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

Matthew T. Parker[1]*, Katarzyna Knop[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

*Matthew T. Parker: m.t.parker@dundee.ac.uk

Katarzyna Knop: k.knop@dundee.ac.uk

Geoffrey J. Barton: g.j.barton@dundee.ac.uk

*Gordon G. Simpson: g.g.simpson@dundee.ac.uk

*Corresponding authors

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 sequencing technologies can reduce the accuracy of intron identification. Here we apply alignment metrics and machine-learning-derived sequence information to filter spurious splice junctions from long read alignments and use the remaining junctions to guide realignment in a two-pass approach. This method, available in the software package 2passtools (https://github.com/bartongroup/2passtools), improves the accuracy of spliced alignment and transcriptome assembly for species both with and without existing high-quality annotations.

Keywords

splicing, long read sequencing, spliced alignment, RNA-seq, gene expression, transcriptome assembly, machine learning, nanopore

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).

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

74 Software tools for long and short RNAseq data analysis incorporate several approaches to address the
75 challenges presented by pre-mRNA splicing. Biologically relevant information can aid the alignment of
76 transcriptomic sequences to the genome. For example, the vast majority of eukaryotic splicing events
77 occur at introns bordered by GU and AG motifs. Making RNAseq read aligners aware of these sequence
78 features (as is the case for the commonly used spliced aligners STAR(16), HISAT2(17) and minimap2(18))
79 can significantly improve the alignment of reads at splice junctions. In addition, where genome and
80 transcriptome annotations exist, many alignment tools allow users to provide sets of correct splice
81 junctions to guide alignment(16-19). Introns containing these guide splice junctions are penalised less
82 than novel introns, resulting in fewer alignment errors. For long reads, software tools such as FLAIR(10)
83 use post-alignment correction to improve splice junction detection and quantification. Post-alignment
84 correction tools take long-read alignments and guide splice junctions from either a reference annotation
85 or a set of accurate short RNAseq reads(10). Introns from long-read alignments which are not supported
86 by the guide splice junction set are "corrected" to the nearest supported junction within a user-defined
87 range. It is unclear whether such post-alignment corrections confer any benefit over providing guide
88 splice junctions during alignment. Small errors in spliced alignment can also be corrected during
89 reference-guided transcriptome assembly. Tools such as StringTie2(6) and pinfish (Oxford Nanopore
90 Technologies) identify clusters of similarly aligned reads and correct them to the median junction
91 positions, before outputting annotations.

92 Two-pass alignment has also been used to improve splice junction detection and quantification(16, 19,
93 20). In a two-pass alignment approach, splice junctions detected in a first round of alignment are scored

less negatively in a second round, thereby allowing information sharing between alignments. This approach has been useful for short-read data, where RNA fragmentation may occur close to splice junctions during sequencing library preparation. The two-pass approach enables these short junction overhangs to be aligned to splice junctions detected in other alignments(20). Splice junctions detected in a first pass may also be filtered to remove false positives before second-pass alignment. Existing tools for splice junction filtering, such as finesplice and portcullis(21, 22), use machine learning with training on a range of junction metrics. A model is trained from high-confidence positive and negative examples from training data and then applied to classify the remaining splice junctions at the decision boundary. Splice junctions are then filtered to remove junctions predicted to be spurious. Subsequent second-pass alignment guided by these filtered junctions can then improve the accuracy of alignment(22).

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

Results and Discussion

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

For sequencing experiments designed to interpret RNA from model organisms, a set of reference splice junctions will already be available (e.g. from Ensembl). We therefore asked how providing these reference splice junctions to minimap2 to guide alignment performed compared with post-alignment correction of junctions with FLAIR(10). For this analysis, we used four nanopore DRS datasets generated from Arabidopsis seedlings(11) and four datasets generated from human cell lines(10). Several types of probable alignment error were identifiable in these data, including failure to align terminal exons and short internal exons, spurious terminal exons, and large insertions to the reference genome (Fig. 1). Because these datasets are likely to contain novel splice junctions which do not appear in reference

124 annotations, we simulated full-length reads (i.e. with no 3' bias(11)) using the Arabidopsis and human
125 reference transcriptomes, AtRTD2(23) and GRCh38(24), respectively. Simulated reads were then
126 mapped to the corresponding reference genome using minimap2(18), either with or without guidance
127 from reference splice junctions. Alignments of simulated reads were found to have similar error profiles
128 to genuine nanopore DRS read alignments (Fig. S1). Reads mapped without reference splice junctions
129 were then corrected using FLAIR with reference splice junctions.

130 Although nanopore DRS has some systematic errors in base-calling (particularly at homopolymers), the
131 majority of sequencing errors occur stochastically(25). In contrast, we found that alignment errors were
132 often repeated at similar locations in the alignments of independent reads from equivalent mRNA
133 transcripts (Fig. 1, Fig. 2A). A common alignment error at splice junctions is failure of a short exon to
134 align correctly. Instead, fragments of the exon are aligned to the ends of flanking exons, resulting in a
135 single incorrectly defined intron. A clear example of such an alignment error was detected at the short
136 (42 nt) exon 6 of Arabidopsis *FLM* (*AT1G77080*; Fig. 2A). Minimap2 uses a modified form of the Smith–
137 Waterman algorithm for performing local alignment(18, 26). This method scores alignments using
138 bonuses for matches to the reference sequence and penalties for mismatches or the opening of
139 insertions, including introns. Incorrect alignment of *FLM* exon 6 is likely to occur because the bonus for
140 aligning a short exon with sequencing errors is not sufficient to overcome the penalty for opening the
141 two flanking introns(18). Overall, we found that only 19.3% of simulated *FLM* reads aligned to the
142 correct transcript isoform. Because the sequence distance between the alignment and the genuine
143 reference splice junctions was so great, FLAIR was unable to perform post-alignment correction at *FLM*
144 exon 6, resulting in the reporting of incorrect introns (Fig. 2A). In all, 40.3% of simulated *FLM* reads were
145 aligned to the correct transcript isoform after FLAIR correction of splice junctions using the reference
146 annotation. However, providing reference splice junctions to minimap2 during alignment resulted in the
147 correct identification of *FLM* exons and introns in most cases: 92.1% of simulated *FLM* reads were
148 aligned to the correct transcript isoform. We conclude that for loci with complex splicing patterns,
149 reference-splice-junction-guided alignment performs better than post-alignment correction.

150 Without guidance from a reference annotation, we found that a median of 73.2% of Arabidopsis reads
151 and 44.4% of human reads mapped correctly to the splice junctions of the transcript they were
152 simulated from (Fig. 2B). The difference between the two organisms may be explained by biological
153 differences between the two species (e.g. in intron size, number of exons per transcript, number of
154 intronless transcripts). After post-alignment correction of splice junctions using FLAIR, the number of

155 correctly identified transcripts detected was improved (median of 87.9% and 63.6% for Arabidopsis and
156 human reads, respectively; Fig. 2B). This came at the cost of a small increase in alignment of reads to
157 incorrect reference transcript splice junctions: from a median of 1.79% to 2.62% for Arabidopsis and
158 from 3.86% to 5.45% for human (Fig. S2A). This misclassification may affect the relative quantification of
159 transcripts for some genes, with implications for differential transcript usage analysis. Reference
160 annotation-informed alignment with minimap2 performed better than FLAIR, with a median of 93.8% of
161 Arabidopsis reads and 73.2% of human reads aligning correctly at the splice junctions of the transcript
162 they were simulated from (Fig. 2B), albeit with misclassification rates of 2.61% and 5.49% respectively
163 (Fig. S2A). We conclude that there is a clear benefit to providing reference splice junctions during
164 alignment of long reads with relatively high sequence error rates, and that this is preferable to post-
165 alignment correction.

166

167 *Alignment metrics enable identification of genuine splice junctions*

168 In newly sequenced organisms, suitable reference annotations to guide alignment may not be available.
169 We therefore asked how the spliced alignment of nanopore DRS reads might be improved in the
170 absence of reference annotation. Naïve two-pass alignment has been successfully used to improve the
171 spliced alignment of short reads(20). We applied this approach with our real and simulated nanopore
172 DRS reads. Splice junctions identified by a first-pass alignment of reads were selected and used (without
173 filtering) to inform a second-pass alignment. The method was compared with reference-guided
174 alignment with minimap2, since we find this to be the gold-standard for aligning reads using information
175 from a reference annotation. We found that using the naïve two-pass approach, the median percentage
176 of simulated Arabidopsis DRS alignments which matched the splice junctions of the reference transcript
177 they were simulated from could be increased slightly from 73.2% to 75.8% (Fig. S2B). The increase was
178 similar for reads simulated from human DRS alignments: from 44.4% to 47.3% (Fig. S2B).

179 We next considered whether further improvements in two-pass alignment could be obtained by filtering
180 out likely false-positive splice junctions from first-pass alignments. This would allow us to provide more
181 refined guide junctions for second-pass alignment (Fig. 3A). A similar approach worked for short reads
182 when splice junctions were filtered by using junction metrics to train a classifier in the portcullis
183 software tool(22). By using the presence or absence of a splice junction in the reference annotation as a
184 ground truth, we considered a range of novel or previously introduced junction metrics(21, 22),

185 including junction alignment distance, supporting read count, intron motif and the presence/absence of
186 nearby splice donor and acceptor sites with higher supporting read counts (Fig. S3A-D).

187 The junction alignment distance (JAD) is defined as the minimum distance to the first mismatch,
188 insertion or deletion on either overhang of a read alignment splice junction. This metric is used by both
189 finesplice and portcullis software tools(21, 22). For the simulated nanopore DRS read alignment datasets
190 sequenced from Arabidopsis RNA, we found that 88.9% of splice junctions found in the reference
191 annotation had at least one read alignment with a JAD of 4 nt, compared with only 10.1% of
192 unannotated splice junctions (Fig. 3B). Consequently, using a threshold of at least one read with a JAD of
193 4 nt, we could identify annotated splice junctions with an F1 score of 0.902 (Fig. S3A). Despite the high
194 probability of at least some genuine unannotated splice junctions in the real Arabidopsis data(11), we
195 found that the same JAD threshold could discriminate between annotated and unannotated splice
196 junctions in real datasets to a similar degree (F1 score = 0.899). Similar results were also seen for
197 simulated human datasets, where the same JAD threshold could discriminate between spurious
198 unannotated and genuine annotated splice junctions (F1 score = 0.868). We conclude that the JAD
199 metric is a powerful discriminator of genuine splice junctions across nanopore DRS datasets from
200 different organisms.

201 Of the other metrics we tested, the read count was predictive of genuine splice junctions at a threshold
202 of >1 read (F1 score = 0.833; Fig. S3B). However, read count correlated strongly with the JAD
203 (Spearman's $\rho = 0.776$), suggesting that it does not provide more information. The presence/absence of
204 a canonical intron motif (i.e. GU/AG, GC/AG or AU/AG) had a very high recall, as 99.96% of annotated
205 introns in the simulated alignments were canonical (Fig. S3C). However, the precision was poorer (F1
206 score = 0.783). This is because in spliced alignment mode minimap2 prefers GU/AG motifs, meaning that
207 67.1% of spurious splice junctions are also aligned so as to use canonical motifs.

208 Finally, we developed a primary donor/acceptor metric similar to the one used in portcullis(22). This is
209 calculated by identifying alternative donor or acceptor sites in a 20 nt window around each
210 donor/acceptor and then determining whether they have greater read support than the current site. In
211 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
212 the tie, i.e. sites with the largest maximum per-read JAD are considered most likely to be genuine and
213 labelled as a primary site. We found that the primary donor and acceptor metrics were also predictive of
214 genuine splice junctions (F1 scores = 0.842 and 0.785 respectively). By combining the metrics to select
215 splice junctions which are both primary donors and acceptors, the F1 score can be increased to 0.918

216 (Fig. S3D). It is unclear why the primary donor score is more predictive than the primary acceptor score.
217 A possible reason is that minimap2 is more likely to produce alignment errors at the donor site of splice
218 junctions (e.g. in the case of failure to align small internal exons) or that there are more genuine
219 alternative acceptor sites than donor sites.

220 We chose to use the identified metrics to create a decision tree model, because these models are easy
221 to interpret and can be kept simple (or pruned) to prevent overfitting. A five-node tree using the JAD,
222 primary donor/acceptor and canonical intron motif metrics (Fig. 3C) was best able to predict genuine
223 Arabidopsis splice junctions (F1 score = 0.935; Fig. 3D). The same decision tree also performed well in
224 predicting genuine and spurious splice junctions from simulated human reads (F1 score = 0.934). This
225 indicates that the model might generalise across nanopore DRS datasets from different organisms,
226 despite their differences in splicing complexity.

227

228 *A combination of splice junction alignment metrics and sequence information improves authentic splice*
229 *junction identification*

230 Genuine splice junctions have sequence biases which are defined by their interactions with spliceosomal
231 uridylate-rich small nuclear RNAs(27). We next asked whether machine learning models could identify
232 genuine splice junctions from the flanking genomic sequences alone. For example, genome sequence
233 information might help identify genuine splice junctions with low read alignment coverage that fail to
234 pass the JAD filter due to stochastic sequencing errors. We therefore extracted 128 nt sequences
235 centred on unique donor and acceptor sites and used these to train LR or random forest models with
236 labels generated by the first decision tree model (Fig. 4A). Using 6-fold cross-validation, we were able to
237 train six models on 83.3% of the data each and use them to make predictions for the remaining 16.7%.
238 Using this approach, we could generate predictions for all splice junctions, with no junction being used
239 for both training and prediction from the same model. We found that LR and random forests performed
240 similarly on the data, indicating that there are few important higher-order interactions (i.e. correlated
241 sequence positions). We therefore proceeded with LR models.

242 At a prediction threshold of 0.5, the LR model overclassified positive splice junctions. False positives may
243 be sequences which could in principle act as splice junctions but do not in reality due to effects that the
244 model cannot capture. One such effect could be the presence of alternative splice junctions which are
245 preferentially processed. This is thought to occur under the “first-come-first-served” model of co-

246 transcriptional splicing(28, 29). The model is also unlikely to be able to correctly identify the intron
247 branchpoint motif because this can vary in position relative to the acceptor site(30). Nevertheless, we
248 found that the LR model approach could predict genuine splice junctions from sequence data alone with
249 comparable accuracy to the metric-based decision tree (Fig S4A-C). For example, for the simulated
250 Arabidopsis datasets, using LR on donor and acceptor sequences (with a prediction threshold of 0.5)
251 yielded an F1 score of 0.904 (Fig S4C), which was similar to the F1 score obtained with the JAD or
252 primary donor/acceptor metrics.

253 We next tested whether the information from the junction metrics and reference sequence model was
254 complementary, i.e. if a combination of the two approaches could produce an improvement in splice
255 junction prediction over each individual approach. Use of a second decision tree model, this time
256 including the JAD metric, primary donor/acceptor metrics and new LR prediction scores (Fig. 4B), further
257 increased the F1 score on splice junctions identified from simulated Arabidopsis read alignments to
258 0.954 (Fig. 4C). For splice junctions from simulated human reads, we also saw an increase in the F1 score
259 to 0.957. We conclude that an ensemble approach incorporating both junction metrics and sequence
260 information works best for detecting and filtering spurious splice junctions from alignments.

261

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

263 We next applied the two decision tree filtering methods to perform two-pass alignment of the simulated
264 reads with minimap2(18). As a positive control, we compared the results to reference-guided alignment
265 with minimap2, since this represents the best possible performance that could be achieved by two-pass
266 alignment (i.e. if the filtered splice junction set perfectly matched the reference annotation). Using
267 filtered splice junctions, the percentage of junctions identified in second-pass alignments that matched
268 annotated splice junctions could be increased over first-pass alignment and naïve two-pass alignment
269 (Fig. 5A). For example, using the simulated Arabidopsis datasets, the median percentage of read
270 alignments matching the splice junctions of the reference transcript they were simulated from increased
271 from 73.2% in the first pass, to 88.2% and 89.3% in a second pass, using the first and second decision
272 tree methods respectively (Fig. 5A). Two-pass alignment rescued the large misalignments of exon 6 seen
273 at *FLM* (Fig. S5A): overall, 86.8% of simulated *FLM* reads aligned to the correct reference transcript after
274 filtered two-pass alignment compared with 19.3% for first-pass alignments. A global improvement in
275 correct alignment was also seen in the simulated human datasets: from 44.4% in the first pass to 64.3%
276 and 65.7% for the two decision tree methods, respectively (Fig. 5A).

277 Although two-pass alignment improved the number of reads aligning to the correct transcript model, we
278 also detected a slight increase in the number of reads aligning to the wrong annotated transcript. In the
279 simulated Arabidopsis reads analysis, reads aligned using the second decision tree model performed
280 worst on this metric: 2.74% of reads aligned to the wrong isoform compared with only 1.79% of reads
281 after first-pass alignment (Fig. S5B). To assess whether such misassignment affects the quantitation of
282 transcripts, we calculated Spearman's correlation coefficient (ρ) for estimated versus known transcript
283 level read counts for the simulated data (Fig. 5B). The results indicated that, despite this misassignment,
284 two-pass aligned reads could be quantified accurately, with an overall improvement in median
285 Spearman's ρ for one-pass versus two-pass of from 0.876 to 0.916 for simulated Arabidopsis reads
286 (Fig. 5B) and from 0.778 to 0.859 for simulated human reads (Fig. 5B). However, there may be corner
287 cases where transcript misassignment could have consequences for transcript usage analysis. This
288 should be considered for experiments where quantification is important. Overall, we conclude that two-
289 pass alignment using filtered junctions can improve both the detection of correct splicing patterns and
290 the quantitation of nanopore DRS reads.

291

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

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

300 We benchmarked our methods using StringTie2 because it is reported to be faster and more accurate
301 than FLAIR on simulated nanopore DRS data(6). Using full-length reads simulated from real Arabidopsis
302 and human nanopore DRS data, we could identify the intron-chain-level precision and recall of
303 annotations assembled from reads processed using either one-pass or two-pass alignment. Here,
304 precision is defined as the percentage of assembled transcripts whose combination of introns match a
305 transcript in the reference annotation; and recall is defined as the percentage of annotated transcripts
306 for which at least one read was simulated and whose combination of introns match a transcript

307 assembled from simulated reads. We assessed reads aligned using guide splice junctions from the
308 reference annotation as a positive control.

309 For both Arabidopsis and human datasets, two-pass alignment generally produced a clear improvement
310 in both precision and recall of StringTie2 transcript assembly over first-pass alignment (Fig. 6A). Of the
311 two decision tree methods produced, decision tree 2 (using junction sequence information) performed
312 best (median F1 score was 0.699 for the Arabidopsis data and 0.629 for the human data). There was a
313 particularly large increase in precision for reference annotation-guided alignment of at least 8.7% and
314 9.6% over one-pass alignment for all Arabidopsis and human samples, respectively (Fig. 6A).

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

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

329 When we applied the same approach to the yeast *Saccharomyces cerevisiae*, the results were very
330 different (Fig. 6E). In this species, two-pass alignment resulted in a median increase of only three more
331 annotated transcripts assembled per sample and an increased number of unannotated transcripts
332 assembled, resulting in a median decrease of 0.8% in assembly precision. Splicing complexity in
333 *S. cerevisiae* is relatively low: there are only 364 annotated introns in the Ensembl R64 annotation, most
334 genes are intronless, and most introns are constitutive(33). This led to a high ratio of unannotated splice
335 junctions in first-pass alignments (the median number of junctions identified was 8,056), suggesting that
336 the vast majority of junctions in the dataset are spurious. Furthermore, most *S. cerevisiae* introns occur

337 close to mRNA 5' ends, resulting in typically short upstream exons that present challenges to alignment
338 software. Such a large ratio of spurious to genuine splice junctions is likely to affect the precision of
339 junction filtering. Notably, even when the reference annotation was used to guide alignment, precision
340 was only improved by a median of 1.9% (with a median of six more transcripts assembled correctly).
341 Intron-containing genes are generally more highly expressed (many encode ribosomal proteins) than
342 intronless genes(34). This may mean that the coverage of annotated transcripts is already good and,
343 thus, that the number of true annotated transcripts assembled cannot be much improved. This result
344 suggests that both reference annotation-guided and two-pass alignment methods have limited use for
345 genome-guided transcriptome assembly in organisms with low complexity splicing.

346 Finally, we considered whether filtered two-pass alignment could improve genome-guided annotation of
347 nanopore DRS reads derived from sequencing cDNA copies and from PacBio IsoSeq data (Fig S6A-D). To
348 assess this, we used the recommended alignment parameters for minimap2(18), but with the splice
349 junction filtering parameters that were used for nanopore DRS data. Overall, the precision and recall of
350 transcripts assembled from both nanopore cDNA and PacBio IsoSeq data for human, mouse and
351 Arabidopsis samples could be improved using two-pass alignment. For human and mouse nanopore
352 cDNA samples, two-pass alignment resulted in a median increase of 3.85% and 2.3% in assembly
353 precision, respectively, compared with one-pass alignment (median increase in annotated transcripts
354 assembled per sample of 609.5 and 420.0, respectively; Fig. S6A,B). For Arabidopsis and human PacBio
355 IsoSeq samples, two-pass alignment resulted in a median increase of 8.45% and 1.35% in assembly
356 precision, respectively, compared with one-pass alignment (median increase in annotated transcripts
357 assembled per sample of 63 and 242.5, respectively; Fig. S6C,D). We conclude that a two-pass method
358 can improve genome-guided transcript assembly of the high-error long reads produced using a range of
359 sequencing technologies.

360

361 *Two-pass alignment can aid novel splice-isoform discovery in annotated species*

362 We have shown that a two-pass approach can improve the accuracy of spliced alignment in the absence
363 of a reference annotation. However, even the most well-studied genomes are likely to be incompletely
364 annotated, and so novel splice-junction discovery which builds upon existing annotations is also
365 desirable. We therefore developed an alternative two-pass method which allows users to provide
366 reference annotations. The annotation is used to train random forest models which can then predict

367 novel splice junctions. These models replace the pre-trained decision trees used in the annotation-
368 independent method. We refer to this method hereafter as annotation-aided two-pass alignment.

369 If a reference annotation for a species is truly complete – i.e. there are no new splice-junctions to be
370 discovered, then two-pass alignment can only reduce the accuracy of alignment by introducing false-
371 positive introns into the guide splice junction set. We therefore hypothesise that two-pass alignment
372 will be useful when many genuine splice junctions are missing from the annotation, because genuine
373 novel splice junctions added to the guide junction set will outweigh false-positives that are introduced.
374 We refer to the percentage of genuine splice junctions that are unannotated as the level of annotation
375 "missingness". To test our hypothesis, we performed random subsampling of transcript isoforms in the
376 Arabidopsis reference annotation to simulate an incomplete reference at a range of missingness levels,
377 from 0.1% to 90% missing. We then performed annotation-aided two-pass alignment of the nanopore
378 DRS dataset and assessed the predictive performance on splice junctions which were absent from the
379 subsampled annotation. We found that the annotation-aided method performed best for medium
380 missingness levels. For example, in Arabidopsis DRS data, when between 25% and 66% of reference
381 isoforms were missing, the true positive rate / recall was high (minimum of 0.86), for a low false positive
382 rate (maximum 0.15) and a high precision (minimum 0.85) (Figure 7A,B). This translates to a 1.3-3.9%
383 improvement in the percentage of correctly aligned reads compared to reference-guided alignment
384 (Figure 7C). At missingness levels of less than 25%, the false positive rate increased and precision
385 decreased (Figure 7A,B). The reason for this decrease in performance is because as the reference
386 annotation nears completion, the imbalance between genuine novel splice junctions and false positives
387 caused by alignment errors increases. However, reductions in splice-junction level precision do not
388 translate to a large drop in the percentage of correctly aligned reads – at 0.1% missingness, the
389 reduction was 0.36% (Figure 7C). Furthermore, at lower levels of missingness, the recall remained high,
390 with at least 96.7% of all genuine novel splice junctions being detected. At extremely high levels of
391 annotation missingness, the recall of the two-pass filtering method begins to fall – at 90% missing, recall
392 is only 0.12 (Figure 7B). This is likely to be because when the reference is extremely incomplete, it no
393 longer represents a good training dataset, since a large proportion of junctions missing from the
394 reference will be genuine. For reference missingness levels >75%, it was therefore better to perform
395 two-pass alignment without the reference annotation (Figure 7C,D). With human RNA datasets, we
396 found that annotation-aided two-pass alignment improved the percentage of correctly aligned reads
397 when transcript isoform missingness was at least 25% (Figure 7D). This is likely due to the completeness
398 of human annotation – more junctions are found in more than one transcript isoform. We conclude that

399 annotation-aided two-pass alignment is most useful when a high-quality annotation is available, but
400 where the conditions of the experiment are expected to produce a significant number of novel splice
401 junctions.

402

403 *Two-pass alignment discovers novel splice isoforms in the Arabidopsis RNA exosome mutant hen2-2*

404 To validate the annotation-aided two-pass approach, we performed a case study with Arabidopsis using
405 the *hen2-2* mutant. HEN2 functions as an accessory protein to the nuclear RNA exosome, and is required
406 for the processing and degradation of specific classes of mRNAs and non-coding RNAs(35). As a result,
407 many RNAs, some of which contain novel splice junctions, accumulate in the *hen2-2* mutant compared
408 to wild-type. Many of these transcripts are unannotated because exosome mediated decay means that
409 they are effectively "hidden" in wild-type plants. We have previously performed Illumina RNAseq of
410 *hen2-2* mutants at relatively high depth(11). We therefore generated nanopore DRS reads from similar
411 tissue and performed annotation-aided two-pass alignment to detect novel splice junctions. Of the
412 17,521 unannotated splice junctions detected in first-pass alignment of the nanopore DRS data, only
413 20% (3548) are supported by Illumina RNAseq, and only 24% (4210) passed filtering, indicating that the
414 majority are spurious (Figure 8A). However, of those that pass filtering, 57% (2382) were supported by
415 Illumina RNAseq. This represents 67% of the 3548 unannotated junctions which were supported by both
416 nanopore DRS and Illumina RNAseq. For example, we detected novel isoforms of annotated genes, such
417 as *AT1G19396*, where use of an alternative donor site in a large intron results in a novel exonic region
418 (Figure 8B). We also detected completely unannotated transcripts, such as an antisense RNA at
419 *AT3G12140* with multiple novel splicing events (Figure 8C). We conclude that two-pass alignment is able
420 to detect genuine novel introns in well-annotated species, under less well-annotated conditions.

421

422 ## Conclusions

423 RNA sequencing is a fundamental tool for understanding what genomes really encode. Technological
424 approaches that directly sequence full-length RNA molecules substantially increase the useful
425 information that RNA sequencing can provide. The challenges that alternative splicing, in particular,
426 presents to the interpretation of high-throughput RNA sequencing data means that software
427 development needs to accompany progress in sequencing technology. In this way, knowledge gained

428 from ambitious genome sequencing programmes such as the Earth BioGenome Project, which aims to
429 characterise all eukaryotic life on Earth(36), can be maximised. We have shown that a two-pass
430 alignment approach, informed by splice junction alignment metrics and machine learning of sequence
431 features associated with splicing, can improve the accuracy of intron detection in long-read data.
432 Knowledge of existing splice junctions can also be applied to aid the discovery of novel splicing events
433 when annotations are incomplete - for example, in disease states with altered gene expression.
434 Consequently, this approach can enhance the utility and realise the potential of long-read RNA
435 sequencing.

436

437 Methods

438 *Nanopore and PacBio data*

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

452 *hen2-2 nanopore DRS data*

453 For newly sequenced nanopore DRS data, *hen2-2* seeds were sown on MS10 medium plates, stratified at
454 4°C for 2 days, germinated in a controlled environment at 22°C under 16 hr light/8 hr dark conditions
455 and harvested 14 days after transfer to 22°C. RNA isolation and nanopore direct RNA sequencing were
456 performed as described previously(11).

Preliminary data processing

Pipelines for processing of data were written using snakemake version 5.10.0(37). 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(38) 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 -x splice -L --cs=long. For nanopore cDNA and PacBio datasets the parameters used were -x splice -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 5,000 nt for yeast, to match the known intron length distributions in these organisms. For two-pass alignments using a guide splice junction set, a junction bonus (--junc-bonus) of 12 was also used, as this was found to improve the percentage of correctly aligned simulated reads when performing reference-guided annotation, compared to the default (--junc-bonus 9). Alignments of DRS reads to the reference transcriptome were performed using splicing-free parameters, namely: -k14 --for-only -L --cs=long.

Simulation of DRS reads

To provide a ground truth with a complete set of known splice sites, sequences were simulated from the reference transcriptomes, with length and error profiles matching those of real DRS reads. This was done by modelling the length, homopolymer error and other error profiles of real reads. Only primary alignments were considered. The cs tags of reads aligned to the reference transcriptome were used to recreate pairwise alignments between each read and the reference, ignoring refskips. Alignments were inverted to match the 3' → 5' sequencing direction of nanopore DRS. Aligned basecalls at reference homopolymers of ≥5 nt in length were used to build a probability model of homopolymer calls given the reference homopolymer. To prevent these error profiles being modelled multiple times, the reference homopolymer was then replaced with the aligned basecall in the pairwise alignment. Next, the altered alignment was used to create a Markov chain model of basecalled sequence given the reference sequence. For each base in the reference sequence in the alignment, the aligned portion of the query sequence was identified. The "state" of the alignment (i.e. match, mismatch, insertion or deletion) was also identified. The probability of seeing a query sequence was calculated, given the current and previous four bases of the reference and the previous four states of the alignment.

487 The reference transcriptome was also used to simulate data using these models. The number of primary
488 alignments in the real data for each reference transcript was used as the number of simulated reads per
489 transcript. To simulate basecall errors, sequences were inverted to the 3′ → 5′ direction and reads were
490 generated using Markov chain Monte Carlo simulations with the basecall model. The reference
491 sequences were prepended with a 10 nt oligo(A) sequence to mimic a short poly(A) tail so that the initial
492 state of the Markov chain was always "AAAAA" and "====" (i.e. four matches). Homopolymers in the
493 simulated read were identified and replaced with randomly selected sequences from the homopolymer
494 model. The read was then reverted to the 5′ → 3′ direction for mapping. Because we wanted to assess
495 the alignment of full-length reads, we did not model or simulate the 3′ bias, which is inherent to
496 nanopore DRS data. However, 10 nt of simulated read were subtracted from the 5′ end of reads to
497 simulate loss of signal at the end of sequencing.

498 *Post-alignment splice junction correction with FLAIR*

499 BAM files were converted to the BED12 format using bedtools(39). BED12 files were then corrected
500 using the reference GTF annotation with FLAIR correct version 1.4 and default settings(10).

501 *Junction metric calculations*

502 Splice junctions and junction metrics were extracted from aligned reads using the long form cs tag
503 produced by minimap2 version 2.17(18) using pysam version 0.15.4. The per-read JAD was calculated as
504 the length of the shorter of the two match operations immediately flanking refskip (splicing) operations.
505 Where there were mismatches or indels immediately adjacent to refskips, a JAD of zero was assigned.
506 The per-splice junction JAD was calculated as the maximum of the per-read JADs. Intron motifs were
507 extracted from cs tags. For Arabidopsis, human and mouse samples, GU/AG, GC/AG and AU/AG splice
508 junctions were all considered canonical. For yeast samples, only GU/AG splice junctions were considered
509 canonical. To calculate the primary donor/acceptor metrics, interval trees of donor and acceptor sites
510 were constructed using NCLS(40). Donors were assigned as primary donors if there were no alternative
511 donor sites within 20 nt with higher read counts. Likewise, acceptors were considered primary if there
512 were no alternative acceptors within 20 nt with higher read counts. Ties were broken using the JAD
513 metric, i.e. the splice junctions with higher JADs were assigned primary status. Where there were still
514 ties after read count and JAD comparisons, no splice junctions were assigned primary status. Splice
515 junctions extracted from four replicates of Arabidopsis or human DRS reads were used to build decision
516 tree models with scikit-learn version 0.22.1(41). A minimum depth of 4, minimum number of samples
517 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(41). For each fold, the model was used to generate out-of-bag predictions on the held-out data. The probabilities produced were then used in place of the canonical intron motif to produce the second decision tree, using a maximum depth of 6, a minimum number of samples of 1,000 and a minimum Gini impurity decrease of 0.003. Thresholds for splice scores in the tree were simplified to comprise only a high confidence threshold of 0.6 (for rescuing splice junctions failing the JAD metric threshold) and a low confidence threshold of 0.1 (for removing false positives from junctions passing the JAD metric threshold).

Annotation-aided two-pass alignment

For use cases where high quality annotations are already available, we developed an annotation-aided two-pass approach. Here, annotated junctions are provided along with read alignments. Annotated junctions are labelled as genuine. Unannotated junctions discovered in alignments are assumed to be mainly spurious. These labels are then used to train an extremely random forest model on junction metrics. Out-of-bag predictions for each junction are used as refined labels for LR models to detect splice junction sequence. A final extremely random forest model is trained on refined labels, using junction metrics and splice junction sequence scores determined by LR models. Positive examples which are not in the annotation will be a mixture of false positives and genuine novel splice-junctions. Any false negatives from the annotation are (optionally) retained.

Evaluation of splice junction models

Performance of the metrics and models was evaluated at splice junction level using the reference annotation as a ground truth. For simulated datasets, annotation is the absolute ground truth because

548 all reads are simulated using only splice junctions in the annotation. For real datasets, some "false
549 positives" are likely to be genuine splice junctions and some junctions in the reference, which appear as
550 false negatives, are actually incorrectly annotated or not expressed. Precision is defined as the number
551 of true positives divided by the total number of positive predictions by the model, i.e. true positives ÷
552 (true positives + false positives). Recall is defined as the number of true positives divided by the total
553 number of real positive examples in the dataset, i.e. true positives ÷ (true positives + false negatives).
554 The F1 score is the harmonic mean of the precision and recall.

555 *Evaluation of alignments*

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

567 *Reference-guided assembly*

568 Reference-guided transcriptome assemblies were produced using StringTie2(6) version 2.1.1 in long-
569 read mode, with otherwise default parameters.

570 *Evaluation of assemblies*

571 Reference-guided transcriptome assemblies were evaluated using the precision and recall of intron
572 chains calculated using gffcompare with default settings(42). The input reference GTF files were filtered
573 to include only transcript models for which at least one read had been simulated.

574 *Reference missingness analysis*

575 To simulate incomplete references, transcript isoforms were removed from the Araport11 (Arabidopsis)
576 and GRCh38 (human) reference annotations at rates from 0.1% to 90%. These incomplete references

were then used to perform reference guided alignment of reads simulated using the full reference annotation. Splice junctions from read alignments were then filtered using the annotation-aided method, and reads were realigned using filtered junctions as a guide. Performance on splice-junctions was measured on junctions which were not present in the annotation (i.e. training set) only. Performance at read-alignment level was measured as the change in the percentage of correctly aligned reads compared to using only the incomplete reference annotation to guide alignment.

Illumina RNAseq analysis

hen2-2 Illumina RNAseq data was downloaded from PRJEB32782. Reads were mapped to the TAIR10 genome using STAR, with a splice junction database built from the Araport11 annotation. Splice junction set intersections were identified in Python using pysam, and the visualised using upset plots.

Declarations

Availability of data and materials

Code availability

The methods used to filter splice junctions have been implemented in the "2passtools" python package, which is available on GitHub in repository https://github.com/bartongroup/2passtools. The software used to simulate reads is available on GitHub in repository https://github.com/bartongroup/yanosim. The scripts, pipelines and notebooks used to perform benchmarking and generate figures are available on GitHub in repository https://github.com/bartongroup/two_pass_alignment_pipeline.

Data availability

Basecalled and simulated nanopore DRS datasets are available from Zenodo at https://zenodo.org/record/3773729. Newly generated nanopore DRS FAST5 data has been made available on ENA under accession PRJEB41381.

601

Competing Interests

603 The authors have no competing interests to declare.

604

Funding

606 This work was funded by the University of Dundee Global Challenges Research Fund, a H2020 Marie
607 Skłodowska-Curie Actions (799300) award to Katarzyna Knop, and BBSRC awards BB/M010066/1,
608 BB/J00247X/1 and BB/M004155/1.

609

Author contributions

611 MTP developed the software and performed the data analysis. KK performed the *hen2-2* nanopore DRS
612 sequencing. MTP and GGS wrote the manuscript. All authors commented on the manuscript.

613

Acknowledgments

615 We thank James Abbott for testing the pipeline.

616

617 References

618 1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform
619 regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-6.
620 2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Reviews Genetics.
621 2019;20(11):631-56.
622 3. Mourão K, Schurch NJ, Lucoszek R, Froussios K, MacKinnon K, Duc C, et al. Detection and
623 mitigation of spurious antisense expression with RoSA. F1000Research. 2019;8.
624 4. Houseley J, Tollervey D. Apparent Non-Canonical Trans-Splicing Is Generated by Reverse
625 Transcriptase In Vitro. PLoS ONE. 2010;5(8).
626 5. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq
627 isoform quantification. BMC Genomics. 2017;18(1).
628 6. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from
629 long-read RNA-seq alignments with StringTie2. Genome Biology. 2019;20(1).
630 7. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables
631 improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology.
632 2015;33(3):290-5.
633 8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly
634 and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell
635 differentiation. Nature Biotechnology. 2010;28(5):511-5.
636 9. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA
637 sequencing on an array of nanopores. Nature Methods. 2018;15(3):201-6.
638 10. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA
639 sequencing of a human poly(A) transcriptome. Nature Methods. 2019;16(12):1297-305.
640 11. Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, et al. Nanopore direct
641 RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife.
642 2020;9.
643 12. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing
644 comes of age: applications and utilities for medical diagnostics. Nucleic Acids Research.
645 2018;46(5):2159-68.
646 13. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford
647 Nanopore sequencing. Genome Biology. 2019;20(1).
648 14. Wick RR, Judd LM, Holt KE. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with
649 deep convolutional neural networks. PLOS Computational Biology. 2018;14(11).
650 15. Dehghannasiri R, Szabo L, Salzman J, Birol I. Ambiguous splice sites distinguish circRNA and
651 linear splicing in the human genome. Bioinformatics. 2019;35(8):1263-8.
652 16. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-
653 seq aligner. Bioinformatics. 2013;29(1):15-21.
654 17. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping
655 with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019;37(8):907-15.
656 18. Li H, Birol I. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
657 2018;34(18):3094-100.
658 19. Liu B, Liu Y, Li J, Guo H, Zang T, Wang Y. deSALT: fast and accurate long transcriptomic read
659 alignment with de Bruijn graph-based index. Genome Biology. 2019;20(1).
660 20. Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. Two-pass alignment
661 improves novel splice junction quantification. Bioinformatics. 2016;32(1):43-9.

662 21. Gatto A, Torroja-Fungairiño C, Mazzarotto F, Cook SA, Barton PJR, Sánchez-Cabo F, et al.
663 FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the
664 assessment of diverse RNA-Seq alignment solutions. Nucleic Acids Research. 2014;42(8):e71-e.
665 22. Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice
666 junctions from RNA-seq with Portcullis. GigaScience. 2018;7(12).
667 23. Zhang R, Calixto Cristiane PG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, et al. A high
668 quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic
669 Acids Research. 2017;45(9):5061-73.
670 24. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38
671 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.
672 Genome Research. 2017;27(5):849-64.
673 25. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for
674 improving nanopore sequencing read accuracy. Genome Biology. 2018;19(1).
675 26. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of
676 Molecular Biology. 1981;147(1):195-7.
677 27. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-
678 site analysis using comparative genomics. Nucleic Acids Research. 2006;34(14):3955-67.
679 28. Carrillo Oesterreich F, Herzel L, Straube K, Hujer K, Howard J, Neugebauer Karla M. Splicing of
680 Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell. 2016;165(2):372-81.
681 29. Reimer KA, Mimoso C, Adelman K, Neugebauer KM. Rapid and Efficient Co-Transcriptional
682 Splicing Enhances Mammalian Gene Expression. bioRxiv. 2020.
683 30. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide
684 discovery of human splicing branchpoints. Genome Research. 2015;25(2):290-303.
685 31. Kuo RI, Cheng Y, Smith J, Archibald AL, Burt DW. Illuminating the dark side of the human
686 transcriptome with TAMA Iso-Seq analysis. bioRxiv. 2019.
687 32. Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, et al. Transcriptome
688 profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. Scientific Reports.
689 2019;9(1).
690 33. Spingola M, Grate L, Haussler D, Ares M. Genome-wide bioinformatic and molecular analysis of
691 introns in Saccharomyces cerevisiae. Rna. 1999;5(2):221-34.
692 34. Ares M, Grate L, Pauling MH. A handful of intron-containing genes produces the lion's share of
693 yeast mRNA. Rna. 1999;5(9):1138-9.
694 35. Chen X, Lange H, Zuber H, Sement FM, Chicher J, Kuhn L, et al. The RNA Helicases AtMTR4 and
695 HEN2 Target Specific Subsets of Nuclear Transcripts for Degradation by the Nuclear Exosome in
696 Arabidopsis thaliana. PLoS Genetics. 2014;10(8).
697 36. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome
698 Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences.
699 2018;115(17):4325-33.
700 37. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics.
701 2012;28(19):2520-2.
702 38. Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.
703 Nature. 2000;408(6814):796-815.
704 39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
705 Bioinformatics. 2010;26(6):841-2.
706 40. Stovner EB, Sætrom P, Hancock J. PyRanges: efficient comparison of genomic intervals in
707 Python. Bioinformatics. 2019.
708 41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine
709 Learning in Python. J Mach Learn Res. 2011;12:2825-30.

710 42. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-
711 seq experiments with HISAT, StringTie and Ballgown. Nature Protocols. 2016;11(9):1650-67.

712

713 ## Figure legends

714 **Fig. 1. Assessment of alignment errors in nanopore DRS datasets.** Nanopore DRS read alignments at
715 Arabidopsis *AT5G05010* locus with different types of alignment error presented. Read alignments are
716 shown in dark blue, with soft-clipped (unaligned) ends shown in light blue. Mismatches and indels of
717 <30 nt are not shown. Insertions to the reference of > 30 nt are shown as orange carets.

718 **Fig. 2. Improved spliced alignment of simulated reads using annotation-guided alignment. A** Reference-
719 guided alignment improves the identification of small exons in nanopore DRS reads. Gene track showing
720 the alignment of a sample of simulated nanopore DRS reads at the Arabidopsis *FLM* gene. AtRTD2
721 reference annotation, from which reads were simulated, is shown on top, with unguided minimap2
722 alignments, FLAIR correction of unguided minimap2 alignments and reference-guided minimap2
723 alignments shown below. Only reads where exon 6 failed to align in the initial unguided alignment are
724 shown. Each read alignment is coloured based on the reference transcript it was simulated from, and
725 reads are shown in the same order within each alignment method group. Mismatches and indels are not
726 shown. **B** Reference-guided alignment improves the identification of correct transcripts globally. Boxplots
727 with overlaid strip-plots showing the percentage of alignments which map exactly to the splice junctions
728 of the transcript from which they were simulated, for unguided minimap2 alignments, FLAIR correction
729 of unguided minimap2 alignments using reference annotation, and reference annotation-guided
730 minimap2 alignments. Reads simulated from intronless transcripts which map correctly without splicing
731 were not included in percentage calculations. Reads were simulated from Arabidopsis (left) and human
732 (right) nanopore DRS data aligned to the AtRTD2 and GRCh38 reference transcriptomes, respectively.

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

741 **Fig. 4. Machine learned sequence information improves identification of genuine splice junctions. A**
742 Outline of the LR model training process. Sequences from splice junctions were extracted from the

743 reference genome and used as training data (i.e. explanatory variables). Training labels (i.e. the response
744 variable) were generated by the first decision tree model. Independent models were trained for 5' donor
745 and 3' acceptor sites and cross-validation used to generate out-of-bag predictions for all sites. **B** Flowchart
746 visualisation of the second decision tree model. Nodes (decisions) and leaves (outcomes) are coloured
747 based on the relative ratio of real and spurious splice junctions. **C** Confusion matrix showing the ratios of
748 correct and incorrect predictions of the second decision tree model on splice junctions extracted from
749 simulated Arabidopsis read alignments.

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

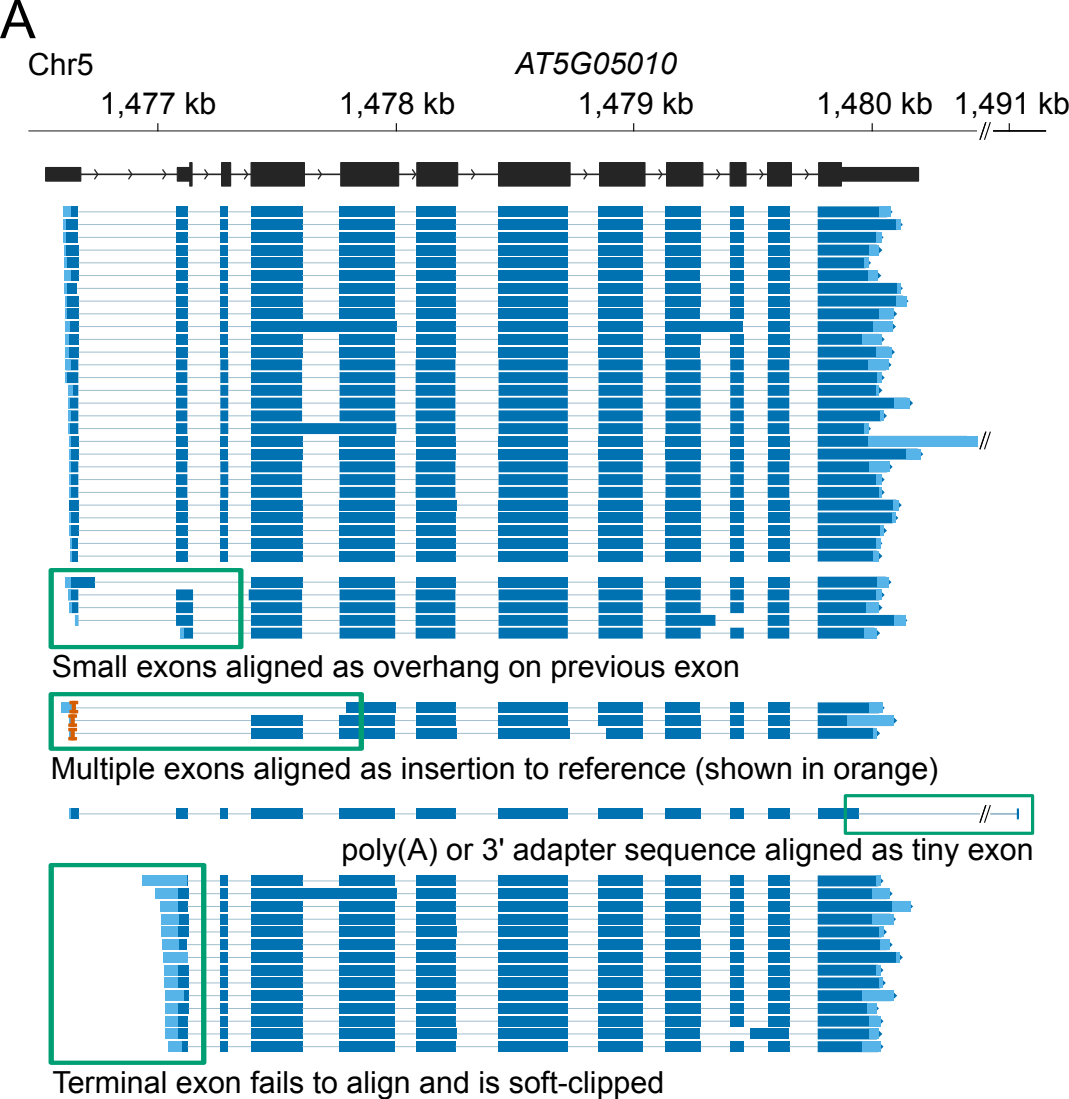
760 **Fig. 6. Filtered two-pass alignment improves genome-guided annotation. A** Scatterplot showing
761 precision against recall for intron chains in genome-guided transcriptome annotations generated from
762 alignments using StringTie2. Precision and recall scores were calculated against reference annotations
763 filtered to include only transcripts for which at least one read was simulated. Reads were simulated from
764 Arabidopsis (left) and human (right) nanopore DRS data aligned to the AtRTD2 and GRCh38 reference
765 transcriptomes, respectively. **B–E** Stripplots with box-and-whiskers showing the number of correct
766 transcripts assembled (left panels) and precision of transcripts assembled (right panels) for genome-
767 guided transcriptome assembly using StringTie2. Two-pass alignment improved the precision and number
768 of transcripts assembled for real nanopore DRS data for **B** Arabidopsis, **C** human, **D** mouse and **E** yeast.
769 For all boxplots, overlaid strip-plots are shown for individual samples. Each sample was assigned a unique
770 marker so that the changes in each sample could be tracked between the one-pass, two-pass and
771 reference-guided alignments. Box-and-whiskers not shown for samples with less than 4 data points. Y
772 limits vary between figures since within-figure (i.e. same species and sequencing technology) comparison
773 is more important than between-figure comparisons.

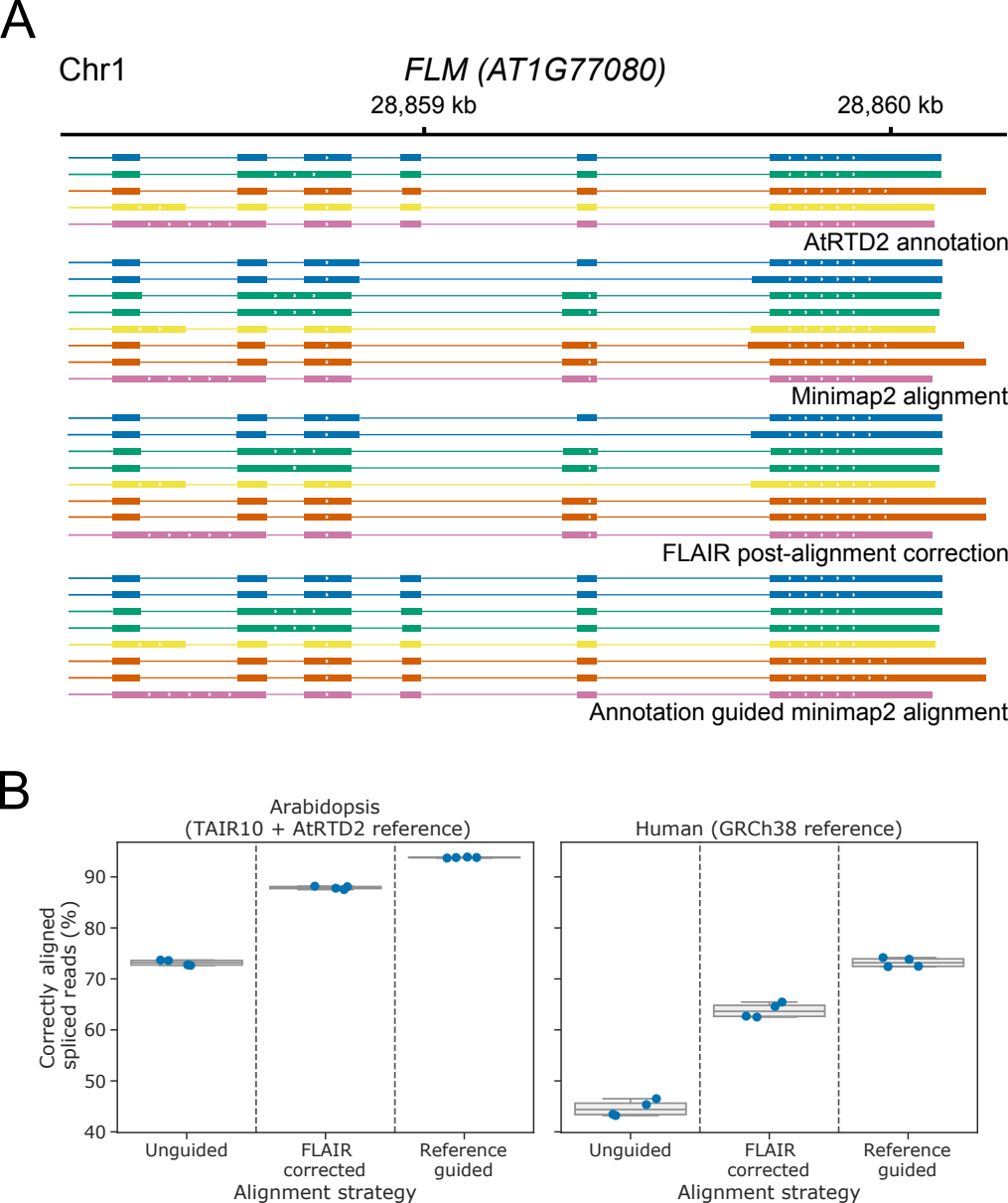
774 **Fig. 7. Annotation-aided two-pass alignment rescues missing splice junctions. A** ROC scatterplot and **B**
775 precision/recall scatterplot showing true positive rate and false positive rate of novel splice junction
776 classification in simulated Arabidopsis read alignments, at different rates reference annotation
777 missingness. Annotated transcript isoforms were subsampled to simulate incomplete reference
778 annotations, and these were used to inform annotation-aided two-pass alignment. **C-D** Line plots showing
779 the improvement in the percentage of correctly aligned reads using two-pass alignment compared to
780 reference-guided alignment at different reference annotation missingness rates for **C** Arabidopsis and **D**
781 humans, respectively. Blue line shows improvement compared to reads aligned using two-pass method
782 only. Orange line shows improvement compared to reads aligned using reference-annotation in first-pass,
783 followed by annotation-aided junction filtering and second pass alignment. Shaded regions represent 95%
784 confidence intervals.

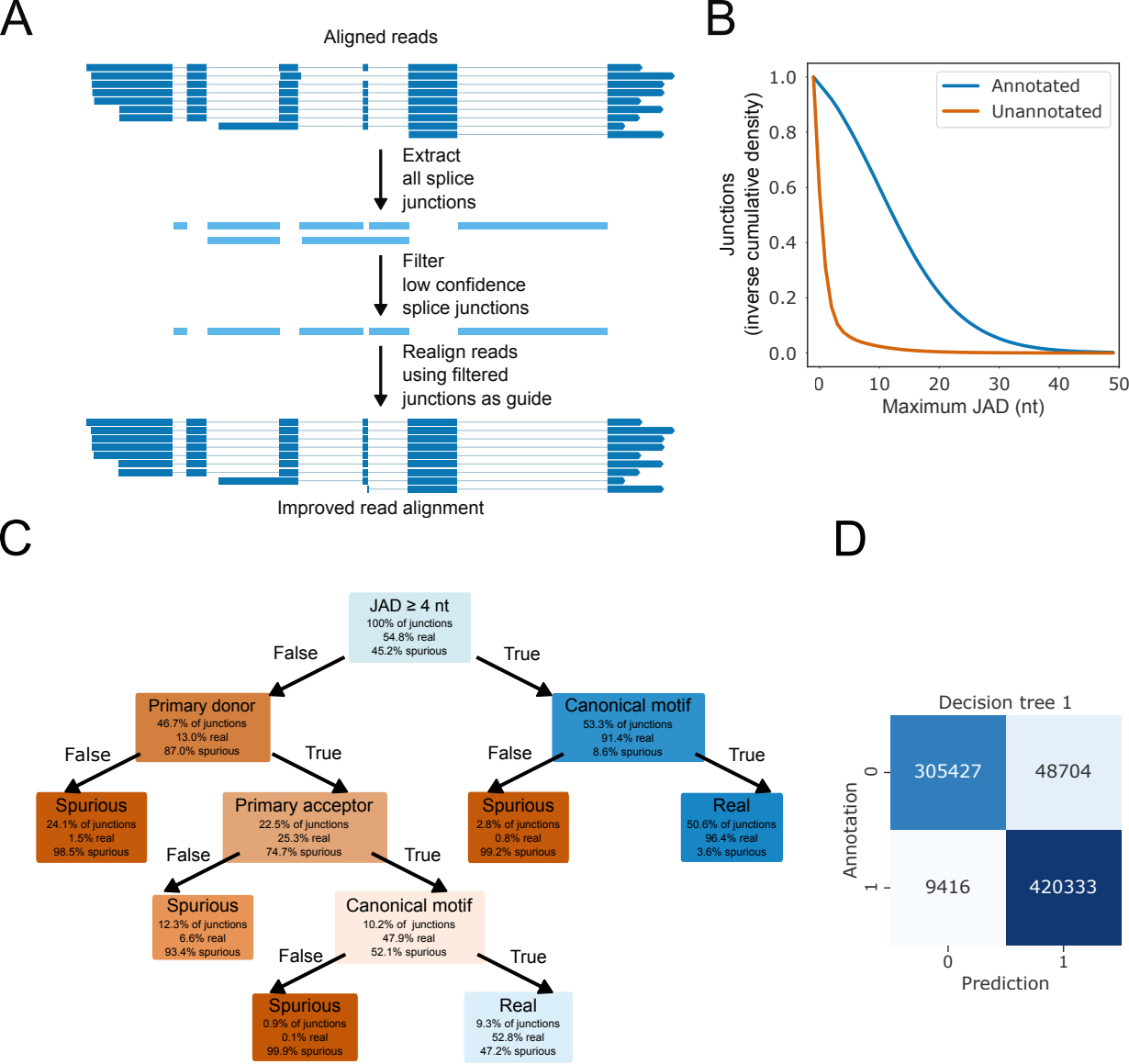
785 **Fig. 8. Annotation-aided two-pass alignment identifies novel splice isoforms in *hen2-2* mutants. A** Upset
786 plot showing the intersection of splice junctions detected using nanopore DRS or Illumina RNAseq, and
787 presence in the AtRTD2 annotation. Horizontal bars show the overall number of junctions detected using
788 each technology/annotation, whilst stacked vertical bars represent set intersections. For nanopore DRS
789 data, splice junctions with one or more supporting read alignment are shown. For Illumina RNAseq, splice
790 junctions with ten or more supporting read alignments are shown. Nanopore DRS junctions which are
791 classified as spurious by the two-pass filtering method are labelled in blue, whilst junctions which are
792 classified as genuine are labelled in orange. Set intersection bars not including nanopore DRS are shown
793 in grey. **B-C** Gene track showing novel splice isoforms detected at **B** *AT1G19396* and **C** *AT3G12140* in *hen2-*
794 *2* nanopore DRS data. AtRTD2 annotation is shown in black. Nanopore DRS reads are shown in blue
795 (positive strand) or light blue (negative strand). Novel splice junctions are shown in orange.

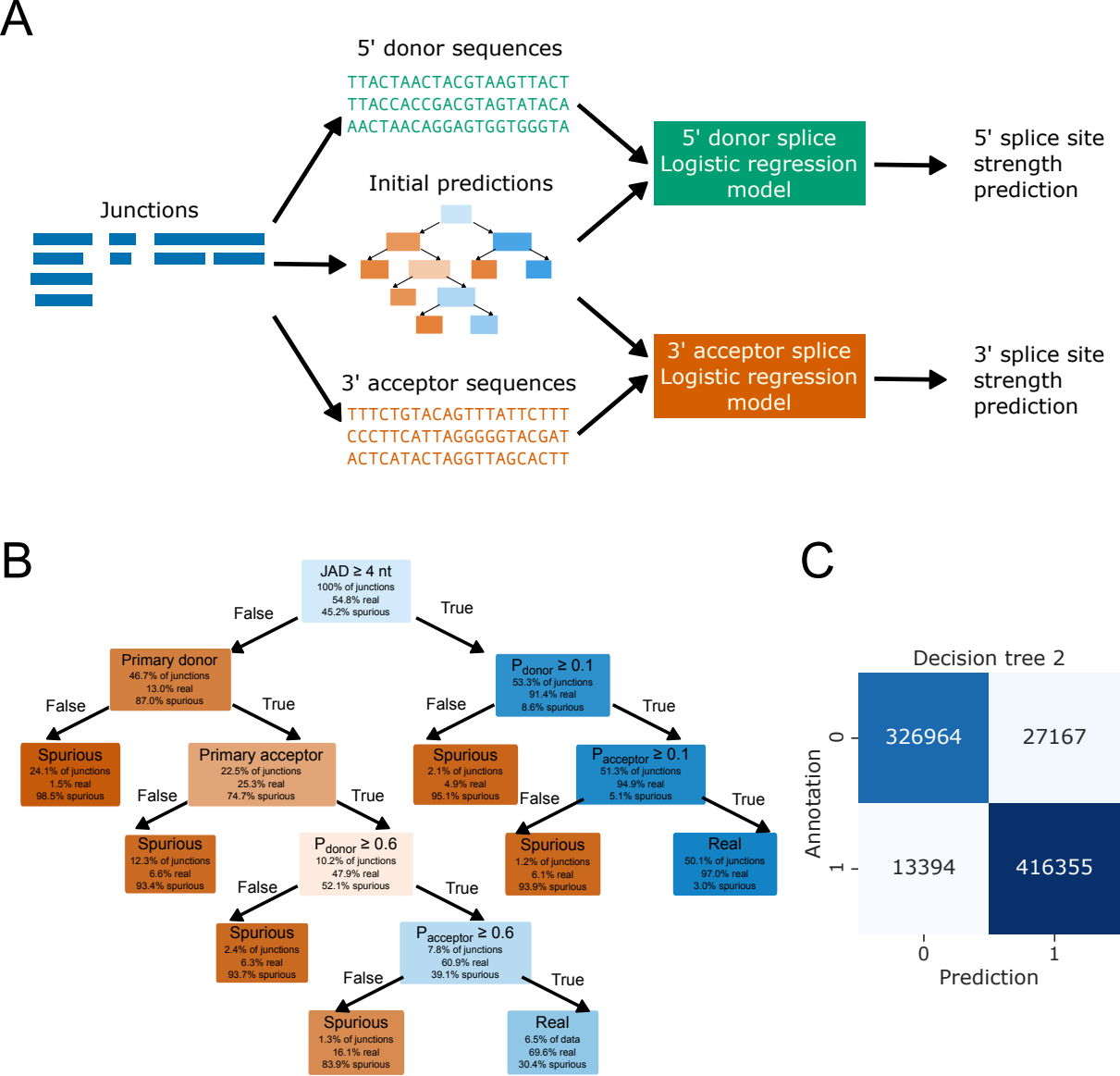
796

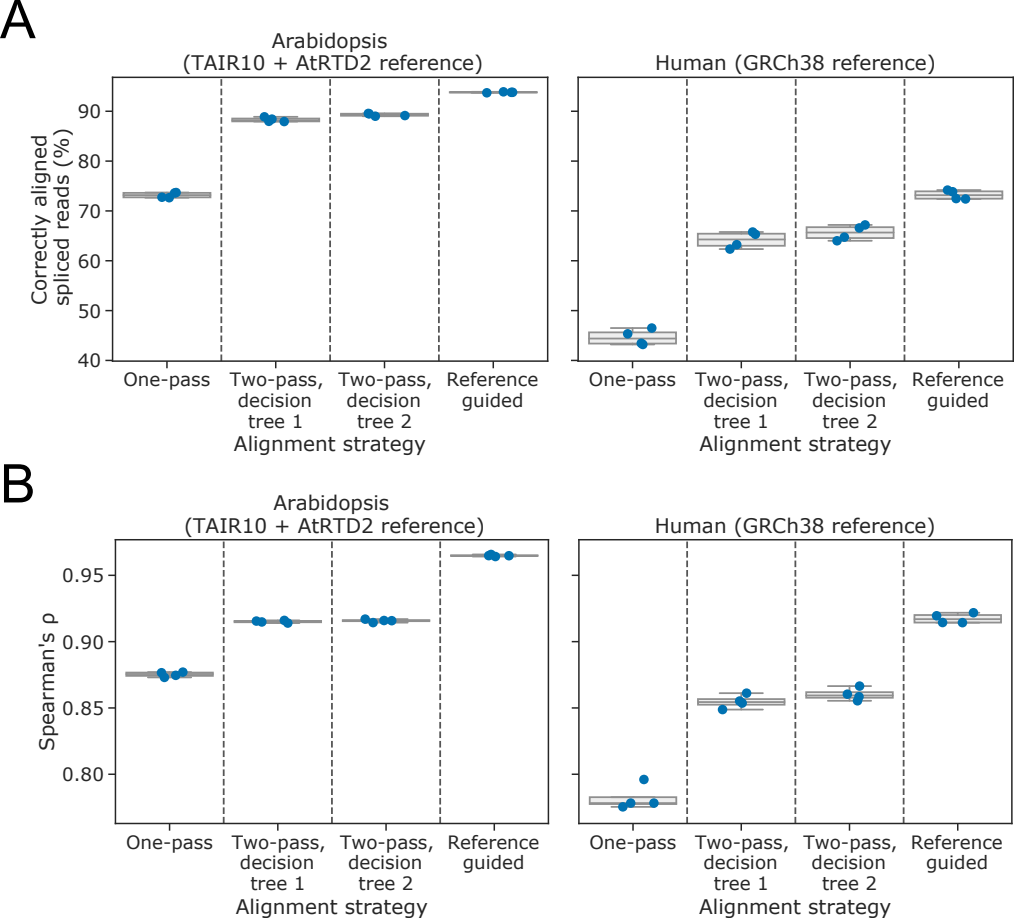
A

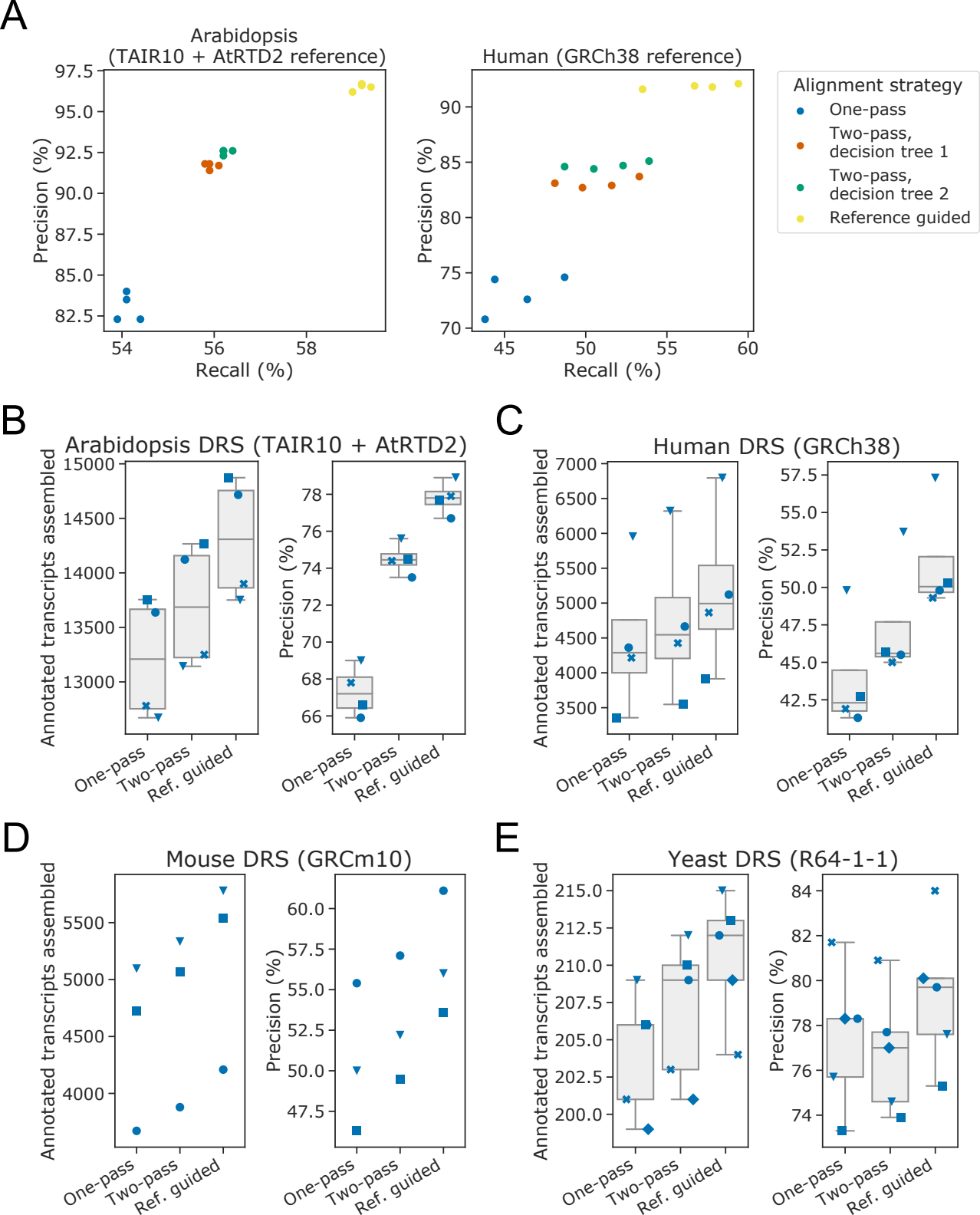

Chr5
AT5G05010
1,477 kb
1,478 kb
1,479 kb
1,480 kb
1,491 kb
Small exons aligned as overhang on previous exon
Multiple exons aligned as insertion to reference (shown in orange)
poly(A) or 3' adapter sequence aligned as tiny exon
Terminal exon fails to align and is soft-clipped

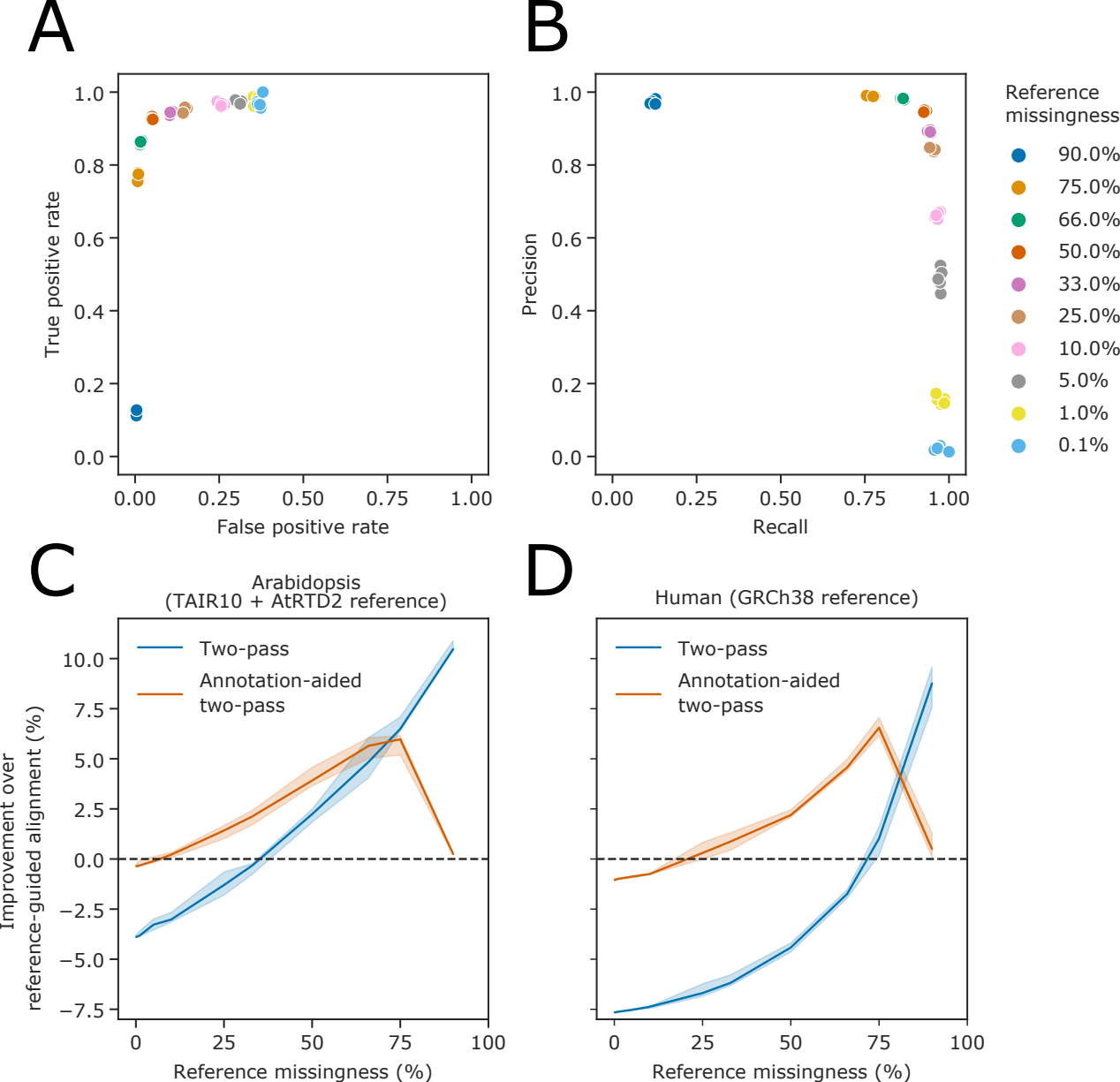

A
Chr1
FLM (AT1G77080)
28,859 kb
28,860 kb
AtRTD2 annotation
Minimap2 alignment
FLAIR post-alignment correction
Annotation guided minimap2 alignment
B
Arabidopsis
(TAIR10 + AtRTD2 reference)
Human (GRCh38 reference)
Correctly aligned
spliced reads (%)
90
80
70
60
50
40
Unguided
FLAIR
corrected
Reference
guided
Alignment strategy

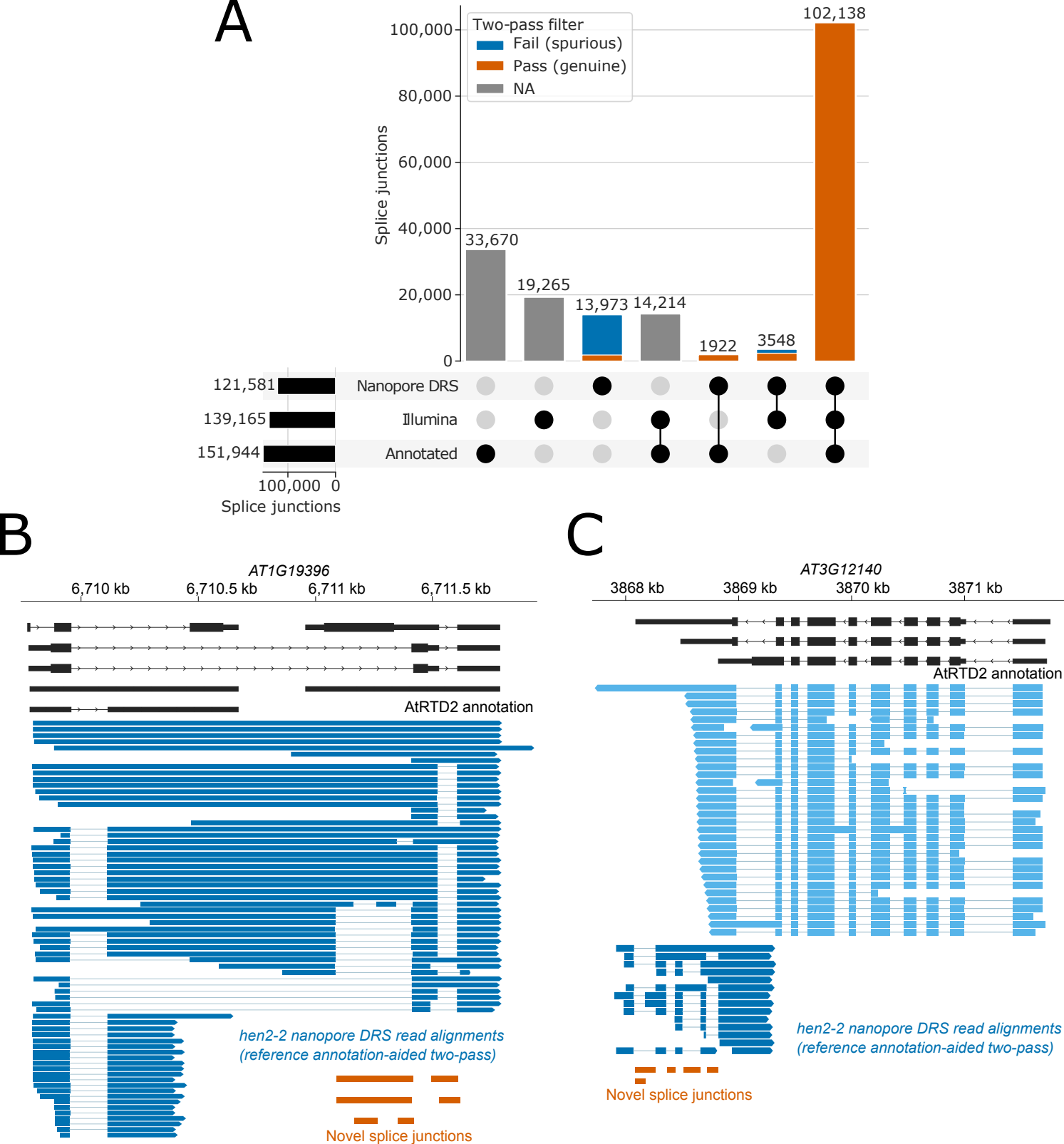

A
Aligned reads
Extract
all splice
junctions
Filter
low confidence
splice junctions
Realign reads
using filtered
junctions as guide
Improved read alignment
B
Annotated
Unannotated
Junctions
(inverse cumulative density)
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Maximum JAD (nt)
C
JAD ≥ 4 nt
100% of junctions
54.8% real
45.2% spurious
False
True
Primary donor
46.7% of junctions
13.0% real
87.0% spurious
Canonical motif
53.3% of junctions
91.4% real
8.6% spurious
Spurious
24.1% of junctions
1.5% real
98.5% spurious
Primary acceptor
22.5% of junctions
25.3% real
74.7% spurious
Spurious
2.8% of junctions
0.8% real
99.2% spurious
Real
50.6% of junctions
96.4% real
3.6% spurious
Spurious
12.3% of junctions
6.6% real
93.4% spurious
Canonical motif
10.2% of junctions
47.9% real
52.1% spurious
Spurious
0.9% of junctions
0.1% real
99.9% spurious
Real
9.3% of junctions
52.8% real
47.2% spurious
D
Decision tree 1
Annotation
0
1
305427
48704
9416
420333
0
1
Prediction


A
5' donor sequences
TTACTAACTACGTAAGTTACT
TTACCACCGACGTAGTATACA
AACTAACAGGAGTGGTGGGTA
Junctions
Initial predictions
3' acceptor sequences
TTTCTGTACAGTTTATTCTTT
CCCTTCATTAGGGGGTACGAT
ACTCATACTAGGTTAGCACTT
5' donor splice Logistic regression model
3' acceptor splice Logistic regression model
5' splice site strength prediction
3' splice site strength prediction
B
JAD ≥ 4 nt
100% of junctions
54.8% real
45.2% spurious
False
True
Primary donor
46.7% of junctions
13.0% real
87.0% spurious
P donor ≥ 0.1
53.3% of junctions
91.4% real
8.6% spurious
Spurious
24.1% of junctions
1.5% real
98.5% spurious
Primary acceptor
22.5% of junctions
25.3% real
74.7% spurious
Spurious
2.1% of junctions
4.9% real
95.1% spurious
P acceptor ≥ 0.1
51.3% of junctions
94.9% real
5.1% spurious
Spurious
12.3% of junctions
6.6% real
93.4% spurious
P donor ≥ 0.6
10.2% of junctions
47.9% real
52.1% spurious
Spurious
1.2% of junctions
6.1% real
93.9% spurious
Real
50.1% of junctions
97.0% real
3.0% spurious
Spurious
2.4% of junctions
6.3% real
93.7% spurious
P acceptor ≥ 0.6
7.8% of junctions
60.9% real
39.1% spurious
Spurious
1.3% of junctions
16.1% real
83.9% spurious
Real
6.5% of data
69.6% real
30.4% spurious
C
Decision tree 2
Annotation
Prediction
0
1
326964
27167
13394
416355


A
Arabidopsis
(TAIR10 + AtRTD2 reference)
Human (GRCh38 reference)
Correctly aligned
spliced reads (%)
90
80
70
60
50
40
One-pass
Two-pass,
decision
tree 1
Two-pass,
decision
tree 2
Reference
guided
Alignment strategy
B
Arabidopsis
(TAIR10 + AtRTD2 reference)
Human (GRCh38 reference)
Spearman's ρ
0.95
0.90
0.85
0.80
One-pass
Two-pass,
decision
tree 1
Two-pass,
decision
tree 2
Reference
guided
Alignment strategy


A
Arabidopsis
(TAIR10 + AtRTD2 reference)
Human (GRCh38 reference)
Precision (%)
Recall (%)
Alignment strategy
One-pass
Two-pass, decision tree 1
Two-pass, decision tree 2
Reference guided
B
Arabidopsis DRS (TAIR10 + AtRTD2)
C
Human DRS (GRCh38)
D
Mouse DRS (GRCm10)
E
Yeast DRS (R64-1-1)
Annotated transcripts assembled
Precision (%)
One-pass
Two-pass
Ref. guided


A
True positive rate
False positive rate
B
Precision
Recall
Reference missingness
90.0%
75.0%
66.0%
50.0%
33.0%
25.0%
10.0%
5.0%
1.0%
0.1%
C
Arabidopsis
(TAIR10 + AtRTD2 reference)
Two-pass
Annotation-aided two-pass
Improvement over reference-guided alignment (%)
Reference missingness (%)
D
Human (GRCh38 reference)
Two-pass
Annotation-aided two-pass
Reference missingness (%)


A
Two-pass filter
Fail (spurious)
Pass (genuine)
NA
Splice junctions
100,000
80,000
60,000
40,000
20,000
0
33,670
19,265
13,973
14,214
1922
3548
102,138
121,581
139,165
151,944
Nanopore DRS
Illumina
Annotated
100,000 0
Splice junctions
B
AT1G19396
6,710 kb
6,710.5 kb
6,711 kb
6,711.5 kb
AtRTD2 annotation
hen2-2 nanopore DRS read alignments
(reference annotation-aided two-pass)
Novel splice junctions
C
AT3G12140
3868 kb
3869 kb
3870 kb
3871 kb
AtRTD2 annotation
hen2-2 nanopore DRS read alignments
(reference annotation-aided two-pass)
Novel splice junctions

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

Supplemental Figures

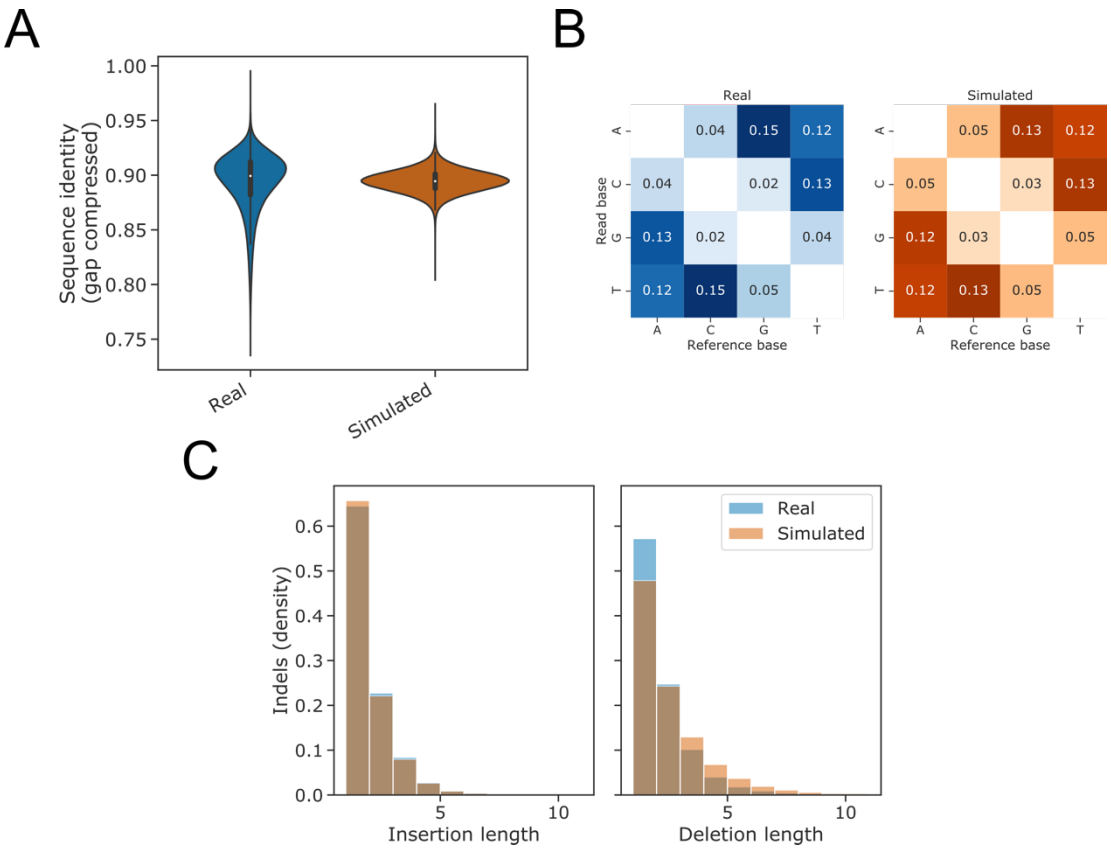


Fig. S1. Simulation of nanopore DRS read alignments. A Violin plot showing the distribution of sequence identity scores for real and simulated Arabidopsis nanopore DRS reads. Simulated reads match the median sequence identity of real reads, although they do not capture the tails of high- and low-quality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

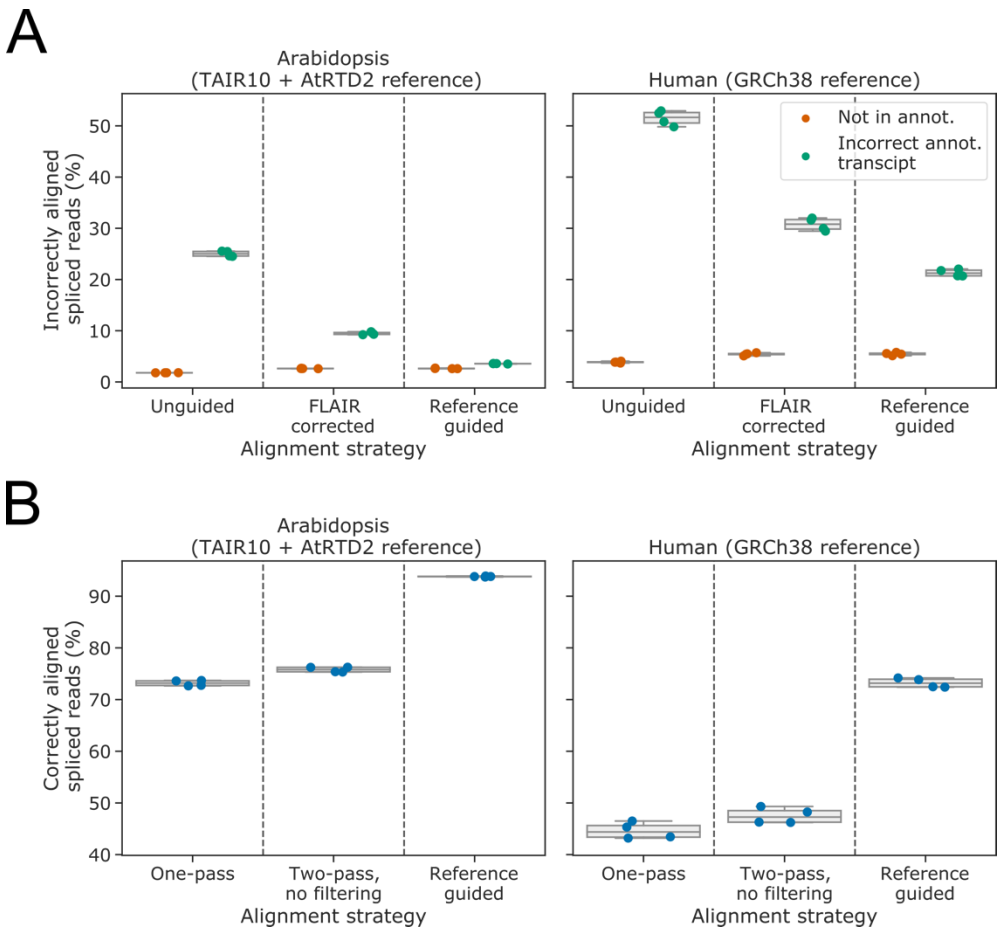


Fig. S2. Annotation-guided alignment improves spliced alignment of simulated reads. **A** Boxplots with overlaid strip-plot showing the percentage of alignments which do not map correctly to the splice junctions of the transcript from which they were simulated, for one-pass unguided minimap2 alignments, FLAIR-corrected alignments and reference annotation-guided minimap2 alignments. Reads that align to unannotated splice junctions or combinations of junctions ("Not in annot.") 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.

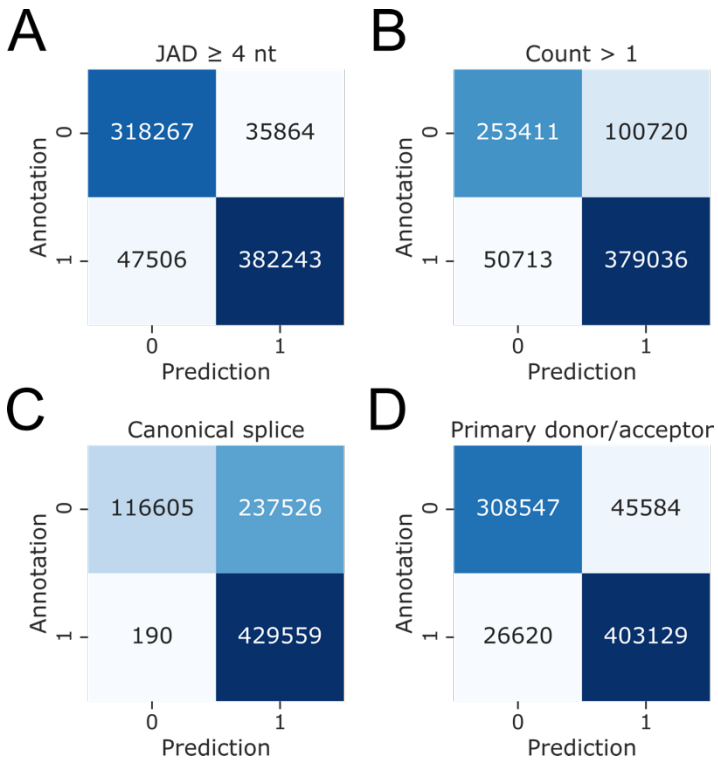


Fig. S3. Junction metrics can identify genuine splice junctions. A-D Confusion matrices showing the ratios of correct and incorrect predictions using: **A** a JAD threshold of 4 nt; **B** a count threshold of 1 nt; **C** the presence of a canonical U2 GU/AG, U12 GC/AG or U12 AU/AG intron motif; and **D** 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.

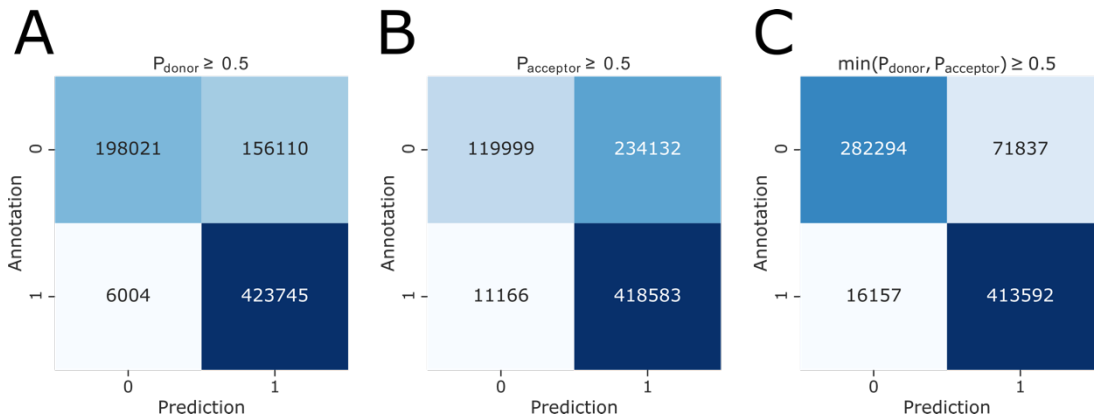


Fig. S4. Machine-learned sequence information can identify genuine splice junctions. A-C Confusion matrices showing the ratios of correct and incorrect predictions using: **A** an LR prediction threshold of 0.5 for splice site strength predictions made on donor site sequences; **B** an LR prediction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

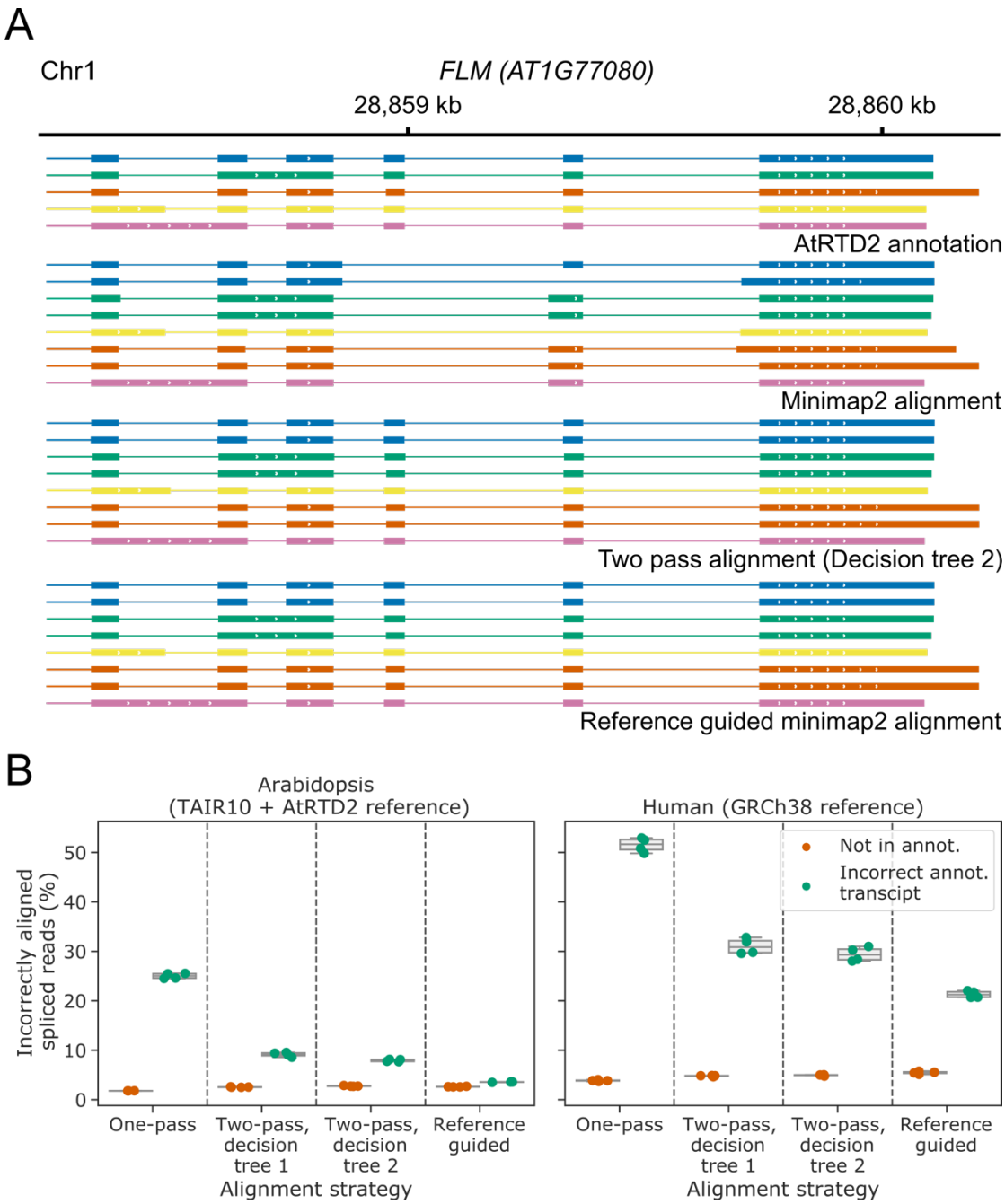


Fig. S5. Filtered two-pass alignment improves the identification and quantification of correct transcripts without a reference annotation. A Gene track showing alignment of a sample of simulated nanopore DRS reads at the Arabidopsis *FLM* gene. The AtRTD2 reference annotation, from which reads were simulated, 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 within each alignment method group. Mismatches and indels are not shown. **B** Boxplots with overlaid

strip-plots 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, two-pass alignment with decision trees one and two, 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 annotated combinations of splice junctions which they were not simulated from 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.

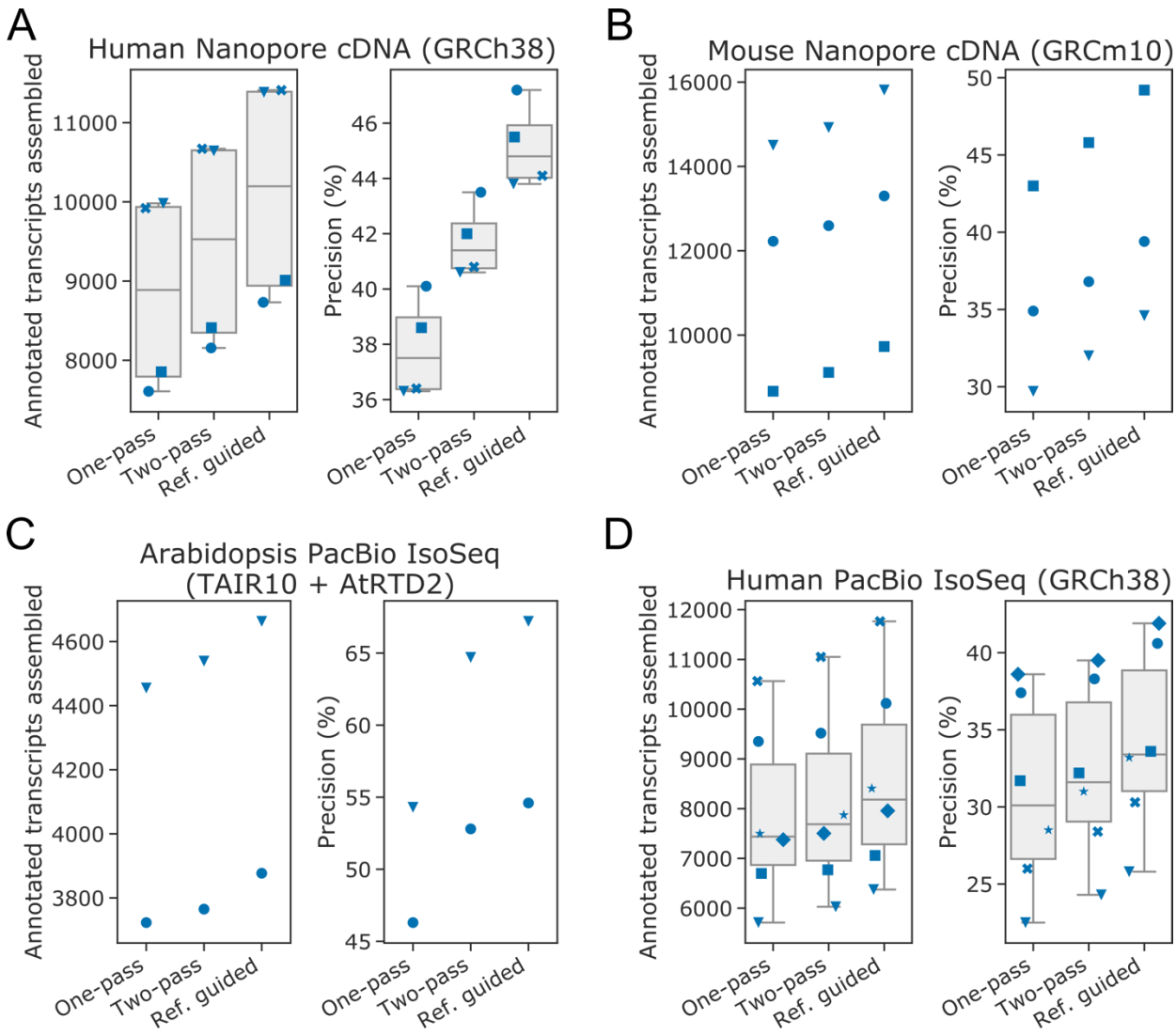


Fig. S6. Filtered two-pass alignment improves genome-guided annotation. A–D Stripplots with box-and-whiskers showing the number of correct transcripts assembled (left panels) and precision of transcripts assembled (right panels) for genome-guided transcriptome assembly using StringTie2. Two-pass alignment improved the precision and number of transcripts assembled from **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. Box-and-whiskers not shown for samples with less than 4 data points.