

1 A comprehensive examination of Nanopore native 2 RNA sequencing for characterization of complex 3 transcriptomes

4
5 Charlotte Soneson^{1,2,*§}, Yao Yao^{1,2}, Anna Bratus-Neuenschwander³, Andrea Patrignani³, Mark
6 D. Robinson^{1,2,*}, Shobbir Hussain^{4,*}

7
8 ¹Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland

9 ²SIB Swiss Institute of Bioinformatics, 8057 Zurich, Switzerland

10 ³Functional Genomics Centre Zurich, ETHZ/University of Zurich, 8057 Zurich, Switzerland

11 ⁴Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, United Kingdom

12
13 *Correspondence to charlotte.soneson@fmi.ch [C.S.], mark.robinson@imls.uzh.ch [M.D.R.], or
14 S.Hussain@bath.ac.uk [S.H.]

15
16 §Current affiliation: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute
17 of Bioinformatics, Basel, Switzerland

20 Abstract

21 A platform for highly parallel direct sequencing of native RNA strands was recently described by
22 Oxford Nanopore Technologies (ONT); in order to assess overall performance in transcript-level
23 investigations, the technology was applied for sequencing sets of synthetic transcripts as well as
24 a yeast transcriptome. However, despite initial efforts it remains crucial to further investigate
25 characteristics of ONT native RNA sequencing when applied to much more complex
26 transcriptomes. Here we thus undertook extensive native RNA sequencing of polyA+ RNA from
27 two human cell lines, and thereby analysed ~5.2 million aligned native RNA reads which
28 consisted of a total of ~4.6 billion bases. To enable informative comparisons, we also performed
29 relevant ONT direct cDNA- and Illumina-sequencing. We find that while native RNA sequencing
30 does enable some of the anticipated advantages, key unexpected aspects hamper its
31 performance, most notably the quite frequent inability to obtain full-length transcripts from single
32 reads, as well as difficulties to unambiguously infer their true transcript of origin. While
33 characterising issues that need to be addressed when investigating more complex
34 transcriptomes, our study highlights that with some defined improvements, native RNA
35 sequencing could be an important addition to the mammalian transcriptomics toolbox.

36 Introduction

37 The extent and observed complexity of cellular mRNA splicing patterns appear to have
38 generally expanded during the course of evolution ¹, and in more advanced species, several
39 subtly different mRNA transcript isoforms are likely to exist for most genes ^{2–4}. Within a
40 biological organism, the observed pattern of mRNA splicing for a given gene also frequently
41 varies between tissues and cell types, and can even respond to external cues or changes to the
42 environment ⁵. Thus, the ability to readily perform transcript-level functional investigations will
43 almost certainly enrich our understanding of a number of important biological processes. To
44 enable this to be accomplished in a reliable manner, methods that can unequivocally distinguish
45 and quantify the presence of transcript isoforms from the raw sequence reads are required.
46

47 Recently, long-read sequencing methodologies have been introduced into the transcriptomics
48 field, offering the opportunity to directly generate individual reads that can span the full length of
49 transcripts ^{6–12}. This could, for example, ameliorate problems associated with earlier
50 technologies' needs for DNA-mediated amplification and computational transcript assembly from
51 short sequence reads ^{13,14}. Notably, the newer long-read Oxford Nanopore Technologies (ONT)
52 platform now also provides the ability to sequence native RNA strands directly ¹⁵. In their study,
53 ONT described the efficient use of native RNA sequencing to yield reliable abundance
54 estimates of full-length transcripts from a yeast polyA+ transcriptome as well as sets of
55 standardized synthetic transcripts. However, larger transcriptome sizes, and in particular the
56 much higher complexity of splicing patterns that can be observed in higher organisms, might
57 potentially pose additional challenges during such transcript-level investigations.
58

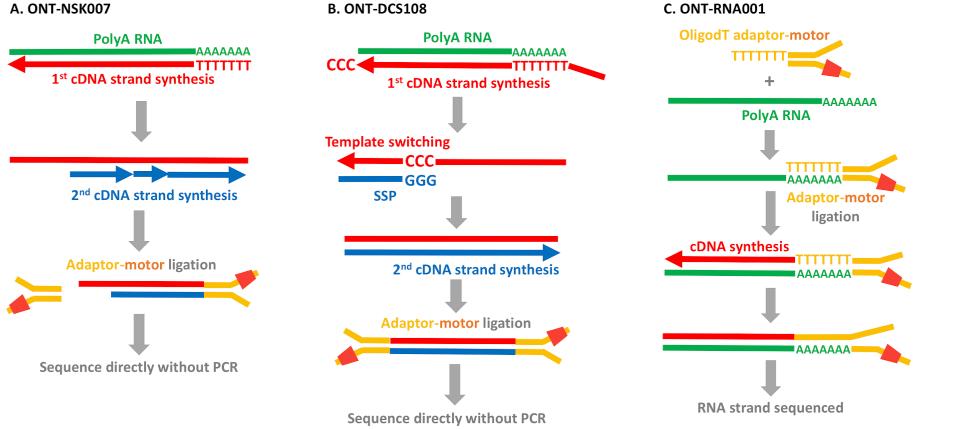
59 With the aim to characterize the gene- and transcript-level composition of complex
60 transcriptomes, in this study, we applied ONT long-read native RNA-sequencing to samples
61 from two human cell lines; HAP1 and HEK293. We also performed matched ONT direct (PCR-
62 free) cDNA sequencing as well as regular Illumina RNA-seq to enable relevant comparisons
63 and assessments. For computational analysis, considering the lower accuracy of Nanopore
64 sequencing, we primarily employed a reference-based approach, estimating abundances of a
65 set of annotated transcript isoforms and genes. An additional motivation for this was that also in
66 situations where a reference-free approach is used for transcript *identification*, reference-based
67 methods are often useful for subsequent *quantification* of transcript abundances. We present
68 our findings relating to differences between the performance of a variety of analysis algorithms,

69 and the potential advantages that current ONT direct RNA-seq brings over the traditional
70 Illumina sequencing, as well as current limitations of the technology.

71 **Results**

72 **Overall data characteristics**

73 We utilized three distinct ONT library preparation workflows in this study, all having in common
74 that RNA or cDNA molecules are sequenced directly without PCR. For our initial efforts, during
75 which direct cDNA sequencing kits were not available from ONT, we modified the regular ONT-
76 NSK007 2D PCR-based workflow in order to enable 1D direct cDNA sequencing (see *Methods*)
77 (Fig. 1A). We also made use of the subsequently released ONT-DCS108 kit for direct cDNA
78 sequencing, incorporating enrichment for full-length cDNAs (Fig. 1B). Most of the data
79 presented in this study, however, was obtained using the ONT-RNA001 kit for native RNA
80 sequencing (Fig. 1C). All ONT sequencing was performed using R9.4 flow cells, and the ONT
81 Albacore package was used for basecalling. We noted concerns of previous studies reporting
82 that filtering of reads during basecalling often results in a significant number of useful good-
83 quality reads being discarded ¹⁰. Indeed, in subsequent versions of available ONT Albacore
84 basecalling packages, filtering was either turned off as default or offered as an option. As
85 sequencing depth would likely be the key limiting factor influencing our downstream analyses,
86 and reasoning that true low-quality reads would be filtered during the alignment step, we thus
87 made use of the Albacore non-filtering option.



88

89 **Figure 1.** Overview of library preparation workflows used in this study. **A.** In the ONT-NSK007 cDNA
90 library preparation method, polyA RNA is used as a template for first strand cDNA synthesis which is
91 initiated from an oligodT primer. The NEB second strand cDNA synthesis module (E6111) is then used to
92 generate double-stranded cDNAs; here random primers are used to initiate cDNA synthesis, the products
93 of which are stitched together by DNA ligase. Note that since priming of second strand synthesis occurs
94 randomly, as depicted here this may not always begin from the very end of the first strand template.
95 Adaptor-motor complexes are then ligated to the double-stranded cDNA ends prior to direct sequencing
96 (the motor is an enzyme which will feed the nucleic strand into the nanopore). Note that instances where
97 the first strand overhang might be particularly long, as in the example depicted here, it is probably unlikely
98 that the adaptor-motor complex will ligate efficiently to enable sequencing of the second strand, though
99 the first strand will still be sequenced. **B.** To better enrich for full length cDNAs, the ONT-DCS108 direct
100 cDNA sequencing kit, which leverages the template switching phenomenon¹⁶, was used. When the first
101 strand cDNA synthesis reaches the end on the RNA molecule, the reverse transcriptase will add a few
102 non-template Cs to the end of the cDNA. A Strand Switching Primer (SSP) present in the reaction binds
103 to these non-templated Cs, and the reverse transcriptase then switches template from the RNA to the
104 SSP. The second cDNA strand, presuming its synthesis continues to the end of the first strand template,
105 will also span the full length of the primary polyA RNA template. Following adaptor ligation, the double
106 stranded cDNAs are then sequenced directly. **C.** The ONT-RNA001 workflow enables sequencing of
107 native RNA strands. Here an oligodT-adaptor-motor complex is ligated to the polyA end of the RNA. In
108 order to relax the secondary structure of the RNA (and thus help ensure efficient translocation of the RNA
109 strand through the nanopore), a cDNA synthesis step is performed. Since only the RNA strand has a
110 motor ligated, the RNA molecule but not the cDNA strand is always sequenced.

111

112

113 The yield from the different ONT protocols varied between approximately 500,000 and
114 1,500,000 unfiltered reads per sample (Supplementary Fig. 1A), and the read length
115 distributions were overall similar among the libraries, with a peak close to 1,000 bases
116 (Supplementary Fig. 1B). The distribution of average base qualities per read varied between the

117 different types of libraries (Supplementary Fig. 1C), with cDNA libraries as expected⁸ showing
118 higher base qualities than native RNA libraries. We also noticed an association between the
119 read length and the average base quality, with both very short and very long reads often having
120 lower quality (Supplementary Fig. 2).

121 Genome and transcriptome alignment

122 The ONT reads were aligned to the human reference genome and transcriptome using
123 minimap2 (see *Methods*). The N50 values for the portion of a read aligned to the genome were
124 907, 1,210, 1,043 and 941 bases for the *ONT-NSK007-HAP*, *ONT-DCS108-HAP*, *ONT-*
125 *RNA001-HAP*, and *ONT-RNA001-HEK* data sets, respectively (median aligned lengths for the
126 respective data sets were 633, 765, 621 and 596 bases, and the longest aligned read parts
127 were 75,756, 20,681, 12,839 and 14,692 bases in the four data sets). As we aligned unfiltered
128 reads, the alignment rates across library types were unsurprisingly only modest, varying
129 between 55 and 76% for the genome alignment, and from 45 to 73% for the transcriptome
130 alignment (Fig. 2A). As expected, the unaligned reads were enriched for low base qualities
131 (Supplementary Fig. 3A), and thus largely represented reads that would have been classified as
132 ‘failed’ during automatic filtering. In comparison, for the four matching Illumina libraries, STAR
133 aligned between 89 and 94% of the reads uniquely to the genome, with an additional 2-2.5% of
134 the reads aligning in multiple locations. The *ONT-DCS108-HAP* libraries showed the largest
135 differences between the genome and transcriptome alignment rates (60-69% vs 45-51%),
136 whereas the rates for the other data sets were more similar. It is possible that one explanation
137 for the large difference in genomic and transcriptomic alignment rates for this data set could be
138 a contamination with genomic DNA. In the *ONT-DCS108-HAP* libraries, compared to the set of
139 all reads aligning to the genome, the reads aligning *exclusively* to the genome showed a slight
140 enrichment for long reads with a lower average base quality but where a larger portion of the
141 read aligned. This pattern, however, was not reproduced in the other data sets, where the reads
142 aligning exclusively to the genome were rather shorter and showed a poorer agreement with the
143 reference (Supplementary Fig. 3).



144

Figure 2. Characterization of aligned reads. **A.** Total number of reads and the number of reads with a primary alignment to the genome or transcriptome, respectively, in each of the ONT libraries. The number displayed in each bar represents the alignment rate in % (the fraction of the total number of reads for which minimap2 reports a primary alignment). **B.** Fraction of the reads with a primary alignment to the genome or transcriptome, respectively, that also have at least one reported secondary or supplementary alignment. The lighter shaded parts of the secondary transcriptome alignment bars correspond to reads where all primary and secondary alignments are to isoforms of the same gene, while the darker shaded parts correspond to reads with reported alignments to transcripts from different genes. **C.** Investigation of supplementary genome alignments. Each supplementary alignment is categorized based on whether it is on the same chromosome and strand as the primary alignment, and if the alignment positions of the primary and supplementary alignments overlap.

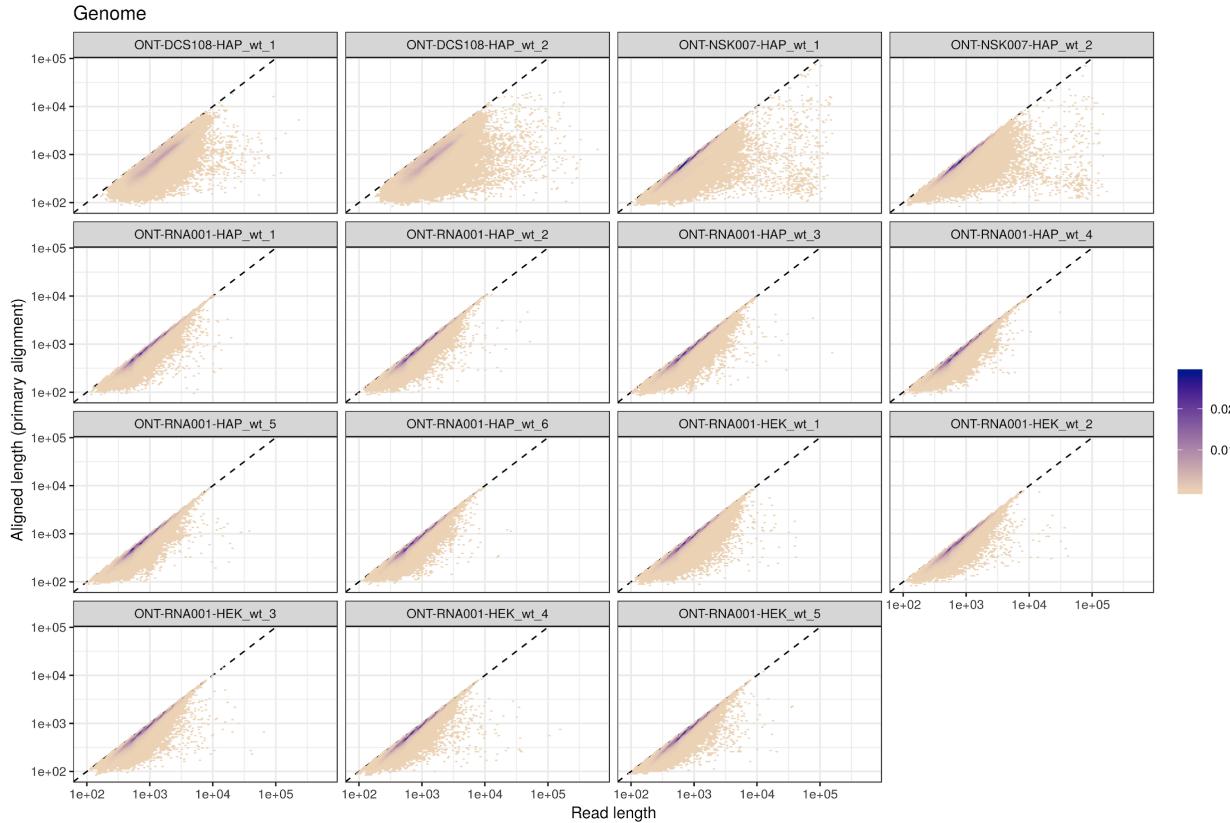
156

157 Approximately 40% of the reads with a primary genome alignment could be mapped to multiple
158 places in the genome, i.e., had also at least one reported secondary genome alignment (Fig.
159 2B). For most libraries, a single secondary alignment was most common, while for the ONT-
160 DCS108-HAP libraries, a larger fraction of reads had more than five secondary genome
161 alignments (Supplementary Fig. 4A). As expected, due to the high similarity among transcripts,
162 the fraction of reads with at least one secondary alignment increased to approximately 80% for

163 the transcriptome alignment (Fig. 2B). Again, a small number of secondary alignments was
164 most common (Supplementary Fig. 4B). The secondary alignment rate was only marginally
165 affected by increasing the $-p$ argument of minimap2, which sets the minimal accepted ratio
166 between the alignment score of secondary and primary alignments, to 0.99 instead of the
167 default 0.8 (Fig. 2B). For a majority of the reads, the target transcripts of all primary and
168 secondary transcriptome alignments were isoforms of the same gene (Fig. 2B), suggesting that
169 the main source of ambiguity is on the individual isoform level rather than on the gene level.
170 Only a small part of the secondary alignments (typically less than 5% of the reads) arose due to
171 the presence of multiple fully identical transcripts in the Ensembl reference catalog; in all
172 remaining cases there was at least some difference between the target transcripts of the
173 reported primary and secondary alignments. ‘Unavoidable’ secondary alignments may also be
174 the result of reads stemming from reference transcripts that are proper subsequences of other
175 reference transcripts. Among the 1,044,960 possible pairs of reference transcripts annotated to
176 the same gene in our annotation catalog, there are 64,437 such pairs (6.2%). In these
177 situations, in theory, a read could still be considered ‘unambiguously assignable’ to the shorter
178 transcript if it is similar enough, under the assumption that all ONT reads represent full-length
179 transcripts. Without this strong assumption, effective automated disambiguation would require a
180 reliable model of the read generation process, accounting for the probability of fragmentation of
181 RNA or cDNA molecules in the library preparation step and/or read truncation during the
182 sequencing-basecalling process. To investigate to what extent the secondary alignments in our
183 libraries could be the result of nested sets of reference transcripts, we extracted all reads with at
184 least one secondary transcriptome alignment, and among all primary and secondary
185 alignments, we selected the one for which the covered portion of the target transcript by the
186 read was highest. If the secondary alignments are the result of the true transcript of origin being
187 contained in the other target transcripts, we expect this maximally covered portion to be close to
188 1. Interestingly, for all the data sets except ONT-NSK007-HAP, while there is a clear peak close
189 to 1, there is also a broad distribution of lower coverage degrees (Supplementary Fig. 5). The
190 large number of secondary transcriptome alignments with alignment scores similar to the
191 reported primary alignment suggests that, despite the long read length, unambiguously inferring
192 the true transcript of origin for any given read is still highly non-trivial, and simply selecting the
193 reported primary alignment for downstream analysis can give misleading results.
194
195 While secondary alignments represent possible mapping positions of a read beyond the one
196 reported in the primary alignment, *supplementary* alignments arise when a read cannot be

197 mapped in a contiguous fashion, and consequently minimap2 splits the alignment into multiple
198 parts. We observed a comparatively large number of supplementary alignments in the *ONT-*
199 *DCS108-HAP* data set, both for genome and transcriptome alignments (Fig. 2B). Further
200 investigation revealed that in this data set, as well as in *ONT-NSK007-HAP*, a relatively large
201 fraction of the supplementary alignments overlapped the corresponding primary alignment, but
202 on the opposite strand (Fig. 2C). This observation is interesting, as we note that the ONT '1D²'
203 sequencing mode (<https://nanoporetech.com/>) exploits the observation that the second strand
204 (which also has a motor enzyme attached) of a double-stranded DNA molecule often enters the
205 sequencing nanopore immediately following the first strand during 1D sequencing. 1D²
206 sequencing chemistry is designed to further promote this observed phenomenon, and the
207 associated 1D² base-caller is specifically designed to efficiently split reads according to each
208 strand sequenced. Thus our findings of frequent overlapping supplementary alignments on
209 opposite strands may reflect un-split reads by the standard 1D basecaller. Accordingly, this type
210 of self-chimeric supplementary alignments were almost completely absent in the native RNA
211 samples where single strands, as opposed to double-stranded cDNAs, are present in libraries.
212 For the *ONT-NSK007-HAP* libraries, where often only one of the strands of the double-stranded
213 cDNAs will have a motor enzyme attached (Fig. 1), the relative frequency of this type of
214 supplementary alignments was somewhat lower than in the *ONT-DCS108-HAP* libraries (in
215 addition to the total rate of supplementary alignments being considerably lower), adding further
216 support to this speculated cause.

217
218 A peak of short low-quality unfiltered reads was consistently observed in the native RNA
219 libraries (Supplementary Fig. 1B), and the majority of these did not align adequately to either the
220 genome or the transcriptome (Supplementary Fig. 3A-B). More generally, for aligned reads, in
221 particular those shorter than 10,000 bases, most of the individual bases could be matched to a
222 position in the reference sequence, indicated by a large fraction of "M"s and consequently a low
223 fraction of insertions, deletions and soft-clipped bases in the CIGAR string (Fig. 3,
224 Supplementary Fig. 3C-D). Reads longer than 10,000 bases, which were mostly found in the
225 cDNA libraries, typically did not align end-to-end (Fig. 3). For the *ONT-DCS108-HAP* libraries, a
226 large fraction of the bases in the primary alignments were soft-clipped, corresponding to the
227 large number of supplementary alignments discussed above.



228

229 **Figure 3.** Total read length (x) vs aligned length (y, the sum of the number of “M” and “I” characters in the
230 CIGAR string) for the primary genome alignment of each read, in each of the ONT libraries. The colour
231 indicates point density.

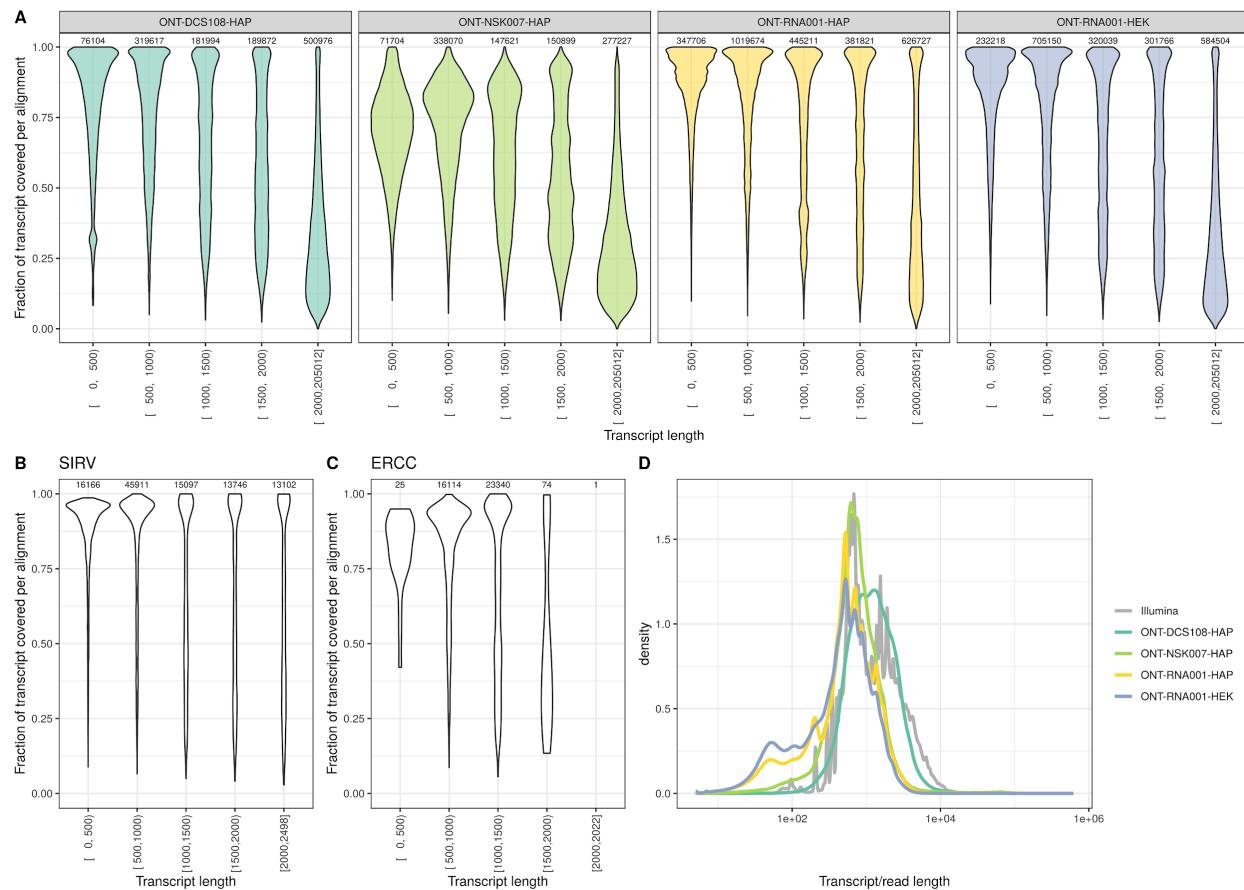
232

233 Incorporating the genomic coordinates of the annotated genes, we observed differences in the
234 gene body read coverage distribution between the libraries (Supplementary Fig. 6), with a
235 stronger 3’ coverage bias in the cDNA libraries than in the native RNA libraries. While given the
236 nature of the library preparation this was expected for the NSK007 cDNA libraries, it is also
237 quite possible that the template switching mechanism does not work to full efficiency in the
238 DCS108 cDNA protocol.

239 Coverage of full-length transcripts by individual ONT reads

240 To investigate to what extent individual ONT reads could be expected to represent full-length
241 transcripts, we selected the “best” target transcript for each read, starting from the set of all
242 primary and secondary transcriptome alignments obtained with minimap2, with $-p$ set to 0.99.
243 For each read, we kept all alignments for which the number of aligned nucleotides was at least
244 90% of the maximal such number across all alignments for the read, and among these, we
245 selected the one with the largest transcript coverage degree (number of “M” and “D” characters

246 in the CIGAR string of the alignment, divided by the annotated transcript length). While this
247 alignment does not necessarily represent the "true" origin of the read, the procedure gives an
248 upper bound of the degree of transcript coverage achieved by individual reads. As expected
249 from the *ONT-NSK007-HAP* library preparation, which does not involve full-length cDNA
250 enrichment (Fig. 1), reads from this sample achieved a lower degree of full-length transcript
251 coverage across the range of transcript lengths (Fig. 4A). While shorter transcripts could often
252 be completely covered by a single read in the *ONT-RNA001-HAP*, *ONT-RNA001-HEK* and
253 *ONT-DCS108-HAP* libraries, this was rarely the case for long transcripts (Fig. 4A,
254 Supplementary Figure 7). This observation, that many of the raw ONT reads do not appear to
255 represent full-length transcripts, needs to be taken into account during transcript identification
256 and quantification. Applying the same procedure to the SIRV and ERCC data sets from Garalde
257 *et al.*¹⁵ revealed that a majority of these synthetic transcripts were well covered by single reads
258 (Fig. 4B-C), confirming observations from previous studies^{9,15}; importantly, however, all
259 transcripts in the SIRV and ERCC catalogs are shorter than 2,500 bases. In the Ensembl
260 GRCh38.90 catalog, approximately 17% of the transcripts are longer than that, and the
261 coverage degree of these transcripts by single reads were generally less than 50%. This
262 suggests that while the synthetic transcript catalogs provide useful information about the
263 performance of long-read transcriptome sequencing and analysis methods, extrapolation of the
264 results to real, complex transcriptomes should be done with care.



265

266 **Figure 4.** Transcript coverage fraction by individual reads. **A.** Distribution of coverage fractions of
267 transcripts by individual reads, for each of the four ONT data sets, stratified by the length of the target
268 transcript. The 'target transcript' was selected to maximize the coverage fraction, among all reported long
269 enough alignments (see text), and thus the reported coverage fractions represent upper bounds of the
270 true ones. The number above each violin indicates the number of processed alignments to transcripts in
271 the corresponding length category. **B-C.** Distribution of coverage fractions of transcripts by individual
272 reads for the SIRV and ERCC data sets. **D.** Observed distribution of raw read lengths (for ONT data sets)
273 and expected distribution of transcript molecule lengths based on annotated transcript lengths and
274 estimated abundances in the Illumina samples. Values are aggregated across all samples within each
275 data set.

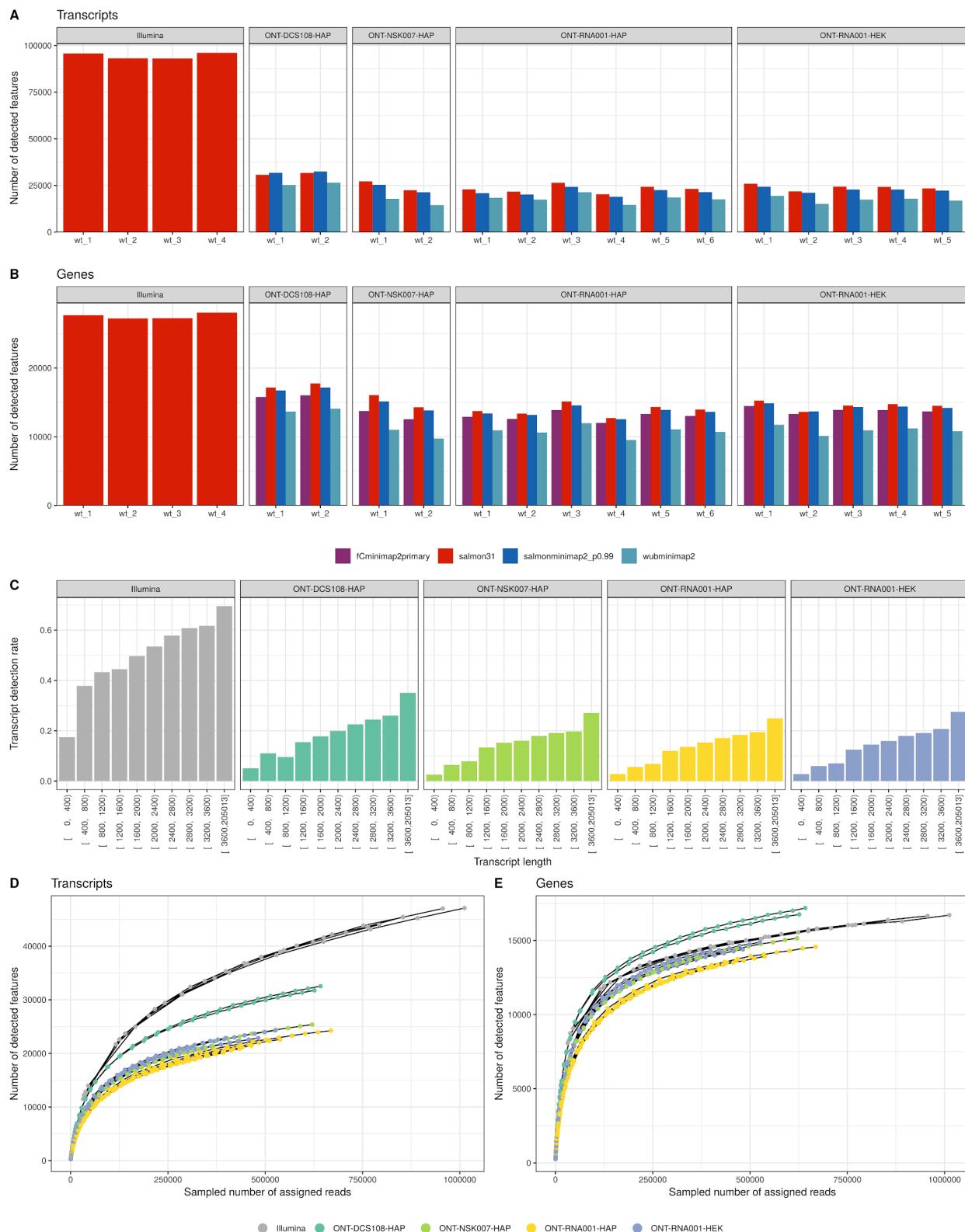
276

277 To further investigate the degree to which individual ONT reads are likely to represent full-length
278 transcripts, we compared the observed raw ONT read length distribution with the 'expected'
279 transcript length distribution in these samples, obtained by weighting the annotated transcript
280 lengths by the estimated transcript abundances (in transcripts per million - TPM) estimated by
281 Salmon in the Illumina samples. This analysis showed an apparent shortage of ONT reads in
282 the length range of the longest transcripts inferred to be expressed in the Illumina data (Fig.
283 4D). The ONT-DCS108-HAP samples were the exception; however, for many of the reads in
284 these libraries, the primary alignment does not cover the entire read (Fig. 3). This holds true
285 both for reads with a supplementary alignment and for those without one (Supplementary Fig.

286 8A-C). Further inspection of the longer annotated transcripts with high estimated abundances in
287 the Illumina samples revealed consistent base pair coverage by Illumina reads along the length
288 of these transcripts (Supplementary Fig. 8D), indicating that these were indeed likely to be truly
289 present in samples. Moreover, these transcripts were from ‘standard’ genes in that the vast
290 majority of the long transcripts with high estimated abundance in the Illumina samples were
291 annotated as protein coding, and they were found on almost all chromosomes. Overall, such
292 observations further illustrate that using current library preparation and sequencing workflows,
293 long transcripts are often not represented by single ONT reads.

294 Reference-based transcript detection and abundance quantification

295 Four reference-based methods were used to estimate transcript and gene abundances in each
296 of the ONT libraries. For two of these methods, we specifically evaluated the impact of data
297 preprocessing: for minimap2 followed by Salmon in alignment-based mode (denoted
298 `salmonminimap2`), we investigated the effect of setting the `-p` argument of minimap2 to different
299 values (the default of 0.8 as well as 0.99) in the transcriptome alignment step, and for Salmon in
300 quasi-mapping mode, we evaluated the effect of providing only the aligned bases of the reads
301 with a primary alignment anywhere in the genome (see *Methods*). Increasing `-p` to 0.99 led to a
302 slightly improved correlation between ONT transcript read counts and estimated transcript
303 abundances from the Illumina samples (obtained by Salmon in quasi-mapping mode), and thus,
304 in the following analyses, we set `-p` equal to 0.99 for Salmon following minimap2
305 (Supplementary Fig. 9). Removing the non-aligned bases before running Salmon did not
306 improve the correlations notably (Supplementary Fig. 9). Since this is a more involved
307 procedure, and further introduces a dependency on the genome alignments, we use the Salmon
308 quantifications obtained using the original, non-truncated reads for the rest of the analyses.



309

310 **Figure 5. Detection of annotated transcripts and genes. A-B.** Number of detected transcripts and genes
311 with the applied abundance estimation methods, in each library. Here, a feature is considered detected if

312 the estimated read count is ≥ 1 . **C**. Fraction of transcripts detected (with estimated count ≥ 1) in at least
313 one sample, stratified by transcript length, in the respective data sets. **D-E**. Saturation of transcript and
314 gene detection, in ONT and Illumina libraries. For each library, we subsampled the reads and recorded
315 the number of transcripts and genes detected with an estimated salmonminimap2 count (ONT libraries) or
316 Salmon count (Illumina libraries) ≥ 1 . The Illumina curves are truncated to the range of read numbers
317 observed in the ONT libraries.

318
319 We observed a large difference between the numbers of reads that were assigned to features
320 by the different quantification methods (Supplementary Fig. 10). The highest assignment rates
321 were consistently obtained with salmonminimap2, where all reads that were aligned to the
322 transcriptome were also subsequently assigned to features. featureCounts assigned a slightly
323 lower fraction of the reads to genes, while Salmon in quasi-mapping mode and Wub assigned
324 considerably fewer reads. However, the relatively low number of reads assigned by Salmon in
325 quasi-mapping mode were distributed across as many, sometimes more, genes and transcripts
326 as the reads assigned by salmonminimap2 (Fig. 5A-B), suggesting that no category of genes or
327 transcripts was consistently missed. In general, the transcript-level detection rate increased with
328 transcript length, both for ONT and Illumina libraries (Fig. 5C). Counting the number of
329 "detected" transcripts and genes, defined as the number of features with an expected read
330 count of at least 1 with salmonminimap2 (ONT) or Salmon (Illumina), at various degrees of
331 subsampling (Fig. 5D-E) suggested that the current sequencing depth of approximately 0.5
332 million mapped ONT reads per library was not enough to detect all expressed genes or
333 transcripts. Furthermore, the number of observed genes were similar to the number observed in
334 the Illumina libraries if these were subsampled to comparable sequencing depths. With the aim
335 of investigating whether there are systematic 'blind spots' in the detection of features in the ONT
336 data (in which case we expect the same set of transcripts to be detected in all libraries) or if the
337 lack of saturation is purely a result of undersampling (in which case we would expect differences
338 in the set of detected transcripts across libraries), we compared the saturation curves obtained
339 from individual samples to that obtained by first pooling the reads across all replicates within a
340 data set, and subsequently sampling from this pool (Supplementary Fig. 11). On the transcript
341 level, pooling the samples improved the degree of saturation for a given number of reads, while
342 no improvement could be seen on the gene level.

343
344 Next, we calculated the correlation between abundance estimates among replicates of the HAP
345 cell line, within and between data sets. As expected, the correlation between replicates was

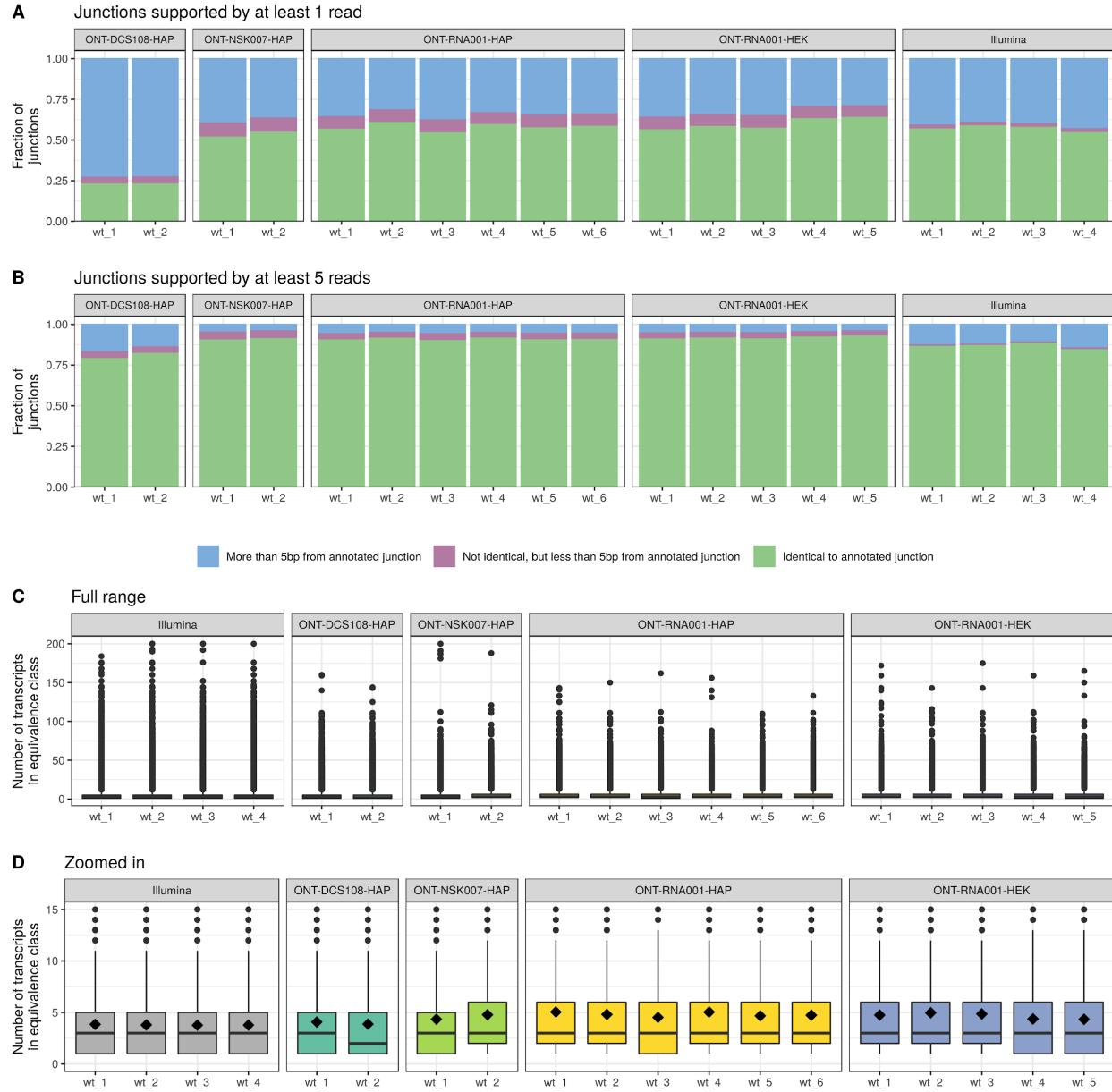
346 higher on the gene level than on the transcript level, and higher within a data set than between
347 data sets (Supplementary Fig. 12). On the gene level, correlation between replicates was
348 almost as high in the ONT data as in the Illumina data, for all quantification methods, while for
349 transcript-level abundances, higher correlations were observed in the Illumina data. Overall,
350 Wub showed the highest correlation of abundance estimates between replicates in the ONT
351 data sets. Notably, correlations between cDNA and native RNA samples were as high as those
352 among samples obtained with different cDNA protocols.

353

354 Comparing the abundance estimates obtained for the same library with different quantification
355 methods showed that, perhaps unsurprisingly, Salmon in quasi-mapping mode and
356 salmonminimap2 had the highest correlation (Supplementary Fig. 13). Stratifying transcripts and
357 genes by the annotated biotype suggested that certain biotypes (in particular, short transcripts
358 such as miRNAs) were consistently assigned very low abundances with ONT, while they were
359 observed in the Illumina libraries (Supplementary Fig. 14).

360 Transcript identifiability

361 Next, we focused on specific transcriptomic features that are useful for discriminating similar
362 isoforms. First, we extracted the junctions observed after aligning the ONT reads to the
363 genome. The majority of the junctions that were covered by at least 5 ONT reads were already
364 annotated in the reference transcriptome, while this was more rarely the case for lowly-covered
365 junctions (Fig. 6A-B). Junctions that were observed in the ONT reads but did not correspond to
366 annotated junctions were less likely than those already annotated to be observed in the Illumina
367 data, and also less likely to harbor a canonical splice junction motif (GT-AG) (Supplementary
368 Fig. 15). Not surprisingly, individual ONT reads generally spanned more junctions than Illumina
369 reads (Supplementary Fig. 16), which should provide improved ability of correct transcript
370 identification.



371

372 **Figure 6. A-B.** Annotation status of junctions observed in each ONT and Illumina library. A junction is
373 considered observed if it is supported by at least 1 (**A**) or 5 (**B**) reads. For each observed junction, the
374 distance to each annotated junction was defined as the absolute difference between the start positions
375 plus the absolute difference between the end positions. This distance was used to find the closest
376 annotated junction. **C.** Distribution of the number of transcripts contained in the Salmon equivalence class
377 that a read is assigned to, across all reads, for each ONT and Illumina library. **D.** As **C**, but zoomed in to
378 the range [0, 15]. The black diamond shape indicates the mean.

379

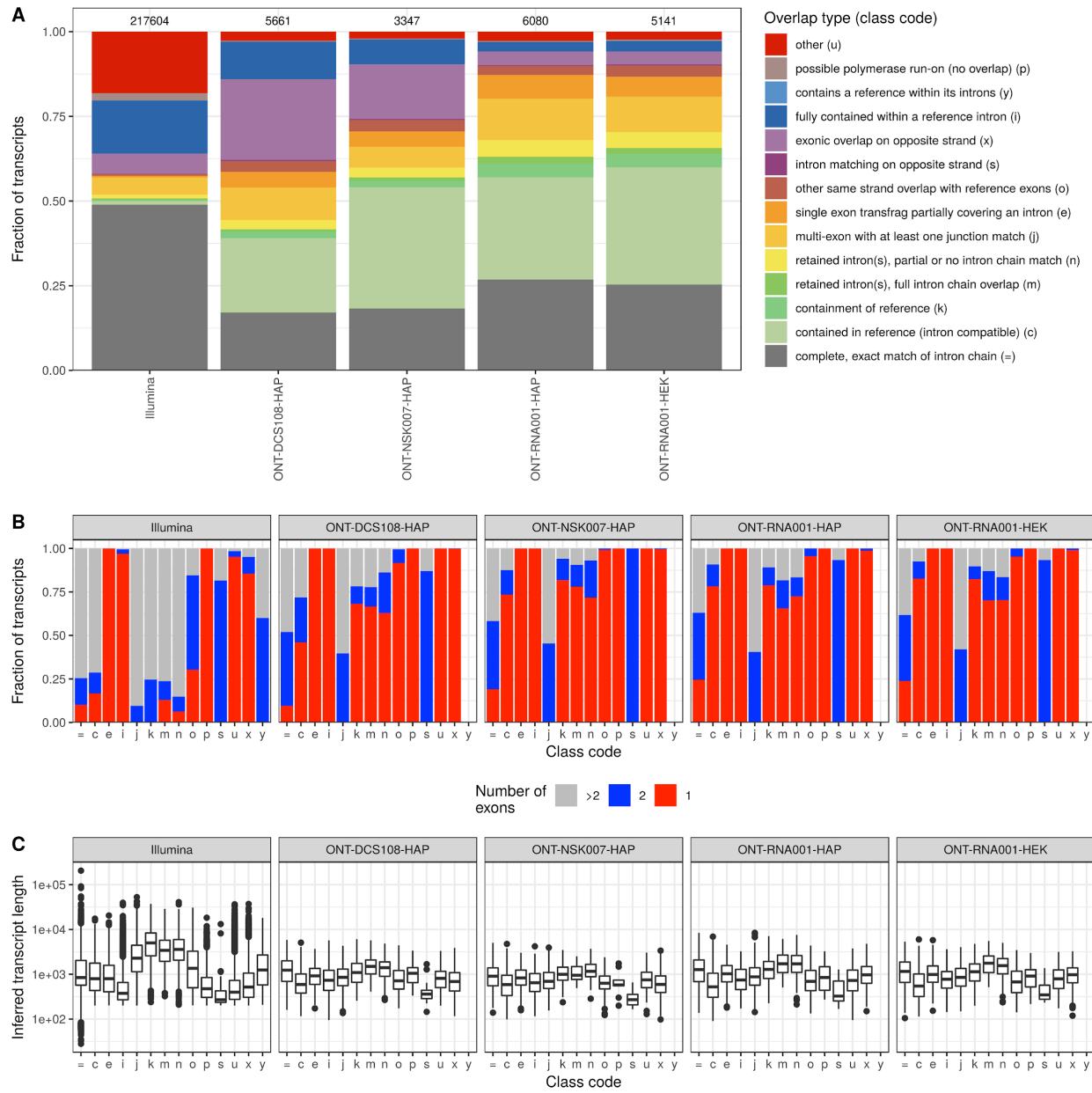
380 In order to further investigate if the longer length of ONT reads compared to Illumina reads in
381 fact improved their unambiguous assignment to specific transcripts, we tabulated the number of
382 transcripts included in the equivalence class that each read was assigned to when running
383 Salmon in quasi-mapping mode. A read being assigned to a large equivalence class indicates

384 that the read sequence is compatible with many annotated transcripts, and consequently that
385 unambiguous assignment is difficult. While fewer ONT reads were assigned to equivalence
386 classes with a very large number of transcripts compared to the Illumina counterparts, the
387 average number of transcripts in the equivalence class, across all reads, was almost identical
388 for the ONT and Illumina libraries (Fig. 6C-D). To investigate to what extent this was an effect of
389 the high redundancy among the annotated transcripts, we ran Salmon with the same index, but
390 using the annotated transcript catalog as a proxy for error-free, full-length ‘reads’. In this case,
391 87% of the reads were assigned to equivalence classes containing a single transcript. This
392 illustrates both that even in this idealized situation, not all reads would be unambiguously
393 assignable to a single annotated transcript, due to redundancies in the annotation catalog, and
394 that for our ONT reads, the ambiguity is still considerably higher than in the ideal situation.
395 Together with the large number of secondary transcriptome alignments observed above, this
396 illustrates the challenging nature of reference-based transcript identification based on ONT
397 reads. Furthermore, the read generation model used by Salmon is adapted to Illumina reads,
398 and thus is likely suboptimal for inferring transcript abundances from ONT reads.

399 Reference-free transcript identification

400 In addition to the reference-based transcript identification and quantification discussed above,
401 we generated a set of high-confidence consensus transcripts for each ONT data set using
402 FLAIR (<https://github.com/BrooksLabUCSC/flair>). For this analysis, only reads with a 5' end
403 close to a known promoter region were considered, and only transcript sequences supported by
404 at least 3 ONT reads were retained. The identified transcripts from FLAIR were compared to the
405 annotated reference transcriptome using gffcompare
406 (<https://ccb.jhu.edu/software/stringtie/gffcompare.shtml>). This comparison identified the most
407 similar reference transcript for each FLAIR transcript that showed at least some overlap with the
408 reference transcriptome, and further assigned a class code describing the type of relationship to
409 this most similar reference transcript (see <https://ccb.jhu.edu/software/stringtie/gffcompare.shtml>
410 for a description of all class codes). Interestingly, only a relatively low fraction of the identified
411 transcripts in each data set contained a junction chain that was identical to that of an annotated
412 transcript (Fig. 7A, class code ‘=’), while a larger fraction of the identified transcripts contained a
413 junction chain that was consistent with an annotated transcript, but only contained a subset of
414 the junctions. This corroborates the previous observations that many ONT reads may not
415 represent full-length transcript sequences. There is a marked difference compared to the set of
416 transcripts assembled with StringTie from the Illumina samples, a larger fraction of which

417 contain a complete intron chain match with an annotated transcript. There is also a larger
418 fraction of Illumina-derived transcripts that do not overlap known transcripts (Fig. 7A, class code
419 ‘u’). FLAIR transcripts with a junction chain perfectly matching an annotated transcript (class
420 code ‘=’) spanned a range of lengths and number of junctions (Fig. 7B-C), suggesting that
421 transcript identification is not limited to, e.g., short isoforms. Overall, the set of transcripts
422 assembled by StringTie from the Illumina data were more often multi-exonic than those from the
423 ONT libraries, and also spanned a broader range of transcript lengths. A random selection of
424 FLAIR transcript sequences (from the *ONT-RNA001-HAP* library) corresponding to annotated
425 transcripts are shown in Supplementary Fig. 17, to illustrate the variety of transcripts that could
426 be identified.



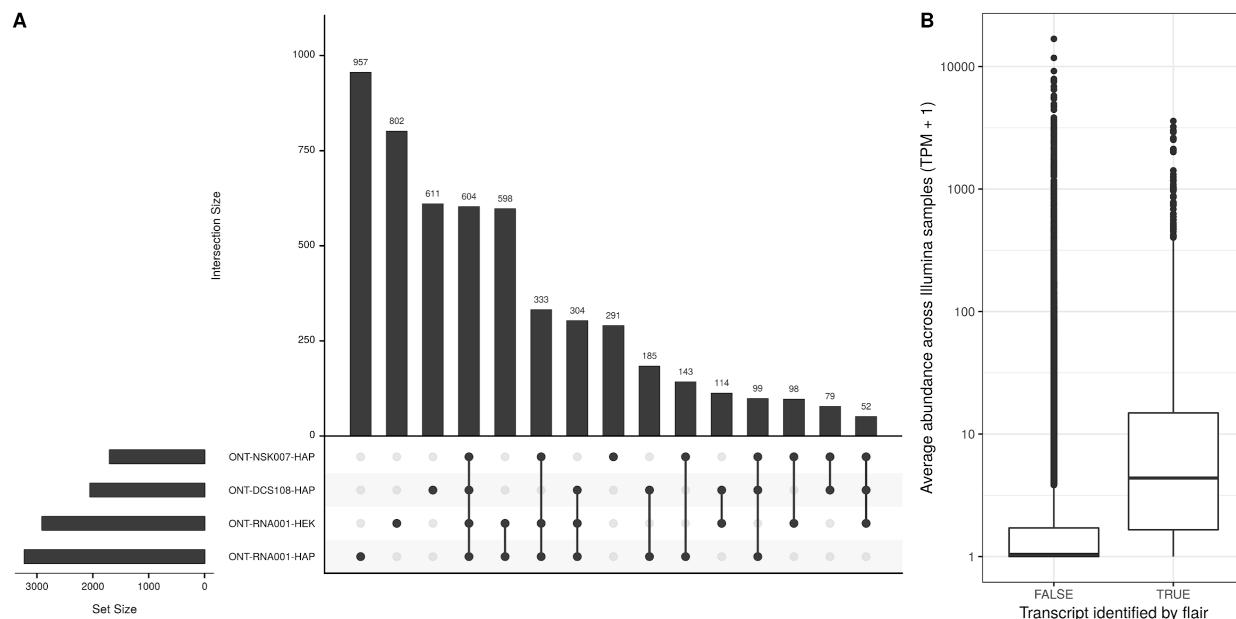
427

428 **Figure 7.** Characterization of transcripts identified by FLAIR. **A.** Class code distribution for de novo
429 identified transcripts from FLAIR (for ONT libraries) or StringTie (for Illumina libraries), compared to the
430 set of annotated transcripts using gffcompare. The number above each bar represents the number of
431 assembled transcripts. The class code for a transcript indicates its relation to the closest annotated
432 transcript. **B.** Number of exons in each transcript identified by FLAIR/StringTie, stratified by the relation to
433 the annotated transcripts (represented by the assigned class code). **C.** Length distribution of transcripts
434 identified by FLAIR/StringTie, stratified by the relation to the annotated transcripts (represented by the
435 assigned class code).

436

437 Comparing the set of annotated reference transcripts that could be identified by at least one
438 FLAIR transcript (class code '=' or 'c') in the respective ONT data sets showed that a large
439 fraction of these transcripts were only identified in a single data set (Fig. 8A). In addition,

440 reference transcripts identified by the native RNA-sequencing protocol in the two different cell
441 lines showed a higher degree of similarity to each other than to those identified with the cDNA
442 protocols in the HAP cell line, suggesting that transcript identification can be strongly affected by
443 the library preparation protocol. Of note, the native RNA protocols provide information about the
444 strandedness of the reads, which is not the case for the cDNA protocols employed here.
445 Reference transcripts with junction chains corresponding to at least one FLAIR transcript
446 generally showed a higher expression level in the Illumina samples than the reference
447 transcripts that were not identified in any ONT data set (Fig. 8B), suggesting that one possible
448 explanation for the discrepancy between the transcripts identified in the different ONT data sets
449 could be the limited sequencing depth, and that a larger number of ONT reads may be
450 necessary to identify a stable set of expressed transcripts.



451
452 **Figure 8.** Comparison of annotated transcripts identified by FLAIR in the four ONT data sets. **A.** UpSet
453 plot representing overlaps between the annotated transcripts that are identified by FLAIR in the different
454 ONT data sets. An annotated transcript is considered to be identified if at least one FLAIR transcript is
455 assigned to it with a class code of either '=' or 'c'. These sets of annotated transcripts are then compared
456 between data sets. Horizontal bars indicate the total number of identified annotated transcripts in the
457 respective data sets, and vertical bars represent the size of each intersection of one or more sets of
458 identified transcripts. **B.** Average abundance across the Illumina samples, for annotated transcripts that
459 are considered 'identified' or not by FLAIR. An annotated transcript is considered to be identified if at least
460 one FLAIR transcript from at least one data set is assigned to it with a class code of either '=' or 'c'.
461

462 Discussion

463 We have performed a detailed evaluation of reads from Nanopore native RNA sequencing as
464 well as complementary direct cDNA sequencing, from the perspective of transcript identification
465 and quantification. The libraries were prepared from human cell lines, which adds a level of
466 complexity compared to many previous studies focusing on either less complex model
467 organisms or synthetic transcripts. In addition, matched Illumina data was generated for
468 comparison.

469

470 We observed that despite the fact that ONT reads are around an order of magnitude longer than
471 typical Illumina reads, identification of their transcript of origin is still highly nontrivial, and a large
472 number of secondary transcriptome alignments with mapping scores very close to the primary
473 alignments were observed for all libraries. This suggests that quantification methods that focus
474 exclusively on the reported primary alignment are likely to be suboptimal, and can be highly
475 biased depending on how the primary alignment is selected among a set of equally-good
476 mappings. We expect that reference-based transcript abundance estimation methods that are
477 able to incorporate information about these multi-mapping reads are more likely to produce
478 reliable abundance estimates; however, to our knowledge no ONT-specific such method, with a
479 read generation model adapted to the ONT library generation, currently exists.

480

481 *De novo* as well as reference-based identification of transcripts suggested that a considerable
482 number of the raw ONT reads are likely to not represent full-length reference transcripts. This
483 can have implications for transcript identification and quantification. For example, it is difficult to
484 determine whether a truly truncated version of a reference transcript is present in a sample, or if
485 the reads rather are fragments of a longer transcript molecule. In addition, by attempting to
486 mitigate this issue, e.g. by filtering the ONT reads to only retain those that overlap a known
487 promoter region, the quantitative nature of the data, as well as the number of usable reads, may
488 be reduced. While our manuscript was in preparation for submission, a preprint authored by
489 Workman and colleagues¹⁷ was published that highlighted some of key benefits of Nanopore
490 native RNA sequencing, but also indeed reported the very frequent presence of truncated
491 Nanopore reads in native RNA libraries. They were also able to estimate that a significant
492 proportion of transcripts may be truncated by nanopore signal noise, caused for example by
493 electrical signals associated with motor enzyme stalls or by otherwise stray current spikes of
494 unknown origin. These surprising findings are supported by our observations that single native

495 RNA reads frequently fail to cover the full length of transcripts. We also agree that nanopore
496 native RNA read truncation is unlikely due to some fundamental limitation of nanopore-based
497 sequencing, especially considering that ONT 1D genomic DNA sequence reads of several
498 kilobases are consistently achieved without issue using the current pore-type^{18–20} used to
499 sequence both DNA and RNA. Further, such problems could conceivably be addressed, to at
500 least some extent, by training basecallers to reliably recognize relevant nanopore signal noise
501 events which might cause single molecule sequence reads to be truncated or split.

502
503 An inability to read approximately 10-15 nucleotides at the 5' end of each strand, and relatively
504 higher error rates, were identified as the two principal drawbacks of Nanopore native RNA
505 sequencing by the Workman et al study, although these are potentially readily addressable¹⁷.
506 Here we highlight that the sequencing depths achieved from native RNA libraries, typically
507 ~0.5M aligned reads per flow cell, are likely not enough to saturate transcript detection, either
508 using reference-based or *de novo* approaches. Further, our attempts at relevant differential
509 expression analyses from native RNA sequencing data during a parallel study suffered from low
510 power and high variability (data not shown), most likely due to the limited coverage within each
511 library replicate. Improving throughput (the amount of sequence rendered per unit cost and unit
512 time) is a critical issue: if ONT sequencing throughput remains low, uptake and thus impact
513 within the transcriptomics field will likely remain limited, even given its distinguished benefits.
514 Although protein-pore sequencing can be scaled to considerably higher levels (i.e. either on the
515 ONT GridION or PromethION instruments), the associated consumable nanopore array costs
516 remain high. Thus, native RNA-seq throughput characteristics that are deemed acceptable by
517 the transcriptomics community at large will likely require a highly-optimized RNA motor enzyme,
518 or ultimately a shift to a lower cost nanopore array type. When characterization of complex
519 transcriptomes at transcript-level comprises the project remit, our study here describes that
520 Nanopore direct RNA-seq remains a roundly promising but fledgling analysis tool.

521 **Methods**

522 **Cell lines and culture**

523 HEK293 cells (ATCC) were cultured in Dulbecco's Modified Eagles Medium (DMEM)
524 supplemented with 10% FBS and penicillin/streptomycin. HAP1 cells (Horizon Discovery) were
525 grown in Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 10% FBS and

526 penicillin/streptomycin. All cultures were maintained at a temperature of 37°C in a humidified
527 incubator with 5% CO₂. When required, exponentially growing cells were harvested by washing
528 in Phosphate Buffered Saline (PBS) and then incubating with Trypsin-EDTA, followed by further
529 washing of pelleted cells in PBS.

530 Library preparation and sequencing

531 For the Nanopore libraries, total RNA was extracted from cell pellets using Trizol, and the
532 polyA+ fraction isolated using oligodT dynabeads (Invitrogen). The ONT kits NSK007, DCS108,
533 and RNA001 were then used for PCR-free 1D library preparations. For RNA001, 500ng of input
534 polyA+ RNA was used per sample and the libraries were made following ONT instructions. For
535 DCS108, 100ng of input polyA+ RNA was used per sample and the libraries were prepared
536 according to ONT instructions. For NSK007, 100 ng of input polyA+ RNA was used per sample
537 and libraries were made according to ONT instructions, except that the hairpin adaptor (HPA)
538 ligation and PCR steps were omitted as described previously ¹⁸, in order to enable 1D and direct
539 cDNA sequencing respectively. The prepared libraries were sequenced on the MinION using
540 R9.4 flow cells with the relevant MinKNOW script to generate fast5 files. All generated fast5
541 reads were then basecalled in Albacore (version 1.2 for NSK007 libraries and version 2.1 for
542 DCS108 and RNA001 libraries) using the relevant script to yield fastq files. As Albacore only
543 contained a 2D script for NSK007 basecalling, only the generated NSK007 fastq 'raw' reads (i.e.
544 complement and template) were taken forward for analysis, while any attempted 'consensus'
545 reads present were discarded.

546
547 For the Illumina samples, all libraries were made using the Illumina TruSeq stranded mRNA kit.
548 The mRNA libraries were prepared from 500 ng of Trizol-extracted total RNA using the Illumina
549 TruSeq® Stranded mRNA Sample Preparation Kit with 15 PCR cycles applied. Libraries were
550 quantified and quality checked using qPCR with Illumina adapter specific primers and Agilent
551 2200 TapeStation, respectively. Diluted indexed mRNA-seq (10nM) libraries were pooled, used
552 for cluster generation (Illumina TruSeq PE Cluster Kit v4-cBot-HS) and sequenced [Illumina
553 HiSeq 4000, Illumina TruSeq SBS Kit v4-HS reagents, paired-end approach (2x150bp) with 40-
554 55 million reads per sample].

555 Genome and transcriptome alignment

556 ONT reads were aligned to the human genome (Ensembl primary assembly GRCh38) and
557 transcriptome (combined cDNA and ncRNA reference fasta files from Ensembl GRCh38.90)
558 using minimap2 v2.12²¹. The genome alignments were performed with the arguments –ax
559 splice -N 10, to allow spliced alignments and up to 10 secondary alignments per read.
560 Alignment files from minimap2 were converted to bam format, sorted and indexed using
561 samtools v1.6²². The Bioconductor package GenomicAlignments (v1.32.0)²³ was used to
562 extract junctions from the alignments. For each observed junction, we calculated the distance
563 (the absolute difference between the start positions plus the absolute difference between the
564 end positions) to the closest annotated junction. For the transcriptome alignment, we used the
565 arguments –ax map-ont -N 100 to allow more secondary alignments, given the high
566 similarity among transcript isoforms. The minimap2 –p argument, representing the minimal ratio
567 of the secondary to primary alignment score that is allowed in order to report the secondary
568 mapping, has a default value of 0.8. For transcriptome alignment, we investigated the effect of
569 increasing this value in order to restrict the number of reported “suboptimal” secondary
570 alignments. To evaluate the alignments, we recorded the alignment rates, defined as the
571 fraction of reads with a reported primary alignment, as well as the aligned fraction of each read,
572 which we defined as the sum of the number of “M” and “I” characters in the CIGAR string,
573 divided by the full length of the read. For some reads, minimap2 also reported supplementary
574 alignments. For each supplementary genome alignment, we compared the alignment position to
575 that of the corresponding primary alignment, and recorded whether these were on the same or
576 different chromosome and/or strand, and whether the primary and supplementary alignments
577 overlapped each other. Finally, we generated reduced FASTQ files by retaining only reads with
578 a primary alignment to the genome, and for each such read, we removed all bases that were
579 (soft-)clipped in the primary alignment. The resulting bam files were converted to FASTQ format
580 using bedtools bamtofastq v2.27.0²⁴, and the reads were subsequently shuffled using bbmap
581 v38.02 (<https://sourceforge.net/projects/bbmap/>). RSeQC v2.6.5²⁵ was used to examine the
582 coverage profile along gene bodies for each library, based on the GENCODE basic v24 bed file
583 downloaded from https://sourceforge.net/projects/rseqc/files/BED/Human_Homo_sapiens/ on
584 October 23, 2018.

585 Gene and transcript abundance estimation

586 Four different computational methods were used to estimate transcript and gene abundances
587 for the ONT libraries. First, we applied Salmon v0.11.0²⁶ in quasi-mapping mode, with an index
588 generated from the combined Ensembl cDNA and ncRNA reference fasta files and using the
589 default k value of 31 (denoted **salmon31** below). For comparability across pipelines, we
590 retained any duplicate transcripts in the index generation. The mean and maximal fragment
591 lengths were set to 600 and 230,000, respectively, and the flag `--dumpEq` was set to retain
592 equivalence class information. Salmon was also run in quasi-mapping mode on the modified
593 FASTQ files, containing only the aligned part of the primary alignments as described above.
594 Second, we applied Salmon in alignment-based mode to the output bam files from the
595 minimap2 transcriptome alignment, using the flag `--noErrorModel` to disable the default
596 short-read error model of Salmon in the quantification (denoted **salmonminimap2**). Third, we
597 applied the `bam_count_reads.py` script from the Wub package
[\(https://github.com/nanoporetech/wub\)](https://github.com/nanoporetech/wub) to the output files from the transcriptome alignment,
599 setting the minimal mapping quality (`-a` argument) to 5 (denoted **wubminimap2**). Finally, we
600 applied featureCounts (from subread v1.6.0)^{27,28} to the primary genome alignments, requiring a
601 minimum overlap of 10 bases and using the `-L` argument to enable the long-read mode
602 (denoted **fCminimap2primary**). While the Salmon variants and Wub provided transcript-level
603 abundance estimates, which were also aggregated to the gene level, featureCounts provided
604 only gene-level counts and was therefore not considered for transcript quantification.

605 De novo transcript identification

606 In addition to the reference-based quantification described above, we also performed reference-
607 free, *de novo* transcript identification using FLAIR (obtained from
608 <https://github.com/BrooksLabUCSC/flair> on December 16, 2018), applied to the combined
609 primary genome alignments from all libraries in each ONT data set. The minimap2 bam files
610 were converted to bed format using the `bam2bed12.py` script provided with FLAIR, and
611 identified junctions were subsequently corrected by comparison to the reference annotation,
612 using the default window size of 10. Next, the corrected reads were collapsed using FLAIR,
613 requiring that the 5' end of the read falls close to a promoter and retaining only transcripts
614 represented by at least 3 reads. The promoter bed file was obtained by combining active, weak
615 and poised promoters identified in nine cell lines by the ENCODE consortium (obtained from
616 <https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm> and lifted over to

617 hg38 coordinates using the UCSC Genome Browser liftOver tool) The identified transcripts from
618 each data set were compared to the annotated transcripts using gffcompare
619 (<https://ccb.jhu.edu/software/stringtie/gffcompare.shtml>), whereby each FLAIR transcript was
620 assigned a *class code*, detailing the way in which it is related to the most similar reference
621 transcript.

622 Processing of Illumina libraries

623 Sequencing adapters were removed from the Illumina libraries with TrimGalore! v0.4.4
624 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, using cutadapt v1.13²⁹), with
625 quality and length cutoffs both set to 20, and aligned to the Ensembl GRCh38.90 primary
626 genome assembly using STAR v2.5.1b³⁰. Abundances of annotated transcripts were estimated
627 using two different methods: first, with StringTie v1.3.3b³¹ using reads aligned with HISAT2
628 v2.1.0³² (with the --dta flag set and using a known splice site file), and second, with Salmon in
629 quasi-mapping mode, using the same index as for the ONT libraries, and including adjustments
630 for GC content and sequence bias. Abundances were read into R using tximport (v1.8.0)³³. In
631 addition, we used StringTie to assemble new transcripts (without the -e flag, provided with the
632 reference gtf file) for comparison with the transcripts identified by FLAIR from the ONT libraries.
633 For this analysis, we merged the HISAT2 bam files from all four Illumina samples to use as the
634 input for StringTie. We used the default coverage cutoff of 2.5 to determine which assembled
635 transcripts to retain in the output file.

636 Public data

637 In addition to the ONT and Illumina data generated in-house, we processed the SIRV E0 (SRA
638 accession number SRR6058584) and ERCC Mix1 (SRA accession number SRR6058582) ONT
639 dRNA libraries from Garalde *et al.*¹⁵. The reads were aligned to the respective transcriptomes
640 using minimap2 with the same settings as above. The SIRV data set was also aligned to the
641 corresponding genome using minimap2 with the settings described above, and additionally
642 setting --splice-flank=no to accommodate the non-canonical splice sites present in this
643 data.

644 Acknowledgements

645 The authors would like to thank Botond Sipos, Michael Stadler and Giovanni d'Ario for
646 constructive discussions and feedback on the manuscript. Research in the S.H. laboratory is
647 funded by the Biotechnology and Biosciences Research Council, UK. C.S. was supported by a
648 Pilot Project grant from the University Research Priority Program Evolution in Action of the
649 University of Zurich. M.D.R. acknowledges support from the University Research Priority
650 Program Evolution in Action at the University of Zurich and the Swiss National Science
651 Foundation (310030_175841).

652 Author contributions

653 C.S.: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology,
654 Software, Visualization, Writing - original draft, Writing - review & editing. Y.Y.: Formal analysis.
655 A.B-N.: Methodology, Writing - review & editing. A.P.: Methodology. M.D.R.: Conceptualization,
656 Data curation, Formal analysis, Funding acquisition, Methodology, Writing - review & editing.
657 S.H.: Conceptualization, Funding acquisition, Investigation, Methodology, Writing - original draft,
658 Writing - review & editing

659 Competing interests

660 The authors declare that they have no competing interests.

661 Data availability

662 The raw sequence files have been uploaded to ArrayExpress under accession numbers E-
663 MTAB-7757 (Illumina) and E-MTAB-7778 (ONT).

664 Code availability

665 The code used to perform the analyses in the paper is available on GitHub:
666 <https://github.com/csoneson/NativeRNAseqComplexTranscriptome>.

667 References

- 668 1. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon

- 669 definition and function. *Nat. Rev. Genet.* **11**, 345–355 (2010).
- 670 2. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative
671 splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*
672 **40**, 1413 (2008).
- 673 3. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature*
674 **456**, 470–476 (2008).
- 675 4. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human
676 transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2011).
- 677 5. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through
678 the lens of local splicing variations. *Elife* **5**, e11752 (2016).
- 679 6. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the
680 human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
- 681 7. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation
682 among the surface receptors of individual B cells. *Nat. Commun.* **8**, ncomms16027 (2017).
- 683 8. Seki, M. *et al.* Evaluation and application of RNA-Seq by MinION. *DNA Res.* (2018).
684 doi:10.1093/dnares/dsy038
- 685 9. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J.
686 Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative
687 assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).
- 688 10. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford
689 Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100
690 (2017).
- 691 11. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and
692 single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9869–9874
693 (2014).
- 694 12. Gonzalez-Garay, M. L. Introduction to Isoform Sequencing Using Pacific Biosciences

- 695 Technology (Iso-Seq). in *Transcriptomics and Gene Regulation* (ed. Wu, J.) 141–160
696 (Springer Netherlands, 2016).
- 697 13. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing
698 libraries. *Genome Biol.* **12**, R18 (2011).
- 699 14. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat.*
700 *Methods* **10**, 1177 (2013).
- 701 15. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat.*
702 *Methods* (2018). doi:10.1038/nmeth.4577
- 703 16. Chenchik, A. *et al.* RT-PCR Methods for Gene Cloning and Analysis. (1998).
- 704 17. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome.
705 *bioRxiv* 459529 (2018). doi:10.1101/459529
- 706 18. Carter, J.-M. & Hussain, S. Robust long-read native DNA sequencing using the ONT CsgG
707 Nanopore system. *Wellcome Open Res* **2**, 23 (2017).
- 708 19. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long
709 reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
- 710 20. Tyson, J. R. *et al.* MinION-based long-read sequencing and assembly extends the
711 *Caenorhabditis elegans* reference genome. *Genome Res.* **28**, 266–274 (2018).
- 712 21. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (2018).
713 doi:10.1093/bioinformatics/bty191
- 714 22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
715 2079 (2009).
- 716 23. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput.*
717 *Biol.* **9**, e1003118 (2013).
- 718 24. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
719 features. *Bioinformatics* **26**, 841–842 (2010).
- 720 25. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments.

- 721 *Bioinformatics* **28**, 2184–2185 (2012).
- 722 26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and
723 bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
- 724 27. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read
725 mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
- 726 28. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
727 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- 728 29. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
729 *EMBnet.journal* **17**, 10–12 (2011).
- 730 30. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- 731 31. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-
732 seq reads. *Nat. Biotechnol.* **33**, (2015).
- 733 32. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
734 requirements. *Nat. Methods* **12**, 357 (2015).
- 735 33. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-
736 level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).