Two-pass alignment using machine-learning-filtered splice junctions increases the accuracy of intron detection in long-read RNA sequencing

Matthew T. Parker[1]*, Geoffrey J. Barton[1], Gordon G. Simpson[1,2]*

[1]School of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH, UK

[2]James Hutton Institute, Invergowrie, DD2 5DA, UK

*Correspondence: m.t.parker@dundee.ac.uk; g.g.simpson@dundee.ac.uk

Abstract

Transcription of eukaryotic genomes involves complex alternative processing of RNAs. Sequencing of full-length RNAs using long-reads reveals the true complexity of processing, however the relatively high error rates of long-read technologies can reduce the accuracy of intron identification. Here we present a two-pass approach, combining alignment metrics and machine-learning-derived sequence information to filter spurious examples from splice junctions identified in long-read alignments. The remaining junctions are then used to guide realignment. This method, available in the software package 2passtools (https://github.com/bartongroup/2passtools), improves the accuracy of spliced alignment and transcriptome annotation without requiring orthogonal information from short read RNAseq or existing annotations.

Background

Understanding eukaryotic genomes requires knowing not only the DNA sequence but also which RNAs are transcribed from it. Eukaryotic transcription by DNA-dependent RNA polymerase II is associated with multiple alternative RNA processing events that diversify the coding and regulatory potential of the genome. Alternative processing choices include distinct transcription start sites, the alternative splicing of different intron and exon combinations, alternative sites of cleavage and polyadenylation, and base modifications such as methylation of adenosine. Patterns of alternative processing can be extensive. For example, more than 90% of human protein-coding genes have at least two splice isoforms[1]. Changes in RNA processing can reflect the reprogramming of gene expression patterns during development or in response to stress or result from genetic mutation or disease. Consequently, the identification and quantification of different RNA processing events is crucial to understand not only what genomes encode but also the biology of whole organisms[2].

The sequencing of RNAs (RNAseq) can reveal gene expression patterns in specific cells, tissues or whole organism. The success of this approach depends upon sequencing methodology and the computational analyses used in interpreting the sequence data. High-throughput sequencing of RNA rarely involves direct RNA sequencing (DRS): instead, copies of complementary DNA (cDNA) produced by reverse transcription of RNA molecules are sequenced[2]. However, template strand switching by reverse transcriptase (RT) during the copying process can produce spurious splicing patterns and antisense RNA signals[3,4]. Three current technologies use RT-based RNA sequencing library preparation: Illumina, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Illumina RNAseq can generate hundreds of millions of highly accurate short sequencing reads, each representing a 50–250 nt fragment of full-length RNA[2]. Methods exist for quantifying known alternative splicing events from short reads[5]. However, when the transcript models are unknown, for example in a non-model organism or a mutant or disease with altered RNA processing, new transcript models must be generated, either *de novo* or with the aid of the reference genome. Because Illumina reads are short, they are unlikely to overlap multiple splice junctions, meaning that phasing of splicing events is difficult and requires complex computational reconstruction[6-8]. PacBio and ONT can sequence full-length cDNA copies without fragmentation, thus allowing whole transcript isoforms to be identified unambiguously[2]. Most recently, ONT introduced a direct sequencing method for RNA[9-11]. Using this approach, it is now possible to capture information on the splicing, 5’ and 3’ ends, poly(A) tail length, and RNA modifications of full-length RNA molecules in a single experiment, without RT-associated artefacts[11].

54 The development of technologies for sequencing full-length RNA molecules makes the identification of
55 authentic processing events possible in principle, but software tools are also needed to interpret the
56 RNA processing complexity. PacBio and ONT sequencing reads have a higher error rate than Illumina[10-14].
57 Consequently, alignment accuracy for long sequence reads at splice junctions is often compromised[9-11].
58 This is a problem for genome-guided transcriptome annotation because the incorrect identification of
59 splice junctions leads to mis-annotated open reading frames and incorrectly truncated protein
60 predictions. In addition, if alignment errors are systematic (i.e. occur for transcripts with specific
61 characteristics), then quantification of transcripts will be compromised. Even with completely error-free
62 reads, alignment at splice junctions is often confounded by multiple equally plausible alternatives[15].
63 Accordingly, computational methods for improving the splice-aware alignment of long reads are
64 required.

65 Software tools for long and short RNAseq data analysis incorporate several approaches to address the
66 challenges presented by pre-mRNA splicing. Biologically relevant information can aid the alignment of
67 transcriptomic sequences to the genome. For example, the vast majority of eukaryotic splicing events
68 occur at introns bordered by GU and AG motifs. Making RNAseq read aligners aware of these sequence
69 features (as is the case for the commonly used spliced aligners STAR[16], HISAT2[17] and minimap2[18]) can
70 significantly improve the alignment of reads at splice junctions. In addition, where genome and
71 transcriptome annotations exist, many alignment tools allow users to provide sets of correct splice
72 junctions to guide alignment[16-19]. Introns containing these guide splice junctions are penalised less than
73 novel introns, resulting in fewer alignment errors. For long reads, software tools such as FLAIR[10] use
74 post-alignment correction to improve splice junction detection and quantification. Post-alignment
75 correction tools take long-read alignments and guide splice junctions from either a reference annotation
76 or set of accurate short RNAseq reads[10]. Introns from long-read alignments which are not supported by
77 the guide splice junction set are "corrected" to the nearest supported junction within a user-defined
78 range. It is unclear whether such post-alignment corrections confer any benefit over providing guide
79 splice junctions during alignment.

80 Two-pass alignment has also been used to improve splice junction detection and quantification[16,19,20]. In
81 a two-pass alignment approach, splice junctions detected in a first round of alignment are scored less
82 negatively in a second round, thereby allowing information sharing between alignments. This approach
83 has been useful for short-read data, where RNA fragmentation may occur close to splice junctions
84 during sequencing library preparation. The two-pass approach enables these short junction overhangs

85 to be aligned to splice junctions detected in other alignments[20]. Splice junctions detected in a first pass
86 may also be filtered to remove false positives before second-pass alignment. Existing tools for splice
87 junction filtering, such as finesplice and portcullis[21,22], use machine learning with training on a range of
88 junction metrics. A model is trained from high-confidence positive and negative examples from training
89 data and then applied to classify the remaining splice junctions at the decision boundary. Splice
90 junctions are then filtered to remove junctions predicted to be spurious. Subsequent second-pass
91 alignment guided by these filtered junctions can then improve the accuracy of alignment[22].

92 In this study, we develop a method for filtered two-pass alignment of the relatively high-error long reads
93 generated by techniques such as nanopore DRS. The resulting software, which we have named
94 2passtools, uses a rule-based approach to identify probable genuine and spurious splice junctions from
95 first-pass read alignments. These can then be used to train a logistic regression (LR) model to identify
96 the biological sequence signatures of genuine splice junctions. We found that integrating the alignment
97 and sequence information extracted in this manner produced the largest improvement in splice junction
98 alignment and subsequent genome-guided annotation. As a result, we can improve the utility of long-
99 read sequencing technologies in revealing the complexity of RNA processing and annotating newly
100 sequenced organisms.

101

102 Results and Discussion

103 *Reference-splice-junction-aware alignment is more accurate than post-alignment junction correction*

104 For sequencing experiments designed to interpret RNA from model organisms, a set of reference splice
105 junctions will already be available (e.g. from Ensembl). We therefore asked how providing these
106 reference splice junctions to minimap2 to guide alignment performed compared with post-alignment
107 correction of junctions with FLAIR[10]. For this, we used four nanopore DRS datasets generated from
108 Arabidopsis seedlings[11] and four datasets generated from human cell lines[10]. Several types of probable
109 alignment error were identifiable in these data, including failure to align terminal exons and short
110 internal exons, spurious terminal exons, and large insertions to the reference genome (Fig. 1). Because
111 these datasets are likely to contain novel splice junctions which do not appear in reference annotations,
112 we simulated full-length reads (i.e. with no 3' bias[11]) using the Arabidopsis and human reference
113 transcriptomes, AtRTD2[23] and GRCh38[24], respectively. Simulated reads were then mapped to the
114 corresponding reference genome using minimap2[18], either with or without guidance from reference

splice junctions. Alignments of simulated reads were found to have similar error profiles to genuine nanopore DRS read alignments (Fig. S1A–C). Reads mapped without reference splice junctions were then corrected using FLAIR with reference splice junctions.

Although nanopore DRS has some systematic errors in base-calling (particularly at homopolymers), the majority of sequencing errors occur stochastically[25]. In contrast, we found that alignment errors were often repeated at similar locations in the alignments of independent reads from equivalent mRNA transcripts (Fig. 1, Fig. 2A). A common alignment error at splice junctions is failure of a short exon to align correctly: instead, fragments of the exon are aligned to the ends of flanking exons, resulting in a single incorrect intron. A clear example of such an alignment error was detected at the short (42 nt) exon 6 of Arabidopsis *FLM* (*AT1G77080*; Fig. 2A). Minimap2 uses a modified form of the Smith–Waterman algorithm for performing local alignment[18,26]. This method scores alignments using bonuses for matches to the reference sequence and penalties for mismatches or the opening of insertions, including introns. Incorrect alignment of *FLM* exon 6 is likely to occur because the bonus for aligning a short exon with sequencing errors is not sufficient to overcome the penalty for opening the two flanking introns[18]. Overall, we found that only 19.3% of simulated *FLM* reads aligned to the correct transcript isoform. Because the sequence distance between the alignment and the genuine reference splice junctions was so great, FLAIR was unable to perform post-alignment correction at *FLM* exon 6, resulting in the reporting of incorrect introns (Fig. 2A). In all, 40.3% of simulated *FLM* reads were aligned to the correct transcript isoform after FLAIR correction of splice junctions using the reference annotation. However, providing reference splice junctions to minimap2 during alignment resulted in the correct identification of *FLM* exons and introns in most cases: 92.1% of simulated *FLM* reads were aligned to the correct transcript isoform. We conclude that for loci with complex splicing patterns, reference-splice-junction-guided alignment performs better than post-alignment correction.

Without guidance from a reference annotation, we found that a median of 73.2% of Arabidopsis reads and 44.4% of human reads mapped correctly to the splice junctions of the transcript they were simulated from (Fig. 2B). The difference between the two organisms may be explained by biological differences between the two species (e.g. in intron size, number of exons per transcript, number of intronless transcripts). After post-alignment correction of splice junctions using FLAIR, the number of correctly identified transcripts detected was improved (median of 87.9% and 63.6% for Arabidopsis and human reads, respectively; Fig. 2B). This came at the cost of a small increase in alignment of reads to incorrect reference transcript splice junctions: from a median of 1.79% to 2.62% for Arabidopsis and

146 from 3.86% to 5.45% for human (Fig. S2A). This misclassification may affect the relative quantification of
147 transcripts for some genes, with implications for differential transcript usage analysis. Reference
148 annotation-informed alignment with minimap2 performed better than FLAIR, with a median of 93.8% of
149 Arabidopsis reads and 73.2% of human reads aligning correctly at the splice junctions of the transcript
150 they were simulated from (Fig. 2B), albeit with misclassification rates of 2.61% and 5.49% respectively
151 (Fig. S2A). We conclude that there is a clear benefit to providing reference splice junctions during
152 alignment of long reads with relatively high sequence error rates, and that this is preferable to post-
153 alignment correction.

154

155 *Alignment metrics enable identification of genuine splice junctions*

156 In newly sequenced organisms or conditions with novel RNA processing events, suitable reference
157 annotations to guide alignment may not be available. We therefore asked how the spliced alignment of
158 nanopore DRS reads might be improved in the absence of reference annotation. Naive two-pass
159 alignment has been successfully used to improve the spliced alignment of short reads[20]. We applied this
160 approach with our real and simulated nanopore DRS reads. Splice junctions identified by a first-pass
161 alignment of reads were selected and used (without filtering) to inform a second-pass alignment. We
162 found that using this simple approach the median percentage of simulated Arabidopsis DRS alignments
163 which matched the splice junctions of the reference transcript they were simulated from could be
164 increased slightly from 73.2% to 75.8% (Fig. S2B). The increase was similar for reads simulated from
165 human DRS alignments: from 44.4% to 47.3% (Fig. S2B).

166 We next considered whether further improvements in two-pass alignment could be obtained by filtering
167 out likely false-positive splice junctions from first-pass alignments. This would allow us to provide more
168 refined guide junctions for second-pass alignment (Fig. 3A). A similar approach worked for short reads
169 when splice junctions were filtered by using junction metrics to train a classifier in the portcullis
170 software tool[22]. By using the presence or absence of a splice junction in the reference annotation as a
171 ground truth, we considered a range of novel or previously introduced junction metrics[21,22], including
172 junction alignment distance, supporting read count, intron motif and the presence/absence of nearby
173 splice donor and acceptor sites with higher supporting read counts (Fig. S3A-D).

174 The junction alignment distance (JAD) is defined as the minimum distance to the first mismatch,
175 insertion or deletion on either overhang of a read alignment splice junction. This metric is used by both

176 finesplice and portcullis software tools[21,22]. For the simulated nanopore DRS read alignment datasets
177 sequenced from Arabidopsis RNA, we found that 88.9% of splice junctions found in the reference
178 annotation had at least one read alignment with a JAD of 4 nt, compared with only 10.1% of
179 unannotated splice junctions (Fig. 3B). Consequently, using a threshold of at least one read with a JAD of
180 4 nt, we could identify annotated splice junctions with an F1 score of 0.902 (Fig. S3A). Despite the high
181 probability of at least some genuine unannotated splice junctions in the real Arabidopsis data[11], we
182 found that the same JAD threshold could discriminate between annotated and unannotated splice
183 junctions in real datasets to a similar degree (F1 score = 0.899). Similar results were also seen for
184 simulated human datasets, where the same JAD threshold could discriminate between spurious
185 unannotated and genuine annotated splice junctions (F1 score = 0.868). We conclude that the JAD
186 metric is a powerful discriminator of genuine splice junctions across nanopore DRS datasets from
187 different organisms.

188 Of the other metrics we tested, the read count was predictive of genuine splice junctions at a threshold
189 of >1 read (F1 score = 0.833; Fig. S3B). However, read count correlated strongly with the JAD
190 (Spearman's $\rho = 0.776$), suggesting that it does not provide more information. The presence/absence of
191 a canonical intron motif (i.e. GU/AG, GC/AG or AU/AG) had a very high recall, as 99.96% of annotated
192 introns in the simulated alignments were canonical (Fig. S3C). However, the precision was poorer (F1
193 score = 0.783). This is because in spliced alignment mode minimap2 prefers GU/AG motifs, meaning that
194 67.1% of spurious splice junctions are also aligned so as to use canonical motifs.

195 Finally, we developed a primary donor/acceptor metric similar to the one used in portcullis[22]. This is
196 calculated by identifying alternative donor or acceptor sites in a 20 nt window around each
197 donor/acceptor and then determining whether they have greater read support than the current site. In
198 case of a tie for read support (e.g. if all splice junctions have a read count of 1), the JAD is used to break
199 the tie, i.e. sites with the largest maximum per-read JAD are considered most likely to be genuine and
200 labelled as a primary site. We found that the primary donor and acceptor metrics were also predictive of
201 genuine splice junctions (F1 scores = 0.842 and 0.785 respectively). By combining the metrics to select
202 splice junctions which are both primary donors and acceptors, the F1 score can be increased to 0.918
203 (Fig. S3D). It is unclear why the primary donor score is more predictive than the primary acceptor score.
204 A possible reason is that minimap2 is more likely to produce alignment errors at the donor site of splice
205 junctions (e.g. in the case of failure to align small internal exons) or that there are more genuine
206 alternative acceptor sites than donor sites.

207 We used the identified metrics to create a simple decision tree model. A five-node tree using the JAD,
208 primary donor/acceptor and canonical intron motif metrics (Fig. 3C) was best able to predict genuine
209 Arabidopsis splice junctions (F1 score = 0.935; Fig. 3D). The same decision tree also performed well in
210 predicting genuine and spurious splice junctions from simulated human reads (F1 score = 0.934). This
211 indicates that the model might generalise across nanopore DRS datasets from different organisms,
212 despite their differences in splicing complexity.

213

214 *A combination of splice junction alignment metrics and sequence information improves authentic splice*
215 *junction identification*

216 Genuine splice junctions have sequence biases which are defined by their interactions with spliceosomal
217 uridylate-rich small nuclear RNAs[27]. We next asked whether machine learning models could identify
218 genuine splice junctions from the flanking genomic sequences alone. For example, genome sequence
219 information might help identify genuine splice junctions with low read alignment coverage that fail to
220 pass the JAD filter due to stochastic sequencing errors. We therefore extracted 128 nt sequences
221 centred on unique donor and acceptor sites and used these to train LR models with labels generated by
222 the first decision tree model (Fig. 4A). Using 6-fold cross-validation, we were able to train six models on
223 83.3% of the data each and use them to make predictions for the remaining 16.7%. Using this approach,
224 we could generate predictions for all splice junctions, with no junction being used for both training and
225 prediction from the same model.

226 At a prediction threshold of 0.5, the model overclassified positive splice junctions. False positives may be
227 sequences which could in principle act as splice junctions but do not in reality due to effects that the
228 model cannot capture. One such effect could be the presence of alternative splice junctions which are
229 preferentially processed. This is thought to occur under the "first-come-first-served" model of co-
230 transcriptional splicing[28,29]. The model is also unlikely to be able to correctly identify the intron
231 branchpoint motif because this can vary in position relative to the acceptor site[30]. Nevertheless, we
232 found that the LR model approach could predict genuine splice junctions from sequence data alone with
233 comparable accuracy to the metric-based decision tree (Fig S4A-C). For example, for the simulated
234 Arabidopsis datasets, using LR on donor and acceptor sequences (with a prediction threshold of 0.5)
235 yielded an F1 score of 0.904 (Fig S4C), which was similar to the F1 score obtained with the JAD or
236 primary donor/acceptor metrics.

237 We next tested whether the information from the junction metrics and reference sequence model was
238 complementary, i.e. if a combination of the two approaches could produce an improvement in splice
239 junction prediction over each individual approach. Use of a second decision tree model, this time
240 including the JAD metric, primary donor/acceptor metrics and new LR prediction scores (Fig. 4B), further
241 increased the F1 score on splice junctions identified from simulated Arabidopsis read alignments to
242 0.954 (Fig. 4C). For splice junctions from simulated human reads, we also saw an increase in the F1 score
243 to 0.957. We conclude that an ensemble approach incorporating both junction metrics and sequence
244 information works best for detecting and filtering spurious splice junctions from alignments.

245

246 *Two-pass alignment with filtered splice junctions improves transcript identification*

247 We next applied the two decision tree filtering methods to perform two-pass alignment of the simulated
248 reads with minimap2[18]. Using filtered splice junctions, the percentage of junctions identified in second-
249 pass alignments which matched annotated splice junctions could be increased over first-pass alignment
250 and naive two-pass alignment (Fig. 5A). For example, using the simulated Arabidopsis datasets, the
251 median percentage of read alignments matching the splice junctions of the reference transcript they
252 were simulated from increased from 73.2% in the first pass to 88.2% and 89.3% in a second pass using
253 the first and second decision tree methods respectively (Fig. 5A). Two-pass alignment rescued the large
254 misalignments of exon 6 seen at *FLM* (Fig. S5A): overall, 86.8% of simulated *FLM* reads aligned to the
255 correct reference transcript after filtered two-pass alignment compared with 19.3% for first-pass
256 alignments. A global improvement in correct alignment was also seen in the simulated human datasets:
257 from 44.4% in the first pass to 64.3% and 65.7% for the two decision tree methods, respectively
258 (Fig. 5A).

259 Although two-pass alignment improved the number of reads aligning to the correct transcript model, we
260 also detected a slight increase in the number of reads aligning to the wrong annotated transcript. In the
261 simulated Arabidopsis reads, reads aligned using the second decision tree model performed worst on
262 this metric: 2.74% of reads aligned to the wrong isoform compared with only 1.79% of reads after first-
263 pass alignment (Fig. S5B). To assess whether such misassignment affects the quantitation of transcripts,
264 we calculated Spearman's correlation coefficient (ρ) for estimated versus known transcript level read
265 counts for the simulated data (Fig. 5B). The results indicated that, despite this misassignment, two-pass
266 aligned reads could be quantified accurately, with an overall improvement in median Spearman's ρ for
267 one-pass versus two-pass of from 0.876 to 0.916 for simulated Arabidopsis reads (Fig. 5B) and from

268 0.778 to 0.859 for simulated human reads (Fig. 5B). However, there may be corner cases where
269 transcript misassignment could have consequences for transcript usage analysis. This should be
270 considered for experiments where quantification is important. Overall, we conclude that two-pass
271 alignment using filtered junctions can improve both the detection of correct splicing patterns and the
272 quantitation of nanopore DRS reads.

273

274 *Filtered two-pass alignment improves reference-guided annotation*

275 Summarising read alignments into annotations facilitates transcript level quantification of short and long
276 reads and aids the interpretation of RNA processing complexity. We therefore asked whether two-pass
277 alignment of spliced long reads with relatively high sequence error rates can improve the results of
278 genome-guided annotation tools. Several software tools designed to produce annotations from long
279 reads exist, including FLAIR[10] and pinfish (ONT), which were designed for nanopore DRS data; TAMA[31],
280 which was designed for PacBio IsoSeq data; and StringTie2[6], which was designed as a technology-
281 agnostic long-read assembly tool.

282 We benchmarked our methods using StringTie2 because it is reported to be faster and more accurate
283 than FLAIR on simulated nanopore DRS data[6]. Using full-length reads simulated from real Arabidopsis
284 and human nanopore DRS data, we could identify the intron-chain-level precision and recall of
285 annotations assembled from reads processed using either one-pass or two-pass alignment. Here,
286 precision is defined as the percentage of assembled transcripts whose combination of introns match a
287 transcript in the reference annotation; and recall is defined as the percentage of annotated transcripts
288 for which at least one read was simulated and whose combination of introns match a transcript
289 assembled from simulated reads. Reads aligned using guide splice junctions from the reference
290 annotation were also assessed as a positive control.

291 For both Arabidopsis and human datasets, two-pass alignment generally produced a clear improvement
292 in both precision and recall of StringTie2 transcript assembly over first-pass alignment (Fig. 6A). Of the
293 two decision tree methods produced, decision tree 2 (using junction sequence information) performed
294 best (median F1 score was 0.699 for the Arabidopsis data and 0.629 for the human data). There was a
295 particularly large increase in precision for reference annotation-guided alignment of at least 8.7% and
296 9.6% over one-pass alignment for all Arabidopsis and human samples, respectively (Fig. 6A).

297 We next considered whether two-pass alignment could improve the genome-guided transcriptome
298 assembly performance of Stringtie2 on real datasets, using current reference annotations as a ground
299 truth. However, it is important to note that there may be genuine transcript examples in the datasets
300 that are not yet included in the reference annotation; if so, this will affect the measurement of
301 precision. Furthermore, recall against the reference is likely to depend on the sequencing depth of
302 samples. We therefore report the number of annotated transcripts assembled for each sample, rather
303 than the recall.

304 Two-pass alignment improved both the precision and the number of transcripts assembled for
305 Arabidopsis, human and mouse samples[10,11,32] (Fig. 6B–D). This approach resulted in a median increase in
306 assembly precision compared with one-pass alignment of 7.1% for Arabidopsis samples, 3.5% for human
307 samples and 2.2% for mouse samples (median increase in annotated transcripts assembled per sample
308 of 478.5, 257.5 and 238, respectively). We conclude that for organisms with complex patterns of pre-
309 mRNA splicing, two-pass alignment can improve both the precision and number of correct (annotated)
310 transcripts assembled by StringTie2 from real nanopore DRS data.

311 When we applied the same approach to the yeast *Saccharomyces cerevisiae*, the results were very
312 different (Fig. 6E). In this species, two-pass alignment resulted in a median increase of only three more
313 annotated transcripts assembled per sample and an increased number of unannotated transcripts
314 assembled, resulting in a median decrease of 0.8% in assembly precision. Splicing complexity in
315 *S. cerevisiae* is relatively low: there are only 364 annotated introns in the Ensembl R64 annotation, most
316 genes are intronless, and most introns are constitutive[33]. This led to a high ratio of unannotated splice
317 junctions in first-pass alignments (the median number of junctions identified was 8,056), suggesting that
318 the vast majority of junctions in the dataset are spurious. Furthermore, most introns occur close to
319 mRNA 5' ends, resulting in typically short upstream exons that present challenges to alignment
320 software. Such a large ratio of spurious to genuine splice junctions is likely to affect the precision of
321 junction filtering. Notably, even when the reference annotation was used to guide alignment, precision
322 was only improved by a median of 1.9% (with a median of six more transcripts assembled correctly).
323 Intron-containing genes are generally more highly expressed (many encode ribosomal proteins) than
324 intronless genes[34]. This may mean that the coverage of annotated transcripts is already good and, thus,
325 that the number of true annotated transcripts assembled cannot be much improved. This result
326 suggests that both reference annotation-guided and two-pass alignment methods have limited use for
327 genome-guided transcriptome assembly in organisms with low complexity splicing.

328 Finally, we considered whether filtered two-pass alignment could improve genome-guided annotation of
329 nanopore DRS reads derived from sequencing cDNA copies and from PacBio IsoSeq data (Fig S6A-D). To
330 assess this, we used the recommended alignment parameters for minimap2[18] but with the splice
331 junction filtering parameters as were used for nanopore DRS data. Overall, the precision and recall of
332 transcripts assembled from both nanopore cDNA and PacBio IsoSeq data for human, mouse and
333 Arabidopsis samples could be improved using two-pass alignment. For human and mouse nanopore
334 cDNA samples, two-pass alignment resulted in a median increase of 3.85% and 2.3% in assembly
335 precision, respectively, compared with one-pass alignment (median increase in annotated transcripts
336 assembled per sample of 609.5 and 420.0, respectively; Fig. S6A,B). For Arabidopsis and human PacBio
337 IsoSeq samples, two-pass alignment resulted in a median increase of 8.45% and 1.35% in assembly
338 precision, respectively, compared with one-pass alignment (median increase in annotated transcripts
339 assembled per sample of 63 and 242.5, respectively; Fig. S6C,D). We conclude that a two-pass method
340 can improve genome-guided transcript assembly of the high-error long reads produced using a range of
341 technologies.

342

343 Conclusion

344 RNA sequencing is a fundamental tool for understanding what genomes really encode. Technological
345 approaches that directly sequence full-length RNA molecules substantially increase the useful
346 information that RNA sequencing can provide. The challenges that alternative splicing, in particular,
347 presents to the interpretation of high-throughput RNA sequencing data means that software
348 development needs to accompany progress in sequencing technology. In this way, knowledge gained
349 from ambitious genome sequencing programmes such as the Earth BioGenome Project, which aims to
350 characterise all eukaryotic life on Earth[35], can be maximised. In addition, we can deepen our
351 understanding of how gene expression changes in disease states, for example, in humans and other
352 well-characterised species. We have shown that a two-pass alignment approach, informed by splice
353 junction alignment metrics and machine learning of sequence features associated with splicing, can
354 improve the accuracy of intron detection in long-read data. Consequently, this approach can enhance
355 the utility and realise the potential of long-read RNA sequencing.

356

Acknowledgments

We thank Kasia Knop for comments on the manuscript, and James Abbott for testing the pipeline. This work was funded by the University of Dundee Global Challenges Research Fund.

Methods

Nanopore and PacBio data

Four replicates of nanopore DRS reads derived from Arabidopsis Col-0 RNA were used [11]. These datasets are available in FAST5 format from the European Nucleotide Archive under accession no. PRJEB32782. The first four listed replicates of DRS and cDNA sequencing reads derived from human cell line GM12878 were used: Birmingham DRS samples 1, 2, 3 and 5; Birmingham cDNA samples 1 and 2; Hopkins cDNA samples 1 and 2)[10]. DRS datasets were downloaded in FAST5 format and cDNA datasets in FASTQ format using the links provided on GitHub (http://s3.amazonaws.com/nanopore-human-wgs/rna/links/NA12878-DirectRNA_All.files.txt). Mouse DRS and cDNA datasets in FASTQ format [32] were downloaded from the European Nucleotide Archive (accession no. PRJEB27590). Yeast DRS datasets in FASTQ format [9] were downloaded from the European Nucleotide Archive (accession no. PRJNA408327). Human IsoSeq datasets in FASTQ format were downloaded from the PacBio AWS webserver (http://datasets.pacb.com.s3.amazonaws.com/2014/Iso-seq_Human_Tissues/list.html). Arabidopsis IsoSeq data in FASTQ format was downloaded from the European Nucleotide Archive (accession no. PRJNA371677).

Preliminary data processing

Pipelines for processing of data were written using snakemake version 5.10.0[36]. FAST5 data was re-basecalled locally using guppy version 2.3.1 (ONT). All alignments were performed using minimap2 version 2.17-r963[18]. Arabidopsis reads were aligned to the TAIR10 reference genome[37] and AtRTD2 reference transcriptome[23]. Human, mouse and yeast reads were aligned to the GRCh38, GRCm38 and R64–1-1 primary assemblies and to cDNA transcriptomes from Ensembl, respectively[24]. Alignments to reference genomes were performed using spliced parameters. For DRS datasets, these were: -k14 -w5 --splice -g2000 -A1 -B2 -O2,32 -E1,0 -C9 --splice-flank=yes -uf -z200 -L --cs=long. For nanopore cDNA and PacBio datasets the parameters used were -x splice -g2000 -L --cs=long. The maximum intron size (-G) was set at 10,000 nt for Arabidopsis samples, at 200,000 nt for human and mouse datasets and at

386 5,000 nt for yeast. For two-pass alignments using a guide splice junction set, a junction bonus of 12 was
387 also used. Alignments of DRS reads to the reference transcriptome were performed using splicing-free
388 parameters, namely: -k14 --for-only -L --cs=long.

389 *Simulation of DRS reads*

390 To provide a ground truth with a complete set of known splice sites, sequences were simulated from the
391 reference transcriptomes, with length and error profiles matching those of real DRS reads. This was
392 done by modelling the length, homopolymer error and other error profiles of real reads. Only primary
393 alignments were considered. The cs tags of reads aligned to the reference transcriptome were used to
394 recreate pairwise alignments between each read and the reference, ignoring refskips. Alignments were
395 inverted to match the 3′ → 5′ sequencing direction of nanopore DRS. Aligned basecalls at reference
396 homopolymers of ≥5 nt in length were used to build a probability model of homopolymer calls given the
397 reference homopolymer. To prevent these error profiles being modelled multiple times, the reference
398 homopolymer was then replaced with the aligned basecall in the pairwise alignment. Next, the altered
399 alignment was used to create a Markov chain model of basecalled sequence given the reference
400 sequence. For each base in the reference sequence in the alignment, the aligned portion of the query
401 sequence was identified. The "state" of the alignment (i.e. match, mismatch, insertion or deletion) was
402 also identified. The probability of seeing a query sequence was calculated, given the current and
403 previous four bases of the reference and the previous four states of the alignment.
404 The reference transcriptome was also used to simulate data using these models. The number of primary
405 alignments in the real data for each reference transcript was used as the number of simulated reads per
406 transcript. To simulate basecall errors, sequences were inverted to the 3′ → 5′ direction and reads were
407 generated using Markov chain Monte Carlo simulations with the basecall model. The reference
408 sequences were prepended with a 10 nt oligo(A) sequence to mimic a short poly(A) tail so that the initial
409 state of the Markov chain was always "AAAAA" and "====" (i.e. four matches). Homopolymers in the
410 simulated read were identified and replaced with randomly selected sequences from the homopolymer
411 model. The read was then reverted to the 5′ → 3′ direction for mapping. Because we wanted to assess
412 the alignment of full-length reads, we did not model or simulate the 3′ bias, which is inherent to
413 nanopore DRS data. However, 10 nt of simulated read were subtracted from the 5′ end of reads to
414 simulate loss of signal at the end of sequencing.

Post-alignment splice junction correction with FLAIR

BAM files were converted to the BED12 format using bedtools[38]. BED12 files were then corrected using the reference GTF annotation with FLAIR correct version 1.4 and default settings[10].

Junction metric calculations

Splice junctions and junction metrics were extracted from aligned reads using the long form cs tag produced by minimap2 version 2.17[18] using pysam version 0.15.4. The per-read JAD was calculated as the length of the shorter of the two match operations immediately flanking refskip (splicing) operations. Where there were mismatches or indels immediately adjacent to refskips, a JAD of zero was assigned. The per-splice junction JAD was calculated as the maximum of the per-read JADs. Intron motifs were extracted from cs tags. For Arabidopsis, human and mouse samples, GU/AG, GC/AG and AU/AG splice junctions were all considered canonical. For yeast samples, only GU/AG splice junctions were considered canonical. To calculate the primary donor/acceptor metrics, interval trees of donor and acceptor sites were constructed using NCLS[39]. Donors were assigned as primary donors if there were no alternative donor sites within 20 nt with higher read counts. Likewise, acceptors were considered primary if there were no alternative acceptors within 20 nt with higher read counts. Ties were broken using the JAD metric, i.e. the splice junctions with higher JADs were assigned primary status. Where there were still ties after read count and JAD comparisons, no splice junctions were assigned primary status. Splice junctions extracted from four replicates of Arabidopsis or human DRS reads were used to build decision tree models with scikit-learn version 0.22.1[40]. A minimum depth of 4, minimum number of samples required to split a node of 1000, and minimum Gini impurity decrease required to split a node of 0.005 were used. The decision tree generated from Arabidopsis reads was a subtree of the human tree (i.e. it could be created by pruning the human tree), indicating that the decision function can generalise across samples.

Reference sequence filtering using LR models

Splice junctions obtained from a first-pass alignment were separated into lists of unique donor sites and unique acceptor sites. These were labelled as positive training examples if they participated in at least one donor/acceptor pair which passed the first decision tree function. Sequences of 128 nt for each splice junction (centred on the donor or acceptor site) were extracted from the reference genome using pysam version 0.15.4 and one hot encoded into four binary variables to create a 512-feature training dataset. LR models were trained using 6-fold cross-validation with scikit-learn version 0.22.1[40]. For each

445 fold, the model was used to generate out-of-bag predictions on the held-out data. The probabilities
446 produced were then used in place of the canonical intron motif to produce the second decision tree,
447 using a maximum depth of 6, a minimum number of samples of 1,000 and a minimum Gini impurity
448 decrease of 0.003. Thresholds for splice scores in the tree were simplified to comprise only a high
449 confidence threshold of 0.6 (for rescuing splice junctions failing the JAD metric threshold) and a low
450 confidence threshold of 0.1 (for removing false positives from junctions passing the JAD metric
451 threshold).

452 *Evaluation of splice junction models*

453 Performance of the metrics and models was evaluated at splice junction level using the reference
454 annotation as a ground truth. For simulated datasets, annotation is the absolute ground truth because
455 all reads are simulated using only splice junctions in the annotation. For real datasets, some "false
456 positives" are likely to be genuine splice junctions and some junctions in the reference, which appear as
457 false negatives, are actually incorrectly annotated or not expressed. Precision is defined as the number
458 of true positives divided by the total number of positive predictions by the model, i.e. true positives ÷
459 (true positives + false positives). Recall is defined as the number of true positives divided by the total
460 number of real positive examples in the dataset, i.e. true positives ÷ (true positives + false negatives).
461 The F1 score is the harmonic mean of the precision and recall.

462 *Evaluation of alignments*

463 To evaluate alignments, we used the intron chain of reference transcripts as a ground truth. The intron
464 chain is the pattern of linked splicing in a transcript, disregarding the transcription start and termination
465 sites. Alignments of simulated reads were considered correct if they mapped correctly to the intron
466 chain of the reference transcript they were simulated from, with no mistakes. Simulated reads that were
467 mapped using intron chains not included in the reference or as being intron-less when they should have
468 splicing were considered novel spurious alignments. Simulated reads that were mapped using the intron
469 chain of a reference transcript other than the transcript they were simulated from were considered to
470 be misassigned. For measures of quantification accuracy, alignment counts for transcripts were
471 generated using the number of simulated reads that aligned with the same splice junctions as the
472 reference transcript. Spearman's correlation coefficients were then calculated against the known input
473 transcript counts for simulation.

474 *Reference-guided assembly*

475 Reference-guided transcriptome assemblies were produced using StringTie2[6] version 2.1.1 in long-read
476 mode, with otherwise default parameters.

477 *Evaluation of assemblies*

478 Reference-guided transcriptome assemblies were evaluated using the precision and recall of intron
479 chains calculated using gffcompare with default settings[41]. The input reference GTF files were filtered to
480 include only transcript models for which at least one read had been simulated.

481 *Code availability*

482 The methods used to filter splice junctions have been implemented in the "2passtools" python package,
483 which is available on GitHub in repository https://github.com/bartongroup/2passtools. The software
484 used to simulate reads is available on GitHub in repository https://github.com/bartongroup/yanosim.
485 The scripts, pipelines and notebooks used to perform benchmarking and generate figures are available
486 on GitHub in repository https://github.com/bartongroup/two_pass_alignment_pipeline. Basecalled and
487 simulated nanopore DRS datasets are available from Zenodo at https://zenodo.org/record/3773729.

488

489 # References

490 1 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**,
491 470-476, doi:10.1038/nature07509 (2008).
492 2 Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics*
493 **20**, 631-656, doi:10.1038/s41576-019-0150-2 (2019).
494 3 Mourão, K. *et al.* Detection and mitigation of spurious antisense expression with RoSA.
495 *F1000Research* **8**, doi:10.12688/f1000research.18952.1 (2019).
496 4 Houseley, J. & Tollervey, D. Apparent Non-Canonical Trans-Splicing Is Generated by Reverse
497 Transcriptase In Vitro. *PLoS ONE* **5**, doi:10.1371/journal.pone.0012271 (2010).
498 5 Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for
499 RNA-seq isoform quantification. *BMC Genomics* **18**, doi:10.1186/s12864-017-4002-1 (2017).
500 6 Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2.
501 *Genome Biology* **20**, doi:10.1186/s13059-019-1910-1 (2019).
502 7 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq
503 reads. *Nature Biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
504 8 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated
505 transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515,
506 doi:10.1038/nbt.1621 (2010).
507 9 Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nature*
508 *Methods* **15**, 201-206, doi:10.1038/nmeth.4577 (2018).
509 10 Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome.
510 *Nature Methods* **16**, 1297-1305, doi:10.1038/s41592-019-0617-2 (2019).
511 11 Parker, M. T. *et al.* Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA
512 processing and m6A modification. *eLife* **9**, doi:10.7554/eLife.49658 (2020).
513 12 Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT)
514 sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids*
515 *Research* **46**, 2159-2168, doi:10.1093/nar/gky066 (2018).
516 13 Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford
517 Nanopore sequencing. *Genome Biology* **20**, doi:10.1186/s13059-019-1727-y (2019).
518 14 Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore
519 reads with deep convolutional neural networks. *PLOS Computational Biology* **14**,
520 doi:10.1371/journal.pcbi.1006583 (2018).
521 15 Dehghannasiri, R., Szabo, L., Salzman, J. & Birol, I. Ambiguous splice sites distinguish circRNA and
522 linear splicing in the human genome. *Bioinformatics* **35**, 1263-1268,
523 doi:10.1093/bioinformatics/bty785 (2019).
524 16 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21,
525 doi:10.1093/bioinformatics/bts635 (2013).
526 17 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and
527 genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915,
528 doi:10.1038/s41587-019-0201-4 (2019).
529 18 Li, H. & Birol, I. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
530 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
531 19 Liu, B. *et al.* deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-
532 based index. *Genome Biology* **20**, doi:10.1186/s13059-019-1895-9 (2019).

20 Veeneman, B. A., Shukla, S., Dhanasekaran, S. M., Chinnaiyan, A. M. & Nesvizhskii, A. I. Two-pass alignment improves novel splice junction quantification. *Bioinformatics* **32**, 43-49, doi:10.1093/bioinformatics/btv642 (2016).

21 Gatto, A. *et al.* FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucleic Acids Research* **42**, e71-e71, doi:10.1093/nar/gku166 (2014).

22 Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience* **7**, doi:10.1093/gigascience/giy131 (2018).

23 Zhang, R. *et al.* A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Research* **45**, 5061-5073, doi:10.1093/nar/gkx267 (2017).

24 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**, 849-864, doi:10.1101/gr.213611.116 (2017).

25 Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology* **19**, doi:10.1186/s13059-018-1462-9 (2018).

26 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195-197, doi:10.1016/0022-2836(81)90087-5 (1981).

27 Sheth, N. *et al.* Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research* **34**, 3955-3967, doi:10.1093/nar/gkl556 (2006).

28 Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**, 372-381, doi:10.1016/j.cell.2016.02.045 (2016).

29 Reimer, K. A., Mimoso, C., Adelman, K. & Neugebauer, K. M. Rapid and Efficient Co-Transcriptional Splicing Enhances Mammalian Gene Expression. *bioRxiv*, doi:10.1101/2020.02.11.944595 (2020).

30 Mercer, T. R. *et al.* Genome-wide discovery of human splicing branchpoints. *Genome Research* **25**, 290-303, doi:10.1101/gr.182899.114 (2015).

31 Kuo, R. I., Cheng, Y., Smith, J., Archibald, A. L. & Burt, D. W. Illuminating the dark side of the human transcriptome with TAMA Iso-Seq analysis. *bioRxiv*, doi:10.1101/780015 (2019).

32 Sessegolo, C. *et al.* Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports* **9**, doi:10.1038/s41598-019-51470-9 (2019).

33 Spingola, M., Grate, L., Haussler, D. & Ares, M. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *Rna* **5**, 221-234, doi:10.1017/s1355838299981682 (1999).

34 Ares, M., Grate, L. & Pauling, M. H. A handful of intron-containing genes produces the lion's share of yeast mRNA. *Rna* **5**, 1138-1139, doi:10.1017/s1355838299991379 (1999).

35 Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115**, 4325-4333, doi:10.1073/pnas.1720115115 (2018).

36 Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520-2522, doi:10.1093/bioinformatics/bts480 (2012).

37 Initiative, T. A. G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

38 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

39 Stovner, E. B., Sætrom, P. & Hancock, J. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics*, doi:10.1093/bioinformatics/btz615 (2019).

580 40 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830
581 (2011).
582 41 Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis
583 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* **11**, 1650-1667,
584 doi:10.1038/nprot.2016.095 (2016).

585

586

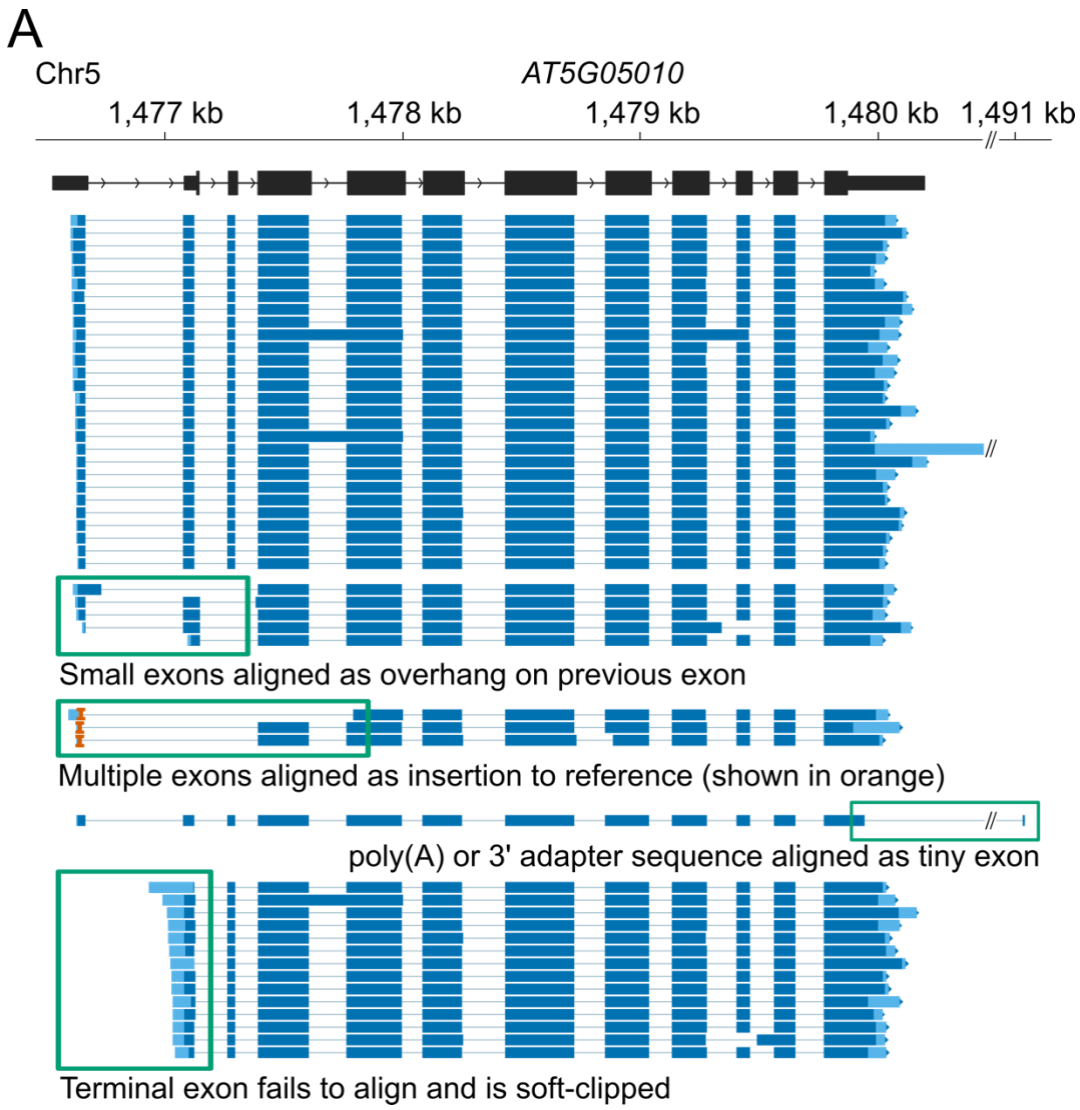
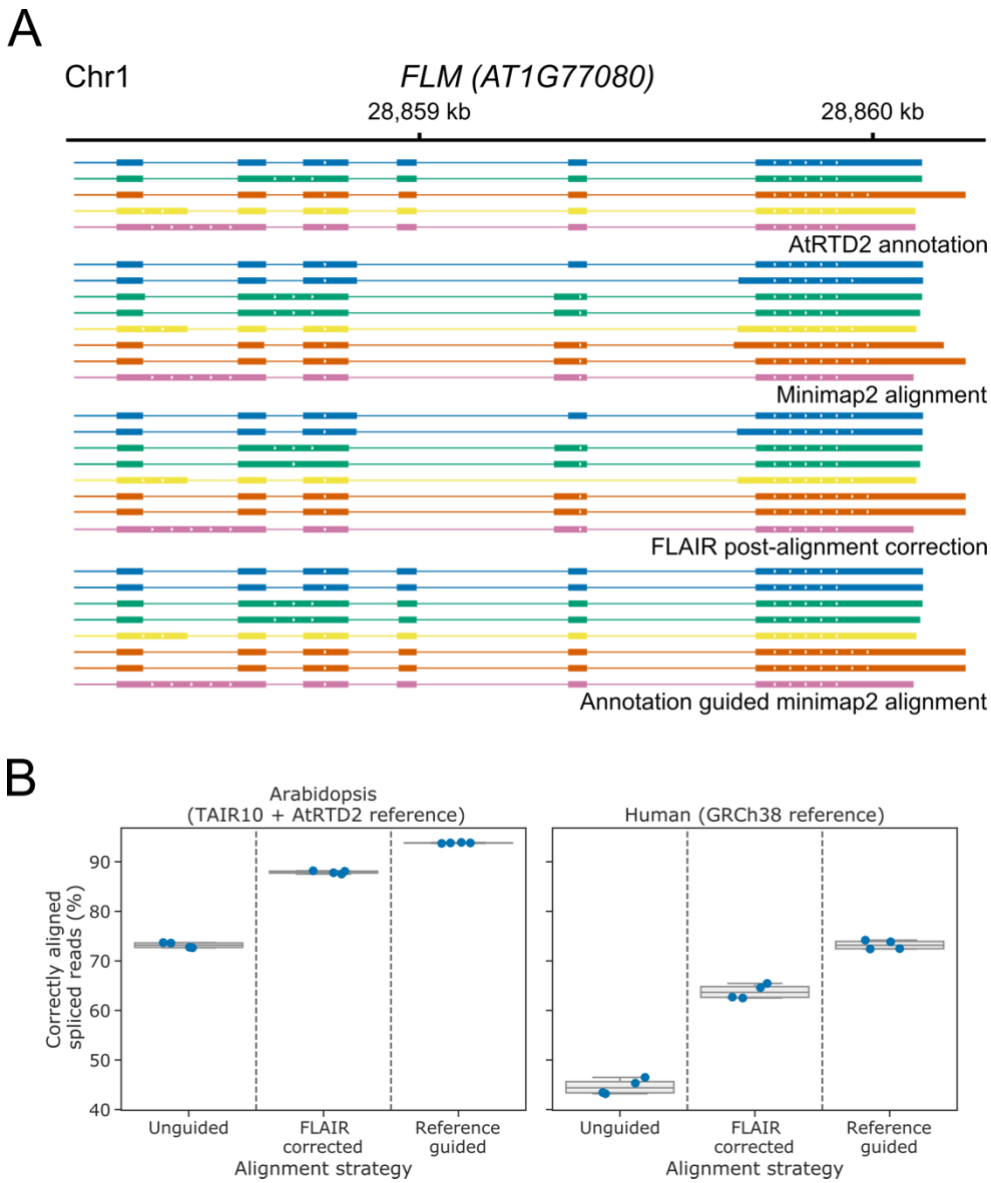


Fig. 1. Assessment of alignment errors in nanopore DRS datasets. Nanopore DRS read alignments at Arabidopsis *AT5G05010* locus with different types of alignment error presented. Read alignments are shown in dark blue, with soft-clipped (unaligned) ends shown in light blue. Mismatches and indels of <30 nt are not shown. Insertions to the reference of > 30 nt are shown as orange carets.



592

593 **Fig. 2. Improved spliced alignment of simulated reads using annotation-guided alignment. A**
594 Reference-guided alignment improves the identification of small exons in nanopore DRS reads. Gene
595 track showing the alignment of a sample of simulated nanopore DRS reads at the Arabidopsis *FLM* gene.
596 AtRTD2 reference annotation, from which reads were simulated, is shown on top, with unguided
597 minimap2 alignments, FLAIR correction of unguided minimap2 alignments and reference-guided
598 minimap2 alignments shown below. Only reads where exon 6 failed to align in the initial unguided
599 alignment are shown. Each read alignment is coloured based on the reference transcript it was
600 simulated from, and reads are shown in the same order within each alignment method group.
601 Mismatches and indels are not shown. **B** Reference-guided alignment improves the identification of
602 correct transcripts globally. Boxplots with overlaid strip-plots showing the percentage of alignments

603 which map exactly to the splice junctions of the transcript from which they were simulated, for
604 unguided minimap2 alignments, FLAIR correction of unguided minimap2 alignments using reference
605 annotation, and reference annotation-guided minimap2 alignments. Reads simulated from intronless
606 transcripts which map correctly without splicing were not included in percentage calculations. Reads
607 were simulated from Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and
608 GRCh38 reference transcriptomes, respectively.

609

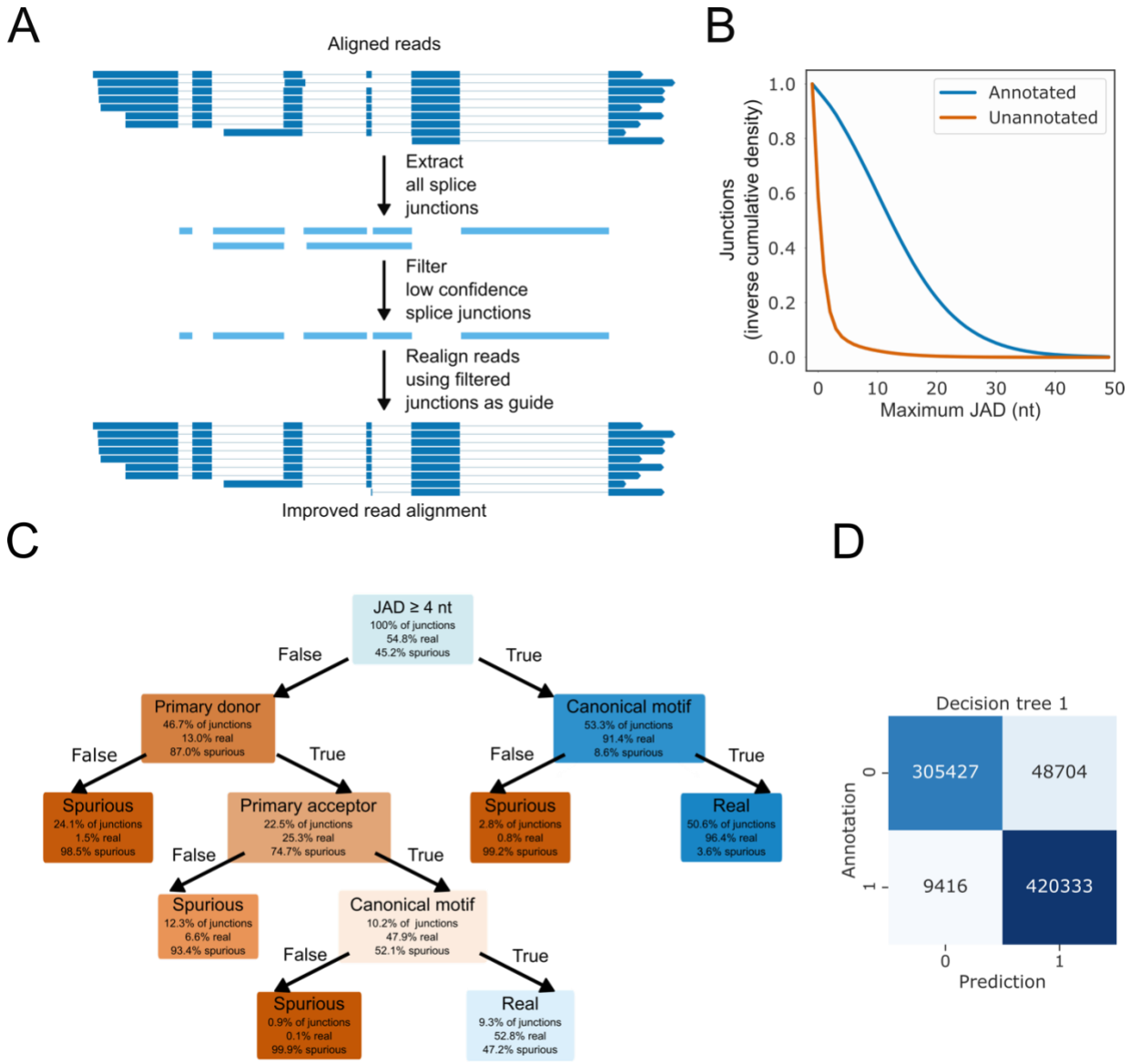


Fig. 3 Junction metrics can identify genuine splice junctions. A Outline of the two-pass method. **B** The JAD metric can discriminate between annotated and unannotated splice junctions in simulated nanopore DRS reads. Inverse cumulative density plot showing the distribution of per-splice junction maximum JAD values for annotated (blue) and unannotated (orange) splice junctions. **C** Flowchart visualisation of the first decision tree model. Nodes (decisions) and leaves (outcomes) are coloured based on the relative ratio of real and spurious splice junctions. **D** Confusion matrix showing the ratios of correct and incorrect predictions of the first decision tree model on splice junctions extracted from simulated Arabidopsis read alignments.

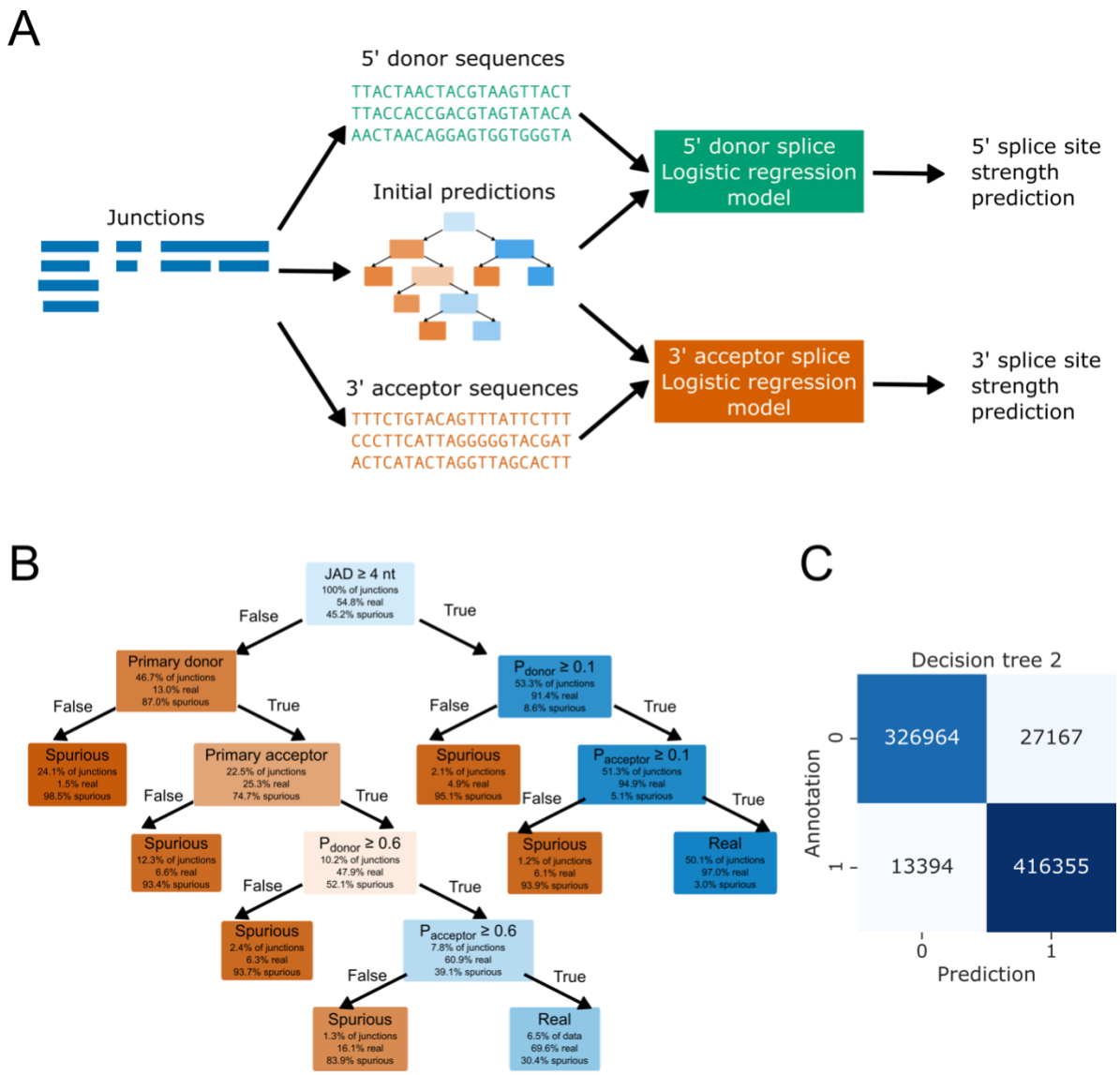


Fig. 4. Machine learned sequence information improves identification of genuine splice junctions. A Outline of the LR model training process. Sequences from splice junctions were extracted from the reference genome and used as training data (i.e. explanatory variables). Training labels (i.e. the response variable) were generated by the first decision tree model. Independent models were trained for 5' donor and 3' acceptor sites and cross-validation used to generate out-of-bag predictions for all sites. **B** Flowchart visualisation of the second decision tree model. Nodes (decisions) and leaves (outcomes) are coloured based on the relative ratio of real and spurious splice junctions. **C** Confusion matrix showing the ratios of correct and incorrect predictions of the second decision tree model on splice junctions extracted from simulated Arabidopsis read alignments.

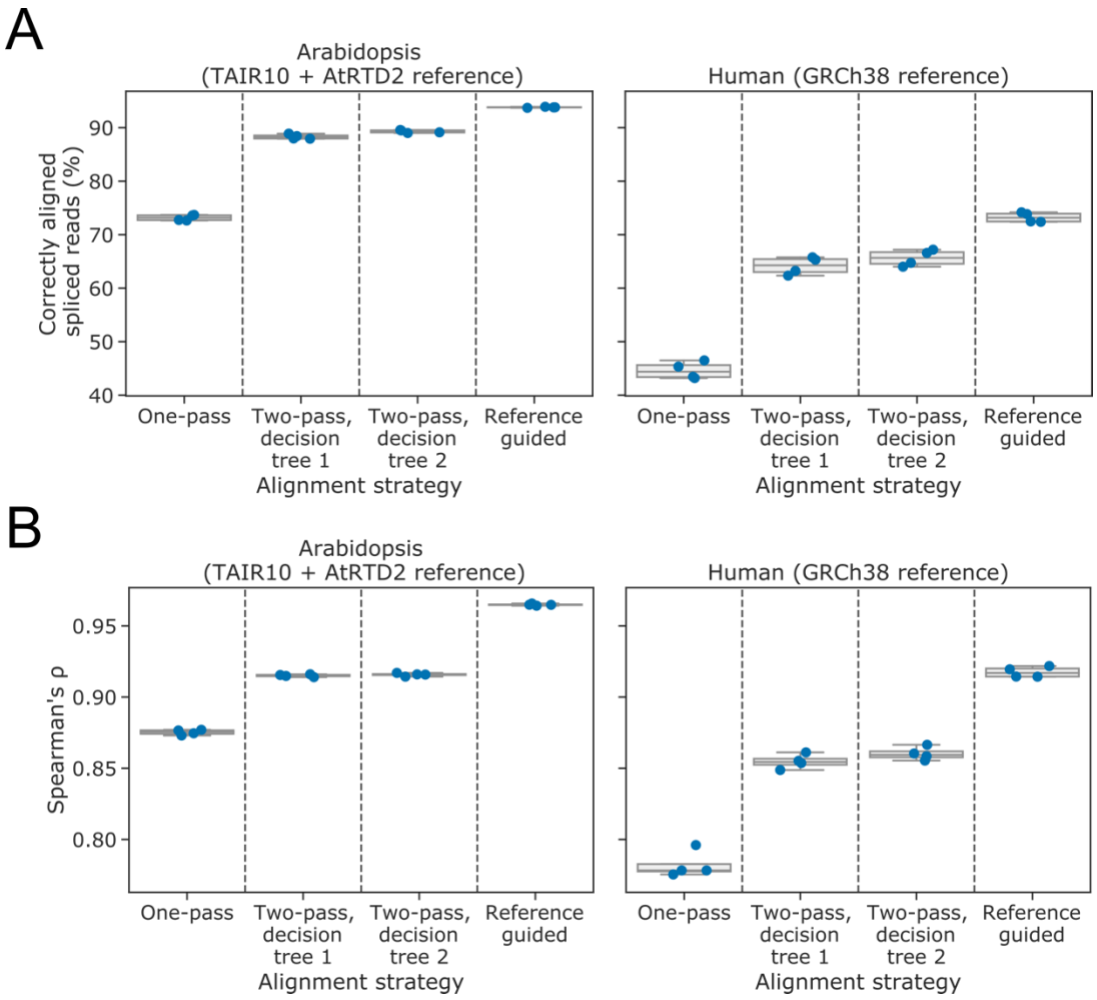


Fig. 5. Filtered two-pass alignment improves the identification and quantification of correct transcripts without a reference annotation. A Boxplots with overlaid strip-plots showing the percentage of alignments which map exactly to the splice junctions of the transcript from which they were simulated, for one-pass unguided minimap2 alignments, two-pass alignments using splice junctions filtered by decision trees one and two, and reference-annotation-guided minimap2 alignments. Reads were simulated from Arabidopsis TAIR10 + AtRTD2 (left) and human GRCh28 (right) nanopore DRS data. **B** Boxplots with overlaid strip-plots showing the Spearman's correlation coefficient for actual transcript level counts from simulated data against counts produced by the alignment methods described in A. Reads were simulated from Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and GRCh38 reference transcriptomes, respectively.

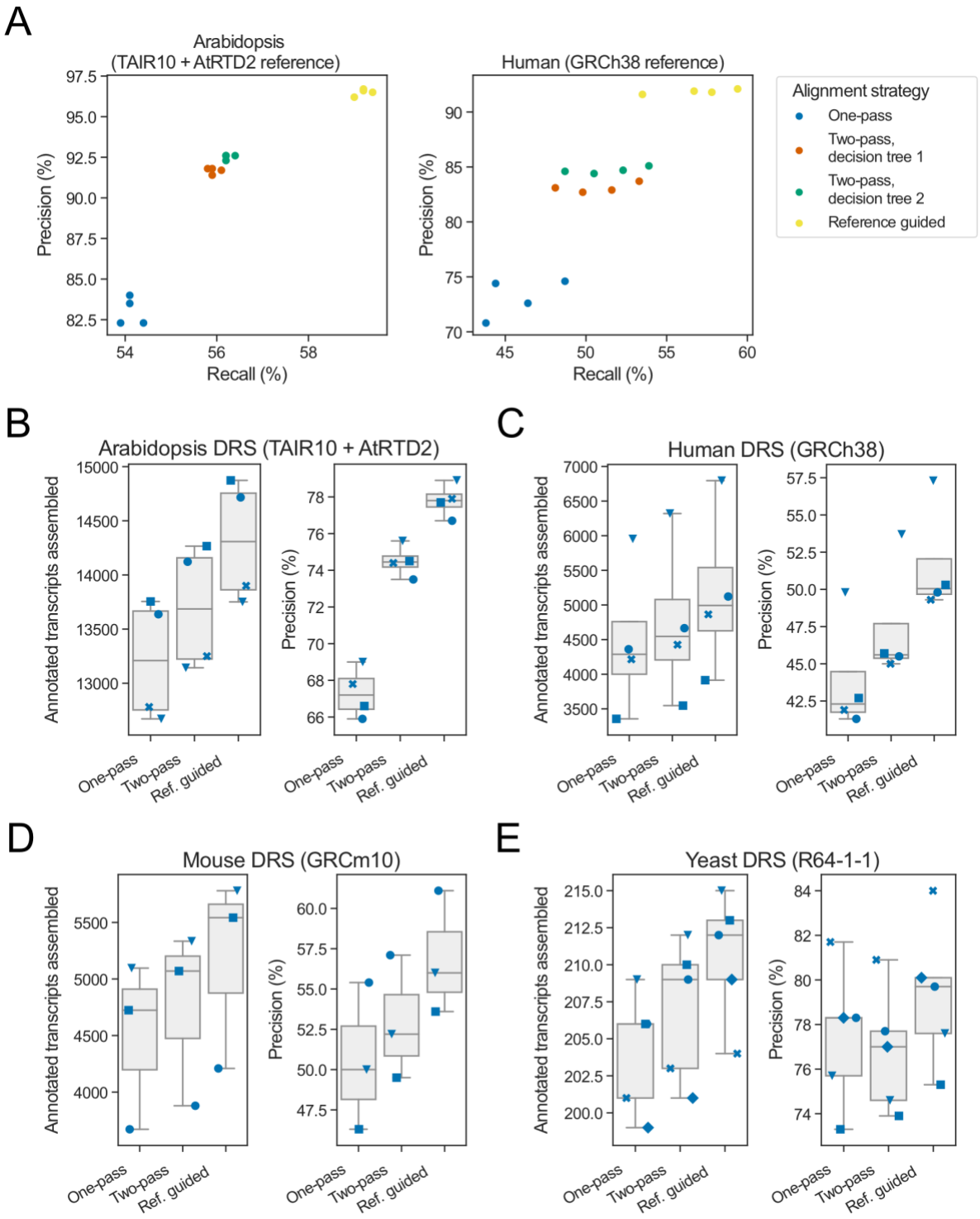


Fig. 6. Filtered two-pass alignment improves genome-guided annotation. A Scatterplot showing precision against recall for intron chains in genome-guided transcriptome annotations generated from alignments using StringTie2. Precision and recall scores were calculated against reference annotations filtered to include only transcripts for which at least one read was simulated. Reads were simulated

648 from Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and GRCh38
649 reference transcriptomes, respectively. **B–E** Boxplots showing the number of correct transcripts
650 assembled (left panels) and precision of transcripts assembled (right panels) for genome-guided
651 transcriptome assembly using StringTie2. Two-pass alignment improved the precision and number of
652 transcripts assembled for real nanopore DRS data for **B** Arabidopsis, **C** human, **D** mouse and **E** yeast. For
653 all boxplots, overlaid strip-plots are shown for individual samples. Each sample was assigned a unique
654 marker so that the changes in each sample could be tracked between the one-pass, two-pass and
655 reference-guided alignments.

656

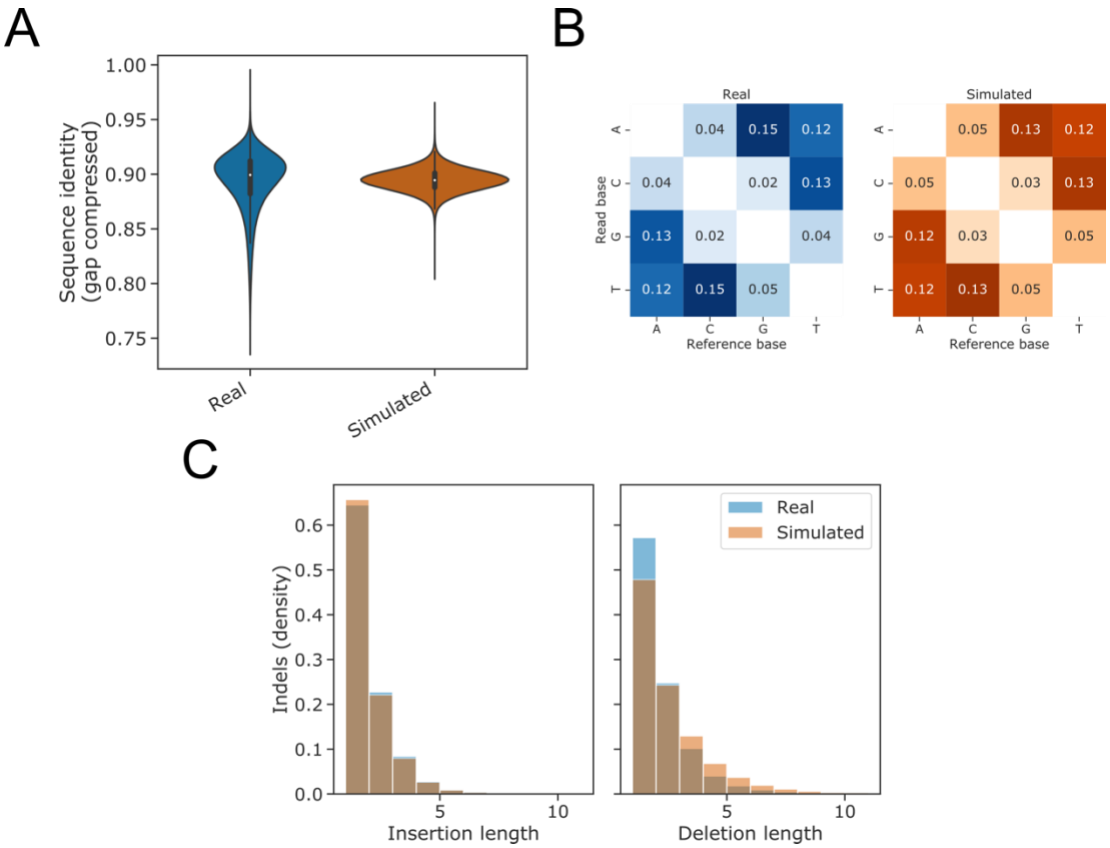


Fig. S1. Simulation of nanopore DRS read alignments. A Violin plot showing the distribution of sequence identity scores for real and simulated Arabidopsis nanopore DRS reads. Simulated reads match the median sequence identity of real reads, although they do not capture the tails of high- and low-quality reads. **B** Insertion and deletion length distributions for real and simulated nanopore DRS reads. **C** Mismatch profiles for real and simulated nanopore DRS reads.

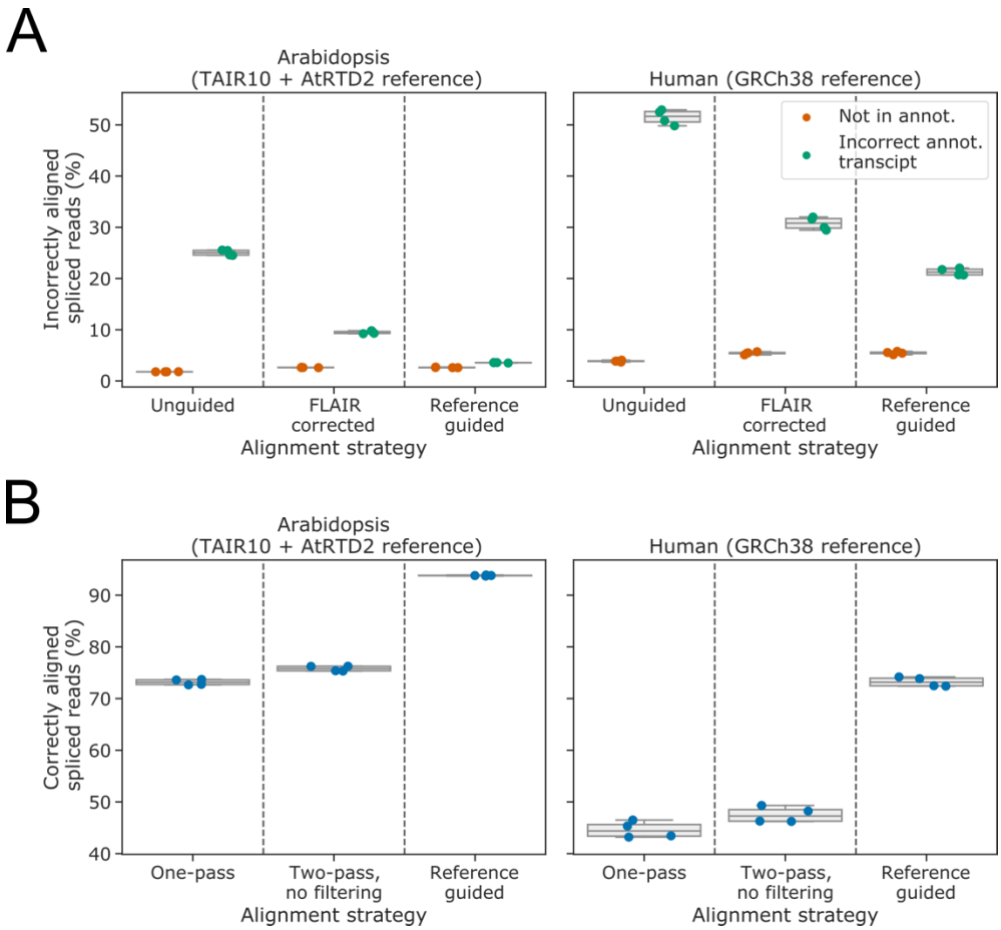


Fig. S2. Annotation-guided alignment improves spliced alignment of simulated reads. A Boxplots with overlaid strip-plot showing the percentage of alignments which do not map correctly to the splice junctions of the transcript from which they were simulated, for one-pass unguided minimap2 alignments, FLAIR-corrected alignments and reference annotation-guided minimap2 alignments. Reads that align to unannotated splice junctions or combinations of junctions are shown in orange. Reads which align to the incorrect annotated combination of splice junctions are shown in green. Reads were simulated from Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and GRCh38 reference transcriptomes, respectively. **B** Boxplots with overlaid strip-plot showing the percentage of alignments which map correctly to the splice junctions of the transcript from which they were simulated, for one-pass unguided minimap2 alignments, two-pass minimap2 alignment and reference annotation-guided minimap2 alignments. Reads were simulated from Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and GRCh38 reference transcriptomes, respectively.

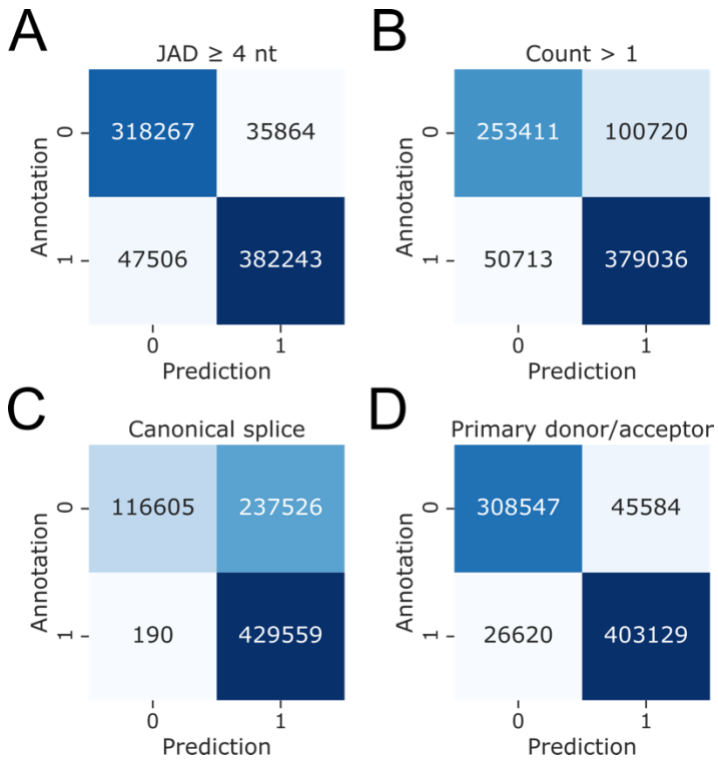


Fig. S3. Junction metrics can identify genuine splice junctions. A-D Confusion matrices showing the ratios of correct and incorrect predictions using: **A** a JAD threshold of 4 nt; **B** a count threshold of 1 nt; **C** the presence of a canonical U2 GU/AG, U12 GC/AG or U12 AU/AG intron motif; and **D** the primary donor/acceptor metric, defined as whether there are no alternate donor or acceptor sites with greater support (i.e. higher count or JAD) within 20 nt.

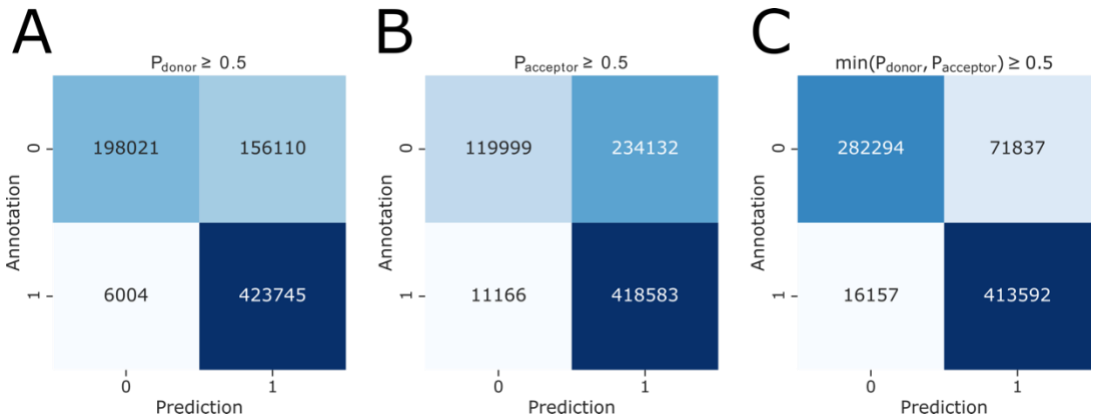


Fig. S4. Machine-learned sequence information can identify genuine splice junctions. A-C Confusion matrices showing the ratios of correct and incorrect predictions using: **A** an LR prediction threshold of 0.5 for splice site strength predictions made on donor site sequences; **B** an LR prediction threshold of 0.5 for splice site strength predictions made on acceptor site sequences; **C** a minimum prediction threshold of 0.5 for both splice donor and acceptor site sequences.

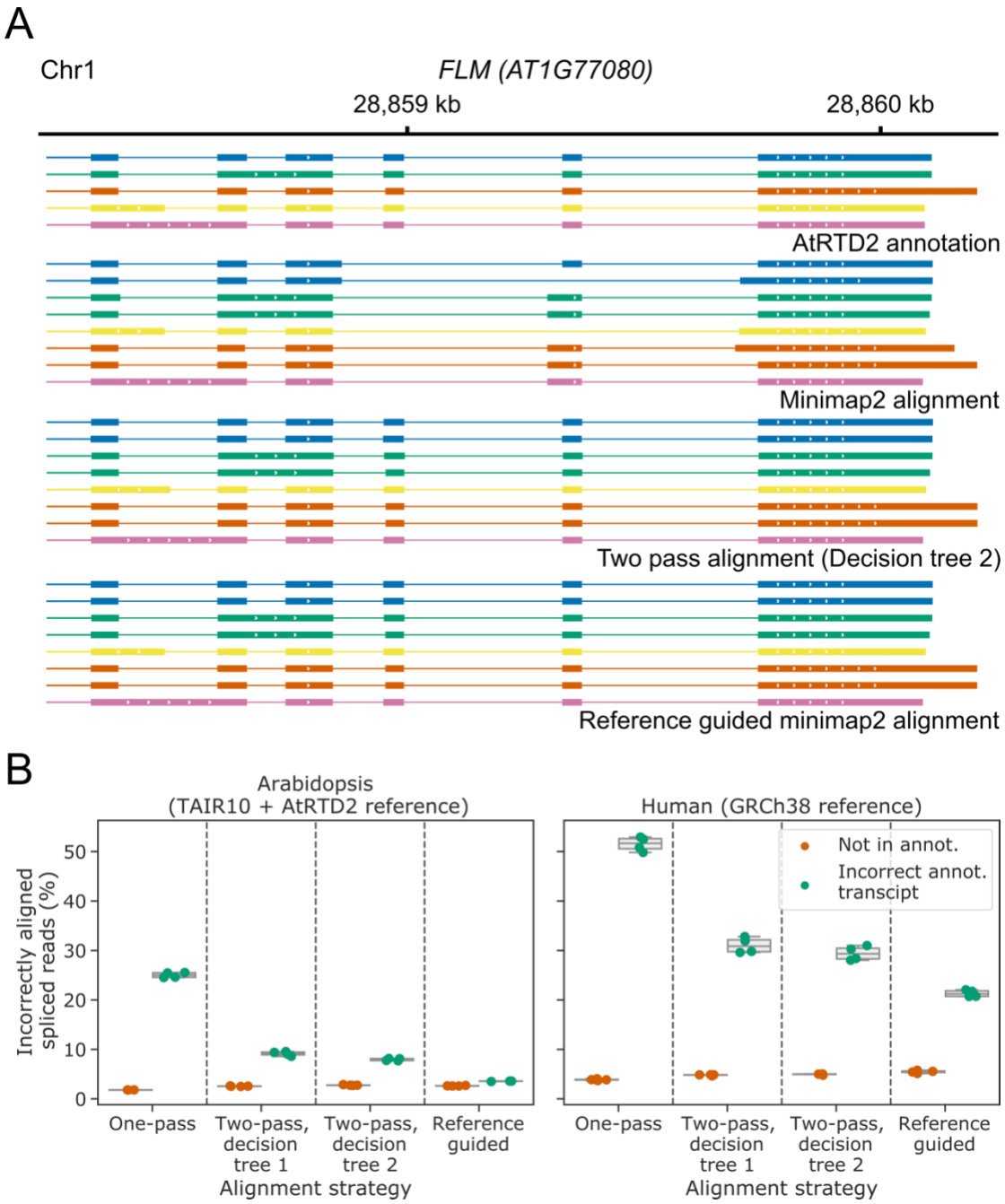


Fig. S5. Filtered two-pass alignment improves the identification and quantification of correct transcripts without a reference annotation. A Gene track showing alignment of a sample of simulated nanopore DRS reads at the Arabidopsis *FLM* gene. The AtRTD2 reference annotation, from which reads were simulated, is shown on top, with unguided minimap2 alignments, two-pass minimap2 alignments using the second decision tree classification, and reference-annotation-guided alignments shown below. Only reads where exon 6 failed to align in the initial unguided alignment are shown. Each read alignment is coloured based on the reference transcript it was simulated from, and reads are in the same order

701 within each alignment method group. Mismatches and indels are not shown. **B** Boxplots with overlaid
702 strip-plots showing the percentage of alignments which do not map correctly to the splice junctions of
703 the transcript from which they were simulated, for one-pass unguided minimap2 alignments, two-pass
704 alignment with decision trees one and two, and reference annotation-guided minimap2 alignments.
705 Reads that align to unannotated splice junctions or combinations of junctions are shown in orange.
706 Reads which align to annotated combinations of splice junctions which they were not simulated from
707 are shown in green. Reads were simulated from Arabidopsis (left) and human (right) nanopore DRS data
708 aligned to the AtRTD2 and GRCh38 reference transcriptomes, respectively.

709

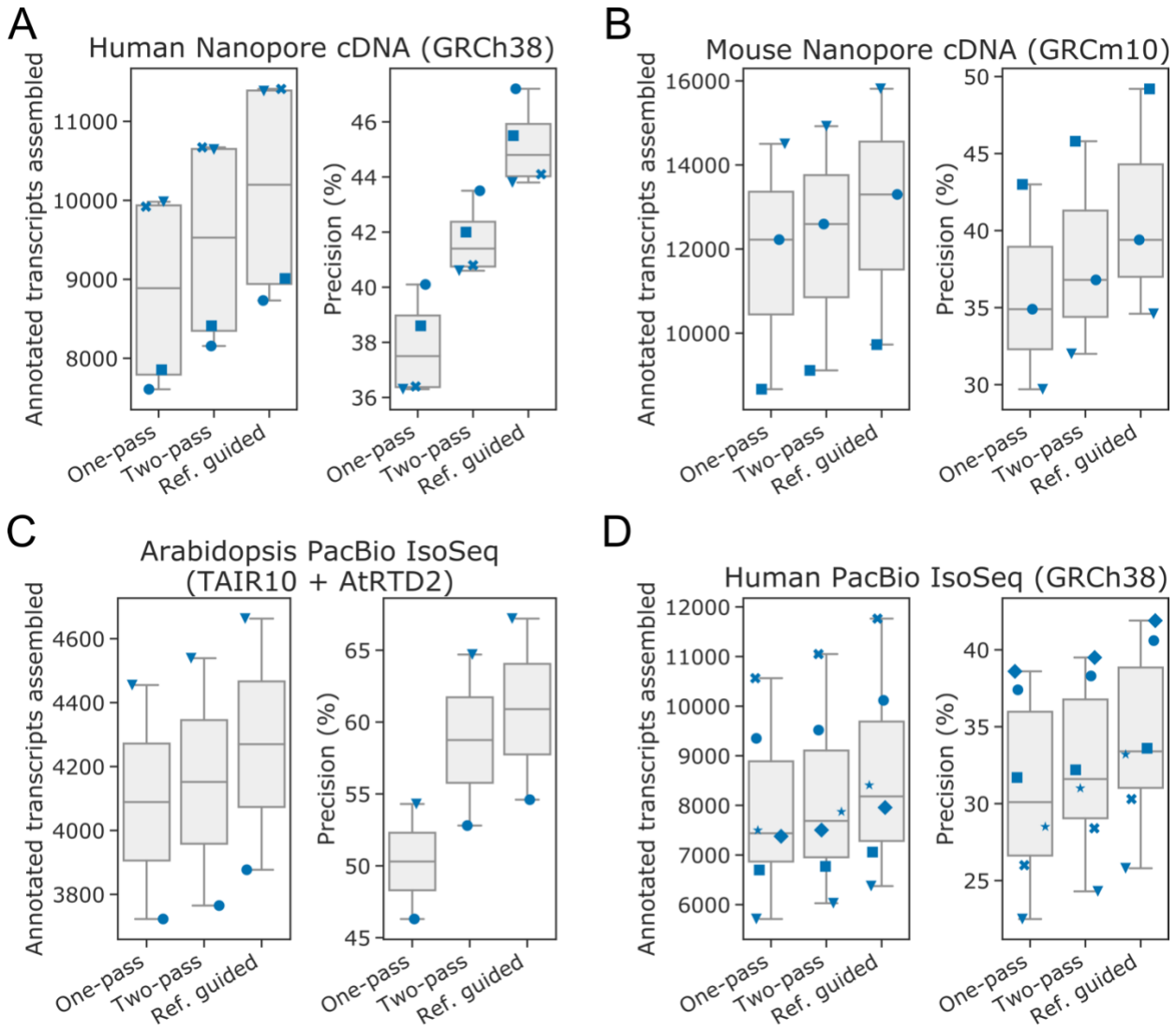


Fig. S6. Filtered two-pass alignment improves genome-guided annotation. A–D Boxplots showing the number of correct transcripts assembled (left panels) and precision of transcripts assembled (right panels) for genome-guided transcriptome assembly using StringTie2. Two-pass alignment improved the precision and number of transcripts assembled for real: **A** human nanopore cDNA; **B** mouse nanopore cDNA; **C** Arabidopsis PacBio IsoSeq; and **D** human PacBio IsoSeq data. For all boxplots, overlaid strip-plots are shown for individual samples. Each sample was assigned a unique marker so that changes in the metrics could be tracked between the one-pass, two-pass and reference-guided alignments.