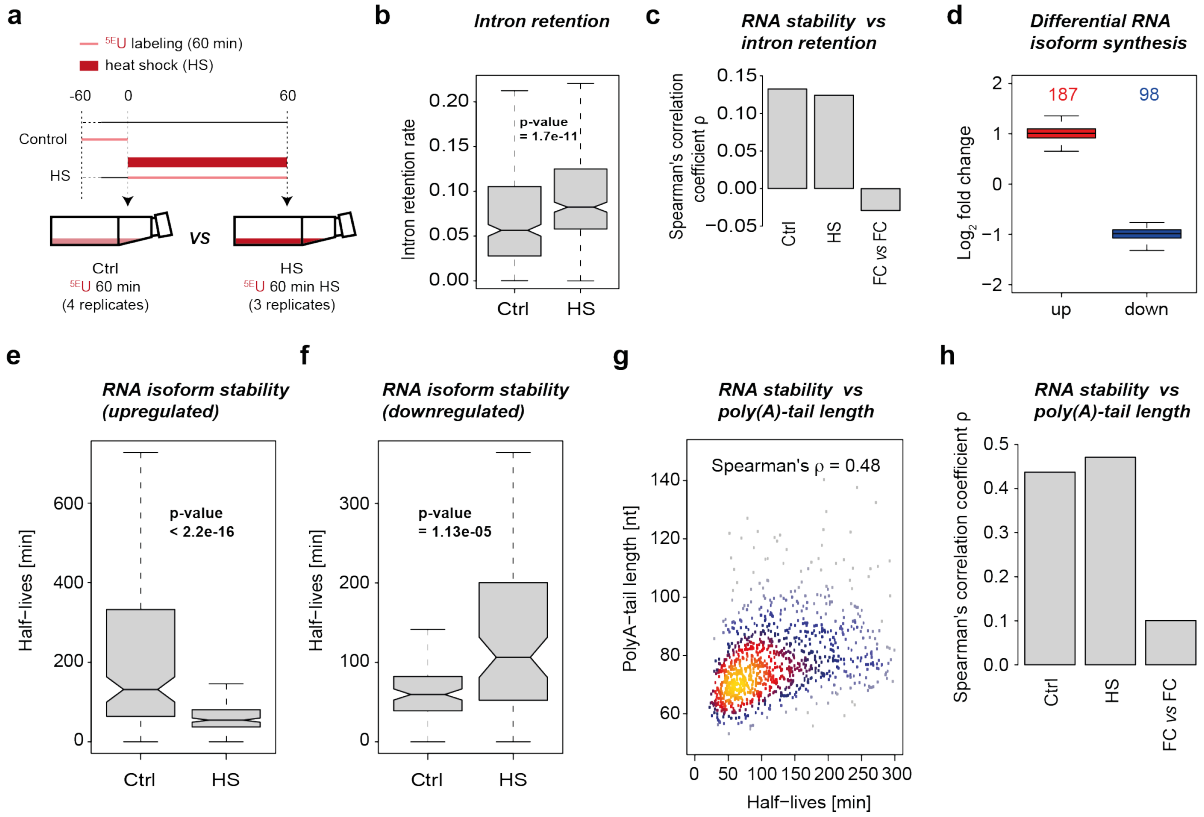
774 larger than 500 nt (grey, alignment length >=500 nt, AUC = 0.96) and for reads with an
775 alignment length larger than 1000 nt (dashed grey, alignment length >=1000 nt, AUC = 0.96). (c)
776 Genome browser view of classified direct RNA 'long-read' nanopore sequencing reads of the
777 human GAPDH gene locus on chromosome 12 (~8 kbp, chr12: 6,532,405-6,540,375) visualized
778 with the Integrative Genomics Viewer (IGV, version 2.4.10; human hg38) [52]. Unlabeled reads
779 are shown in grey, ^{5E}U-labeled reads are shown in red.
780



781

782 **Figure 4. Poly(A)-tail length determination of human RNA isoforms.** (a) Histogram of
783 poly(A)-tail length estimates of 714,536 RNA isoforms (mean: 81 nt, median: 73 nt). (b)
784 Cumulative distribution function of poly(A)-tail length estimates of the RNA calibration strand
785 (RCS, yeast derived spike-in RNAs that are equipped with a poly(A)-tail of exactly 30 adenines
786 (ONT, SQK-RNA001)). Vertical solid black line indicates optimal result of 30 nt (median: 30.6,
787 coefficient of variation: 0.62). Vertical dashed black lines indicate 2-fold in either direction. (c)
788 Genome browser view of classified direct RNA 'long-read' nanopore sequencing reads with
789 poly(A)-tail (green) of the human GAPDH gene locus on chromosome 12 (~8 kbp, chr12:

790 6,532,405-6,540,375) visualized with the Integrative Genomics Viewer (IGV, version 2.4.10;
791 human hg38) [52].
792



793

794 **Figure 5. nano-ID monitors RNA isoform dynamics during heat shock.** (a) Experimental set-
795 up of the heat shock treatment (60 min at 42 °C) in human K562 cells. (b) Boxplot shows intron
796 retention rate (**Methods**, min 5% in either condition) of 358 gene loci comparing heat shock (^{5E}U
797 60 min HS) against control (^{5E}U 60 min). (c) Bar plot shows correlation (Spearman's rank
798 correlation coefficient) of RNA half-lives and intron retention ratios before and after heat shock
799 (1,027 loci). The third bar shows the correlation of their respective folds. (d) Boxplot shows
800 upregulated (red) and downregulated (blue) RNA isoforms upon 60 min of heat shock (42 °C). A
801 minimum fold change of 1.25 and a maximum p-value of 0.1 was set for calling a significant
802 expression change. (e) Boxplot shows half-lives of significantly upregulated RNA isoforms
803 comparing heat shock (^{5E}U 60 min HS) against the control (^{5E}U 60 min). (f) As in (e) for
804 significantly downregulated RNA isoforms. (g) Scatter plot with color-coded density of RNA
805 half-lives and RNA poly(A)-tail lengths in both conditions. Shown are 1,230 highly expressed

806 RefSeq GRCh38 annotated genes. Correlation is calculated as Spearman's rank correlation
807 coefficient (0.48) rounded to the second decimal. (h) As in (c) using the RNA poly(A)-tail
808 lengths (1,230 loci).
809

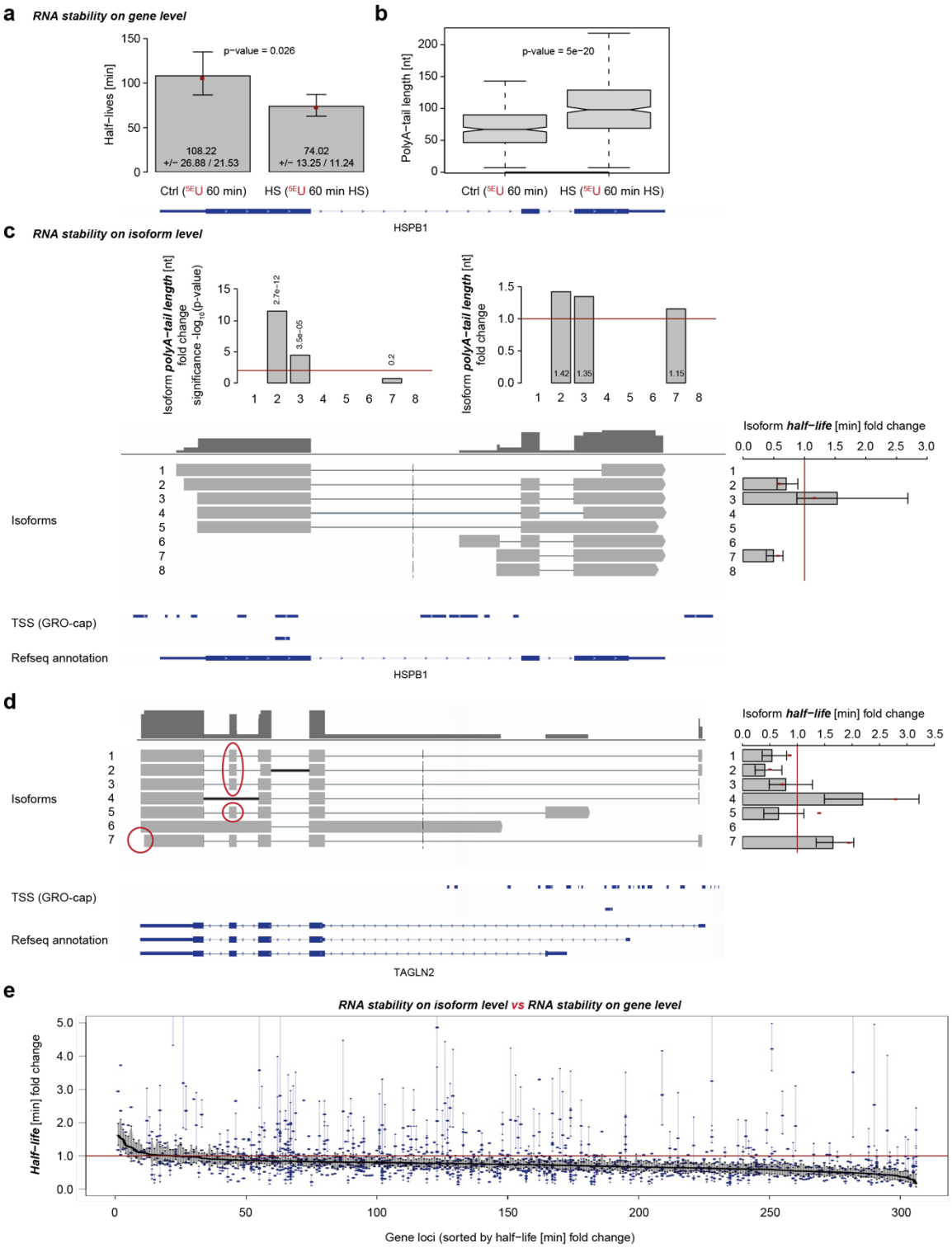


Figure 6. nano-ID resolves the characteristics of individual RNA isoforms. (a) Boxplot shows half-life estimates of RNAs from the human HSPB1 gene locus (chr6:31,813,514-31,819,942) comparing heat shock (HS, ^{5E}U 60 min HS) against control (Ctrl, ^{5E}U 60 min).

814 Standard deviation is shown as error bars. Red points depict half-life estimate of merged
815 replicates in each condition. (b) Boxplot shows the poly(A)-tail length distributions of RNAs
816 from the human HSPB1 gene locus. 437 RNAs from heat shocked samples (HS, ^{5E}U 60 min HS)
817 are compared to 341 RNAs in the respective control sample (Ctrl, ^{5E}U 60 min). (c) Schematic
818 shows direct RNA nanopore sequencing derived RNA isoforms at the human HSPB1 gene locus
819 above annotated transcription start sites (TSSs) from published GRO-cap data generated in K562
820 cells [2] and RefSeq GRCh38 annotation. Bar plots show RNA isoform half-life fold changes,
821 poly(A)-tail length fold changes and their respective significance as standard deviation (error
822 bars) or -$\log_{10}$(p-value). Red lines indicate no fold change or -$\log_{10}$(p-value) with p-value 0.01.
823 (d) As in (c) for RNA isoforms at the human TAGLN2 gene locus (chr1:159,916,107-
824 159,927,542). (e) Half-life fold change (y-axis) depicted for RNAs encoded by 306 high
825 confident gene loci (x-axis). All estimates are supported across biological replicates ($n \geq 3$) and
826 conditions (n=2). Half-life estimates for RNA encoded by the entire gene loci (combined) are
827 depicted as a black line (sorted in decreasing order). Blue dots represent individual RNA isoform
828 half-life estimates at respective gene loci (1,169 isoforms in total). Perpendicular blue and black
829 lines represent standard deviations of individual estimates. For individual RNA isoform half-life
830 estimates, standard deviations are only shown if not overlapping with the standard deviation of
831 the respective combined half-life estimates (black).
832