SpliceGrapherXT: From Splice Graphs to Transcripts Using RNA-Seq

Mark F. Rogers, Christina Boucher, and Asa Ben-Hur
Department of Computer Science
1873 Campus Delivery
Fort Collins, CO 80523
rogersma@cs.colostate.edu, cboucher@cs.colostate.edu, asa@cs.colostate.edu

ABSTRACT

Predicting the structure of genes from RNA-Seq data remains a significant challenge in bioinformatics. Although the amount of data available for analysis is growing at an accelerating rate, the capability to leverage these data to construct complete gene models remains elusive. In addition, the tools that predict novel transcripts exhibit poor accuracy. We present a novel approach to predicting splice graphs from RNA-Seq data that uses patterns of acceptor and donor sites to recognize when novel exons can be predicted unequivocally. This simple approach achieves much higher precision and higher recall than methods like Cufflinks or IsoLasso when predicting novel exons from real and simulated data. The ambiguities that arise from RNA-Seq data can preclude making decisive predictions, so we use a realignment procedure that can predict additional novel exons while maintaining high precision. We show that these accurate splice graph predictions provide a suitable basis for making accurate transcript predictions using tools such as IsoLasso and PSGInfer. Using both real and simulated data, we show that this integrated method predicts transcripts with higher recall and precision than using these other tools alone, and in comparison to Cufflinks. SpliceGrapherXT is available from the SpliceGrapher web page at http://SpliceGrapher.sf.net.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

bioinformatics, RNA-seq, transcriptomics, splice graph

1. INTRODUCTION

Precursor mRNA in eukaryotes undergoes extensive alternative splicing that results in increased transcriptome complexity. This expands the protein-coding capacity of a genome and provides mechanisms for regulating gene expression [30, 20, 11]. Genome-wide sequencing of mRNA (RNA-Seq) is becoming widely used, and is allowing researchers to probe the transcriptome on an unprecedented scale [24, 34, 5, 8, 26]. Although it is relatively inexpensive to obtain hundreds of millions of reads using high-throughput sequencing technologies, transcript prediction methods still suffer from low recall and precision.

The reads produced by high-throughput methods present challenges for analysis. RNA-Seq studies usually rely on two kinds of alignment algorithms to map reads to a reference genome: ungapped alignment algorithms that map whole reads to a genome [16, 15, 13] and spliced alignment algorithms that map reads across exon-intron junctions [2, 36, 31, 10, 33]. These algorithms effectively reverse-engineer the mRNA splicing process, providing evidence for novel exons and splice junctions. However, using this evidence to make predictions remains challenging.

To date, most methods for predicting mRNA splicing have focused on predicting complete transcripts and their relative expression levels simultaneously [18, 7, 25, 4, 17, 23, 27]. For example, IsoLasso estimates isoform expression by simultaneously minimizing abundance error (the difference between the expected and actual number of reads assigned to an isoform) and the number of expressed isoforms [18]. Comparisons on both simulated and real data show that the best of these approaches yields low accuracy when confronted with ambiguous and noisy RNA-Seq data [23].

By focusing on whole transcript prediction instead of novel exon prediction, these methods may not predict novel exons correctly in the presence of ambiguous RNA-Seq evidence. A complementary approach is to focus on predicting *splice graphs* that capture in a single structure the ways in which exons may be combined for a gene [9, 28]. The SpliceGrapher method [28] uses a conservative approach that may denote evidence as unresolvable to avoid making spurious predictions. Previous work with SpliceGrapher showed that it can predict novel exons with higher fidelity than methods like Cufflinks [28]. With SpliceGrapherXT we present a novel, comprehensive method based on simple patterns of acceptor and donor sites that vastly simplifies our approach to splice-graph prediction. Our SpliceGrapherXT method predicts splice graphs from RNA-Seq evidence when exons can be resolved easily; stores details about regions where evidence is ambiguous, and uses a realignment procedure to resolve exons supported by the ambiguous evidence. Combined, these procedures substantially improve Splice-

Grapher's recall and enable it to maintain a high level of precision on both simulated and real data.

Splice graph prediction methods such as SpliceGrapherXT can make accurate predictions for a gene's splicing activity. A relevant question, then, is whether we can use these predictions to improve the accuracy of transcript prediction methods. Accordingly, we present preliminary results for two methods that can convert splice graph predictions into complete transcript predictions. Our results show the potential for these methods to predict accurately the transcripts recapitulated in an RNA-Seq data set. In this work we compare SpliceGrapherXT with the original SpliceGrapher [28] and with three state-of-the-art transcript prediction methods: Cufflinks [32], IsoLasso [18] and Isolasso/CEM [19]. In simulation experiments, SpliceGrapherXT correctly predicts up to 56% more novel exons than our previous method, and more than twice as many exons as these other methods. Results on real data are comparable, with SpliceGrapherXT correctly recalling twice as many exons as other methods on human and plant data. SpliceGrapherXT is available for download at http://SpliceGrapher.sourceforge.net.

2. METHODS

2.1 Predicting splice graphs

SpliceGrapherXT accepts as input a set of RNA-Seq alignments in the Sequence Alignment/Map (SAM) format and a set of annotated gene models (either GFF3 or GTF format), and predicts splice graphs and transcripts based on these data. For each gene, the algorithm proceeds in the following steps:

1. Build an initial splice graph from the gene model.
2. Identify clusters of reads that overlap the gene.
3. Use patterns of donor and acceptor sites to predict novel exons.
4. Use spliced alignments to connect exons in the graph.
5. Update the graph with the new exons.
6. Save information about ambiguous regions that cannot be resolved.
7. (Optional) Resolve ambiguous regions in the splice graph by realigning the reads to a putative transcriptome.
8. (Optional) Employ PSGInfer or IsoLasso to predict isoforms from the resulting splice graph.

There are major differences between SpliceGrapherXT and the original SpliceGrapher in the algorithm used to predict novel exons (step 3 above), the addition of a realignment procedure designed to resolve ambiguous regions in the splice graph (step 7), and a method for using predicted splice graphs to predict novel transcripts (step 8). The original SpliceGrapher focused on predicting alternative splicing (AS) events and thus used separate inference rules for different kinds of AS. These isolated rules are sometimes unable to resolve all the evidence for a gene, as shown in Figure 1. For example, the original SpliceGrapher method is unable to resolve two novel skipped exons that fall within the same

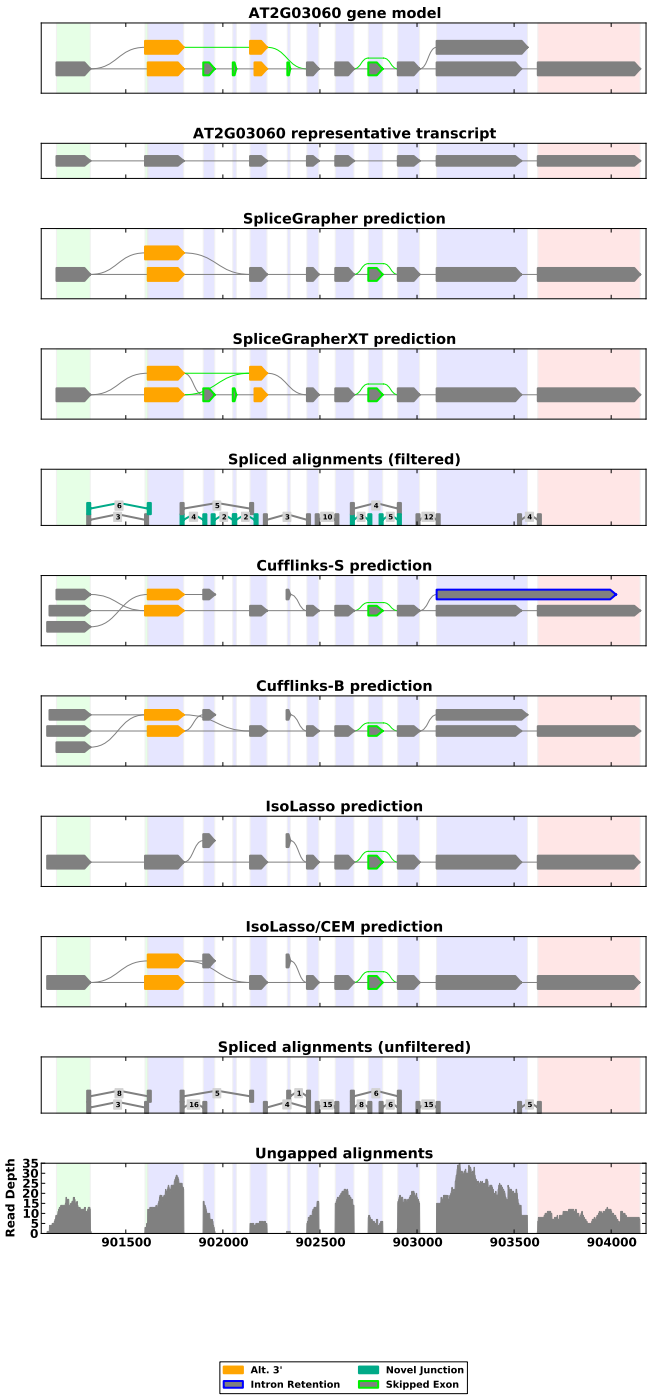

Figure 1: Example where the evidence for novel exons is ambiguous. We used the FluxSimulator tool to simulate reads from the complete gene model, then provided a representative transcript to each method to make its predictions. The original *SpliceGrapher* method was unable to resolve some exons because there is evidence for multiple kinds of AS within the same region. By combining constraints into a single comprehensive framework, *SpliceGrapherXT* is able to resolve these ambiguities. The other methods are unable to disambiguate this evidence.

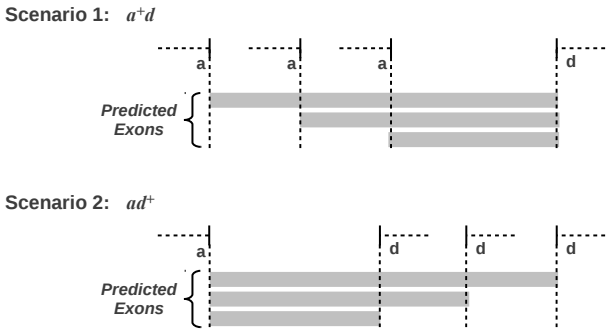

Figure 2: Correspondence between regular expression rules for sequences of acceptor and donor sites within read clusters, and the exon predictions that follow. Scenario 1 predicts splice sites at the 3' end of a putative exon, while scenario 2 predicts splice sites at the 5' end. We combine these expressions to predict the four primary AS events.

intron (Figure 1, third panel down). To predict a novel, skipped exon within an existing intron, the original inference rules require a unique acceptor site upstream of the exon and a unique donor site downstream from it. Here there are two novel exons within the same intron, accompanied by two novel acceptor sites and two donor sites. The original method cannot resolve these novel exons; but extending the method to handle all possible combinations of evidence would make the inference rules overly complex and inefficient.

SpliceGrapherXT predicts novel exons using a cohesive set of constraints that derive from the following observation: when two or more acceptor sites appear upstream of two or more donor sites in the same cluster, the choice of exon boundaries is not well defined. We specify these criteria as two constraints imposed on the splice sites within any cluster of reads before we will predict an exon:

1. At most one donor site can be downstream of multiple acceptor sites.
2. At most one acceptor site can be upstream of multiple donor sites.

Each of these constraints must hold for SpliceGrapherXT to make a prediction.

The RNA-Seq spliced alignments that fall within a genomic region form a sequence of confirmed donor and acceptor sites. To apply the criteria above to these sequences, we use the symbols a and d to denote acceptors and donors, respectively, and convert sequences of confirmed splice sites within a read cluster into strings of a's and d's. For every such string, we express the constraints above as follows:

1. At most one a precedes one or more d's, which may be expressed using a pattern of the form a^+d.
2. At most one d follows one or more a's, which may be expressed using the pattern ad^+.

These two constraints define a regular language to recognize strings that allows us to identify sequences of acceptor and donor sites that meet our requirements (see Figure 2). The original AS-based rules apply these constraints implicitly in the way they resolve putative exons [28]. Using a regular language allows us to make these constraints explicit and eliminates the need for having separate inference rules for each type of AS event.

2.2 Resolving ambiguities using realignment

Initially, SpliceGrapherXT predicts novel exons and records information about those it could not resolve due to ambiguous combinations of acceptor and donor sites. The next step in the algorithm is to resolve those ambiguous loci by realigning the reads to putative transcripts in order to find evidence for exons that result from ambiguous combinations of acceptor and donor sites. For each gene that contains an ambiguous locus, we construct putative exons for each combination of ambiguous acceptor and donor sites, inserting those putative exons into the splice graph. We then construct putative transcripts by traversing all paths through the graph. Although traversing all possible paths can potentially yield an intractably large putative transcriptome, in our experiments with human and Arabidopsis the realignment procedure takes no longer than the initial predictions. We use BWA [15] to realign reads to the putative transcripts and resolve putative exons whenever the read coverage across a transcript covers an entire exon plus anchor regions on either side (the default being 10 bases). Recall that unresolved exons arise from regions with multiple choices for acceptor and donor sites, so the anchor regions allow us to discriminate between them.

iReckon uses a similar realignment procedure [23]. The main difference in our approach is that it is only applied to those genes that contain unresolved exons. The idea of read realignment has also been used in the area of DNA sequence assembly for verification or refinement of an assembly [3, 12, 1, 29].

2.3 Predicting transcripts

SpliceGrapherXT has an optional step of predicting transcripts from its generated splice graphs using one of two existing methods—PSGInfer [14] and IsoLasso [18]. Both methods are designed to predict transcripts from known exons and quantify their expression. Without the benefit of SpliceGrapherXT, PSGInfer can only predict novel isoforms that include novel combinations of existing exons. IsoLasso is able to predict novel exons and isoforms, but in our experiments its recall falls well below that of SpliceGrapherXT.

SpliceGrapherXT's success in predicting novel exons is due largely to the fact that it predicts splice graphs instead of trying to predict complete mRNA transcripts. However, once we make accurate splice graph predictions, we may use the predicted graphs to predict transcripts by integrating our predictions with other methods. Here we present two procedures for predicting transcripts based on predicted splice graphs: one using PSGInfer and one that uses IsoLasso.

PSGInfer annotates a splice graph with edge weights that reflect the frequency with which each edge is used. Each tran-

script can then be assigned a probability that is the product of the edge weights along its path in the graph. SpliceGrapherXT then predicts those transcripts whose probability exceeds some threshold. We note that as a side-effect, using PSGInfer allows us to resolve exons that remain unresolved even after the realignment procedure.

We also explore the integration of SpliceGrapherXT with IsoLasso [18]. We use it to assign expression estimates to a putative transcriptome, and extract those transcripts for which the estimated expression exceeds a given threshold. We use IsoLasso for this procedure due to its popularity as a standard benchmark [18, 21, 23] and the ease of integrating it with our method.

2.4 Simulation experiments

To evaluate the performance of SpliceGrapherXT and compare it to Cufflinks we used reads simulated from the complete gene models. Each method was provided a single representative splice form for each gene; we then compared each method's predictions with the set of transcripts that were not given to it as input to assess its performance. Preliminary experiments have shown that providing the methods with multiple isoforms instead of a single isoform led to similar performance, so we chose to highlight the success of the methods in this somewhat more challenging task.

In these experiments we randomly selected 10 sets of 1000 genes that produce multiple isoforms and used the FluxSimulator [6] to generate reads from those genes. We aligned these reads to the genome using Tophat [31] and then used each pipeline to generate predictions.

The FluxSimulator provides a detailed simulation for each step in the RNA-Seq generation process [6]. Because it simulates some isoforms at higher expression levels than others, some isoforms from the gene models would be impossible to predict from just a few generated reads. At the exon level, our test set includes any exon that has some read coverage. At the isoform level, we include only isoforms for which every splice junction has a read that was generated from it.

2.5 Real data

Evaluation on real data is challenging because gene model annotations are incomplete and because we may have no *a priori* knowledge of which genes and transcripts are represented in an RNA-Seq data set. Similarly to our simulation experiments, we provide each method with a representative isoform from each gene, and compare their performance on the rest of the isoforms.

As with the simulations, we establish test sets of exons and isoforms to assess each method's performance. To identify the isoforms represented in the RNA-Seq data, we use Cufflinks to quantify gene and transcript expression against the full set of gene models. We then use the 95% confidence intervals provided by Cufflinks to identify all isoforms with positive expression levels.

We used data sets with read lengths of at least 75nt in two different species: *H. sapiens* and the model plant *A. thaliana*. For *H. sapiens* we downloaded 28M 75-nt read pairs (56M reads) that were used as a control for comparison with CLIP-seq reads[35] (NCBI accession SRP009861, sample SRX111920). For *A. thaliana* we used 76-nt data consisting of approximately 116M read pairs (232M reads) across five samples (19M to 65M reads per sample, NCBI accession SRA047499) [22]. We expected these reads to yield good results as they were derived from a cDNA library that was normalized to increase coverage across genes.

2.6 Evaluating performance

We measure performance at two levels, the exon and transcript level. SpliceGrapherXT predicts splice graphs that we can evaluate at the exon level. We can also evaluate SpliceGrapherXT with IsoLasso or PSGInfer at the transcript level. At both levels we provide the number of true positive predictions, precision and recall, where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$ and TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. In the case of real data however, precision is underestimated since a prediction that is not part of the known gene models, may be a novel exon or an isoform that is as-of-yet, unknown. This phenomenon is more pronounced in Arabidopsis, whose genome is not as well annotated as the human genome.

2.7 Parameter selection

In our experiments we used Cufflinks version 2.02 and IsoLasso version 2.6.1 that includes the LASSO and CEM methods. Both Cufflinks and IsoLasso have parameters that require tuning. Cufflinks uses a threshold on the predicted expression level to determine when to predict a transcript. The threshold is given as a fraction of the most abundant isoform for a gene, below which other isoforms will not be predicted. Set to 0, Cufflinks will predict any isoform with read coverage and should provide maximum recall; we also found that when we set it to 10% (0.1), Cufflinks achieves a good balance between recall and precision. We refer to these two settings as $Cufflinks_S$ (*sensitive*, threshold=0) and $Cufflinks_B$ (*balanced*, threshold=0.1). IsoLasso performance varies primarily with two parameters: a minimum expression value cutoff (`minexp`) similar to Cufflinks' threshold, and a coverage fraction cutoff (`u`) that controls its sensitivity to multiple isoforms. After extensive testing on simulated data we achieved the best performance using `minexp=0.0034` and `u=0.98` for IsoLasso and `minexp=0.002` and `u=0.78` for IsoLasso/CEM.

3. RESULTS

3.1 Simulation experiments

We ran our simulations on two different species: human and the model plant *A. thaliana* that has a small, compact genome with short introns. Although both species have well-annotated gene models, *H. sapiens* has much more complex gene models with many more isoforms per gene: on average, 6.85 annotated splice forms per gene compared with 1.24 per gene for *A. thaliana*.

Our simulation experiments show that the original SpliceGrapher and SpliceGrapherXT achieve very high precision at the exon level—nearly three times higher than Cufflinks and IsoLasso, while maintaining much higher recall at the same time (Table 1). SpliceGrapherXT recall is twice that

Exon prediction summary for *H. sapiens*				
Method	TP	FP	Recall	Precision
Reference	19,956	—	—	—
Test set	9,349	—	—	—
SpliceGrapher	573	**31**	0.06	**0.95**
SpliceGrapherXT	1,053	80	0.11	0.93
Realigned	**1,086**	82	**0.12**	0.93
Cufflinks$_S$	590	1,254	0.06	0.32
Cufflinks$_B$	574	1,278	0.06	0.31
IsoLasso	550	3,681	0.06	0.13
IsoLasso/CEM	546	4,005	0.06	0.12

Exon prediction summary for *A. thaliana*				
Method	TP	FP	Recall	Precision
Reference	10,168	—	—	—
Test set	2,623	—	—	—
SpliceGrapher	816	**35**	0.31	**0.96**
SpliceGrapherXT	1,207	64	0.46	0.95
Realigned	**1,299**	98	**0.49**	0.93
Cufflinks$_S$	620	1,318	0.24	0.32
Cufflinks$_B$	571	1,160	0.22	0.33
IsoLasso	520	1,844	0.20	0.22
IsoLasso/CEM	607	1,923	0.23	0.24

Table 1: Performance averaged over 10 simulation runs for *H. sapiens* (top) and *A. thaliana* (bottom) with 1,000 randomly-selected genes with multiple transcripts. Approximately 1 million paired-end reads were generated for each run (500,000 pairs). Each method was provided with a single representative isoform. The *Reference* row provides the number of exons present in the full gene models for the randomly selected genes; *test set* provides the number of exons that were removed. Results show the number of correctly predicted exons (TP), and the recall and precision for each method.

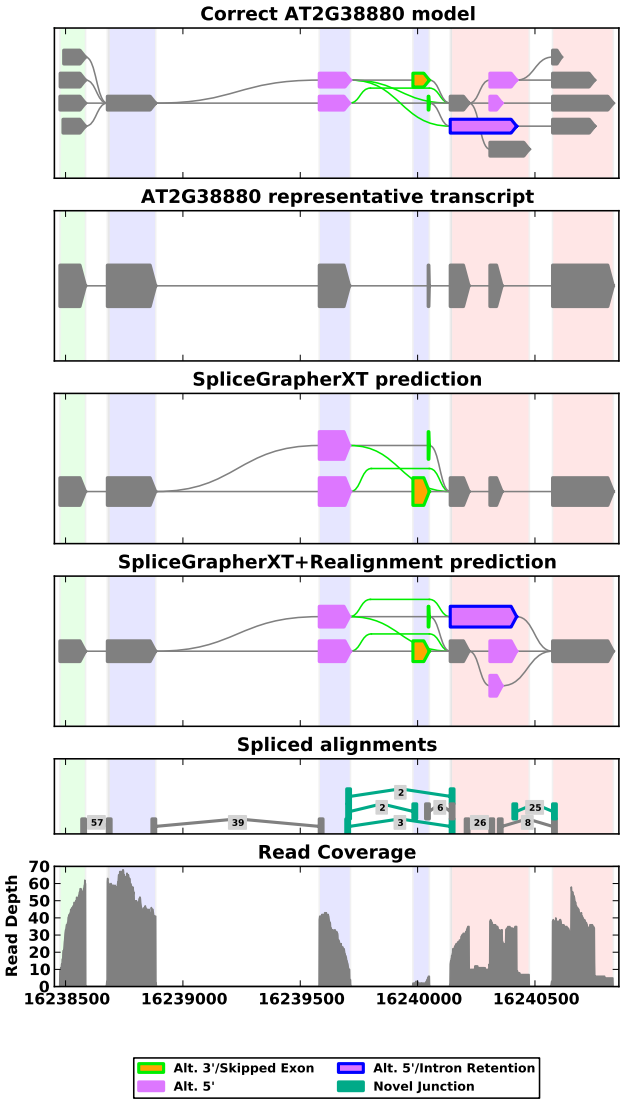

Figure 3: An example of the success of the realignment procedure. In this case, there is evidence for several novel exons near the 3' end of the gene. The splice site pattern is not recognized by SpliceGrapherXT's regular expression and so does not yield an initial prediction. SpliceGrapherXT stores information about this evidence for the realignment procedure, which finds compelling evidence for three of the novel exons that match the original gene model.

of Cufflinks and IsoLasso and also presents a major improvement over the original SpliceGrapher. With the realignment procedure, SpliceGrapherXT's recall increases by up to 8% with little appreciable decrease in precision.

Visual inspection of some genes confirms that SpliceGrapherXT can make predictions where other methods cannot (Figure 1). In this example, evidence for multiple forms of AS in one location in the gene makes it difficult for any of the methods to resolve all of the evidence. SpliceGrapherXT is able to resolve the most evidence, including a new exon that extends the gene on its 5' end. Cufflinks is unable to make sense of this evidence and reverts to the reduced gene model. The IsoLasso and IsoLasso/CEM methods both yield the same prediction in this case, resolving some of the evidence at the 5' end of the gene, although they fail to resolve its 3' end.

We also confirmed that when SpliceGrapherXT cannot predict a definite set of exons, the realignment procedure can resolve them in many cases (Figure 3). Here SpliceGrapherXT is able to resolve some of the evidence for novel exons, but one of the clusters contains the pattern *aaadadd* that does not match the splice site regular expression. Information about this evidence is retained and used in the realignment procedure, which finds compelling evidence for three of the novel exons that match the original gene model.

We expected that paired-end reads would resolve more exons than single-end reads. However, we found that this was not the case, and the reason may be traced to the constraints we placed on paired-end reads: to accept an alignment we require both reads in a pair to align to the same transcript on opposite strands, and the alignment positions must correspond to an insert size that is compatible with the distribution for the data set. Altogether, these constraints reduced the number of paired-end alignments we could use, which may explain why we did not see increased performance from paired-end data.

3.2 Real data

We further tested the ability of SpliceGrapherXT to identify novel exons and isoforms using real paired-end RNA-Seq data. Our results are summarized in Table 2. In *H. sapiens* we find that SpliceGrapherXT predicts nearly twice as many of the left-out exons as Cufflinks at either setting. The realignment procedure further increases the exon recall by 10%. In *A. thaliana* SpliceGrapherXT's advantage is even more noticeable: with the realignment procedure it predicts twice as many left-out exons as Cufflinks.

In our simulations we found that precision was higher in *A. thaliana* than in *H. sapiens* for all methods. This trend is reversed for real data. The reason for this is because the human genome is more thoroughly annotated than that of *A. thaliana*; results on real data in human show few novel exons, whereas in *A. thaliana*, many novel exons are discovered. These novel exons are considered to be false positives, where in practice they may be exons that have not been annotated yet.

3.3 Predicting transcripts

Statistics for *H. sapiens*				
Method	TP	FP	Recall	Precision
Left out	87,799	—	—	—
SpliceGrapher	2,173	**1,071**	0.025	**0.67**
SpliceGrapherXT	3,965	2,756	0.045	0.59
Realigned	**4,367**	3,163	**0.050**	0.58
Cufflinks$_S$	2,006	10,532	0.023	0.16
Cufflinks$_B$	2,015	18,135	0.021	0.10
IsoLasso	1,308	31,392	0.013	0.04
IsoLasso/CEM	1,359	32,616	0.014	0.04

Average statistics for five *A. thaliana* replicates				
Method	TP	FP	Recall	Precision
Left out	14,038	—	—	—
SpliceGrapher	2,041	**6,123**	0.15	**0.25**
SpliceGrapherXT	2,315	12,154	0.17	0.16
Realigned	**2,833**	14,874	**0.20**	0.16
Cufflinks$_S$	1,636	80,164	0.11	0.02
Cufflinks$_B$	996	23,904	0.07	0.04
IsoLasso	890	28,777	0.06	0.03
IsoLasso/CEM	990	32,011	0.07	0.03

Table 2: This table shows the prediction performance on real data for *H. sapiens* (top) and *A. thaliana* (bottom). Shown are the number of exons removed from the original gene models (left out) followed by the statistics for each method.

To test our two approaches to predicting transcripts, we integrated our SpliceGrapherXT predictions with IsoLasso [18] and PSGInfer. To integrate SpliceGrapher with these tools, we create putative gene models from SpliceGrapher's predicted graphs and provide the gene models to each method in the form of a GTF file (PSGInfer) or a BED file (IsoLasso). To augment the original gene models, we create putative transcripts for every path through a graph. We then evaluate each method by setting a minimum threshold (a probability for PSGInfer, an FPKM value for IsoLasso) and comparing the exons and isoforms above the threshold with those in our test set.

In our simulations, all of the SpliceGrapherXT update procedures except IsoLasso provide much higher recall than Cufflinks and all yield higher precision in *A. thaliana* (Table 3). In *H. sapiens* the results are mixed, as both PSGInfer$_{.01}$ and the unassisted IsoLasso/CEM achieve the highest recall, while Cufflinks$_B$ achieves the highest precision. The PSGInfer$_{.01}$ update procedure provides the highest recall of any method in both species, while PSGInfer$_{.15}$ consistently yields the highest precision of the update procedures. The differences between making predictions in *A. thaliana* and in *H. sapiens* are clear from these simulations: all of the SpliceGrapherXT update methods have recall that is much higher in the less well-annotated *A. thaliana* than in human, with a similar increase in precision. With fewer transcripts per gene in the *A. thaliana* gene models, fewer transcripts are represented in the simulated RNA-Seq data, making the prediction task much easier than it is for human.

We tested our isoform prediction procedure on the real data from *A. thaliana* and *H. sapiens* using PSGInfer and Iso-

Transcript prediction summary for *H. sapiens*				
Method	TP	FP	Rec.	Prec.
SGXT updates				
PSGInfer$_{.01}$	**282**	2,538	**0.16**	0.10
PSGInfer$_{.15}$	180	**1,205**	0.10	0.13
IsoLasso	187	1,891	0.10	0.09
CEM	264	2,376	0.15	0.10
Unassisted methods				
IsoLasso	233	4,427	0.13	0.05
CEM	**281**	5,339	**0.16**	0.05
Cufflinks$_S$	250	2,528	0.14	0.09
Cufflinks$_B$	239	1,355	0.13	**0.15**

Transcript prediction summary for *A. thaliana*				
Method	TP	FP	Rec.	Prec.
SGXT updates				
PSGInfer$_{.01}$	**942**	508	**0.74**	0.65
PSGInfer$_{.15}$	796	**130**	0.63	**0.86**
IsoLasso	665	272	0.52	0.71
CEM	920	518	0.72	0.64
Unassisted methods				
IsoLasso	501	752	0.39	0.40
CEM	629	834	0.49	0.43
Cufflinks$_S$	546	1,554	0.43	0.26
Cufflinks$_B$	617	485	0.49	0.56

Table 3: Transcript prediction performance averaged over 10 simulated runs. On average there were 1,800 isoforms removed from a random selection of 1,000 genes in human, and 1,273 isoforms on average in *A. thaliana*. Shown are the number of true-positive and false-positive predicted isoforms along with the recall and precision. Results are shown for the SpliceGrapherXT update methods (SGXT updates): SpliceGrapherXT with PSGInfer, SpliceGrapherXT with IsoLasso and SpliceGrapherXT with IsoLasso/CEM. Also shown are results for unassisted IsoLasso, IsoLasso/CEM, the high-sensitivity version of Cufflinks (Cufflinks$_S$), and the balanced performance version (Cufflinks$_B$).

Transcript prediction summary for *H. sapiens*				
Method	TP	FP	Rec.	Prec.
SGXT updates				
PSGInfer$_{.01}$	1,675	26,715	0.060	0.059
PSGInfer$_{.15}$	1,270	20,627	0.046	0.058
IsoLasso	1,355	**13,536**	0.049	0.091
CEM	1,569	13,966	0.057	**0.101**
Unassisted methods				
IsoLasso	709	117,458	0.026	0.006
CEM	832	207,168	0.030	0.004
Cufflinks$_S$	**2,825**	120,002	**0.102**	0.023
Cufflinks$_B$	1,594	70,861	0.057	0.022

Transcript prediction summary for *A. thaliana*				
Method	TP	FP	Rec.	Prec.
SGXT updates				
PSGInfer$_{.01}$	**1,388**	28,144	**0.269**	0.047
PSGInfer$_{.15}$	991	**5,127**	0.192	**0.162**
IsoLasso	999	14,611	0.194	0.064
CEM	1,283	17,867	0.249	0.067
Unassisted methods				
IsoLasso	627	24,453	0.122	0.025
CEM	749	26,001	0.145	0.028
Cufflinks$_S$	1,143	59,015	0.221	0.019
Cufflinks$_B$	1,004	9,347	0.195	0.097

Table 4: Transcript prediction performance on real data from *A. thaliana* and *H. sapiens*. Shown are the number of left-out isoforms correctly predicted (TP), the number incorrectly predicted (FP), the proportion correctly predicted (recall) and the proportion of true-positives (precision). In *A. thaliana*, the PSGInfer$_{.01}$ and CEM update methods have the highest recall, but only PSGInfer$_{.15}$ has higher precision than Cufflinks$_B$. On human data the update procedures yield lower recall than Cufflinks$_S$, but with up to four times the precision.

Lasso to predict the isoforms recapitulated in the data (Table 4). The results for real data are more subtle than for simulated data. In *A. thaliana*, the PSGInfer and CEM update methods have better recall than the Cufflinks methods, but with lower precision than Cufflinks$_B$. PSGInfer at the high threshold exceeds the precision of Cufflinks$_B$ substantially. On human data the differences between the update procedures and Cufflinks are more noticeable, where the update procedures can yield precision more than four times as high as Cufflinks.

Prediction is much harder with real data than with simulated data. Real data can include a considerable amount of noise generated during the sequencing process, as well as evidence for novel transcripts. As a result, recall and precision on real data are much lower for all methods than with the simulated data. This is especially evident in *A. thaliana*: the methods predict a large number of novel transcripts, so precision and recall are low relative to the gene models.

We are encouraged by the results for the transcript prediction pipeline. Methods such as PSGInfer and IsoLasso appear to work well when we provide them with accurate information about possible transcripts. Recently iReckon has been shown to provide higher precision and recall than IsoLasso [23] for transcript prediction, so we plan to compare SpliceGrapherXT's predictions with iReckon. Our results suggest that an approach that combines SpliceGrapherXT's conservative splice graph predictions with transcript expression estimation methods may give us the most complete picture of a gene's splicing behavior based on existing data.

4. CONCLUSION

We have presented SpliceGrapherXT, a novel method that predicts splice graphs from RNA-Seq data by applying a simple set of constraints to patterns of acceptor and donor sites. Our results show that compared with Cufflinks, IsoLasso and IsoLasso/CEM, SpliceGrapherXT provides much greater precision when predicting novel exons, while providing higher recall as well. In addition, we have presented a realignment method for resolving ambiguities in the RNA-Seq data that inference alone cannot. This procedure can resolve many of the ambiguities in the initial predictions and thus increase the tool's sensitivity to novel exons. Finally, we have explored two methods for converting SpliceGrapherXT's splice graph predictions into transcript predictions. Our results show that these integrated methods can produce transcript predictions with much higher recall and precision than Cufflinks.

5. ACKNOWLEDGMENT

The authors wish to acknowledge Colin Dewey at the University of Wisconsin-Madison for his assistance with PSGInfer.

Funding. This project was supported by NSF ABI grant 0743097 and DOE-USDA grant 207793.

6. REFERENCES

[1] J. A. Chapman, I. Ho, S. Sunkara, S. Luo, G. P. Schroth, and D. S. Rokhsar. Meraculous: De Novo Genome Assembly with Short Paired-End Reads. *PLoS ONE*, 6(8):e23501, 08 2011.

[2] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *BMC Bioinformatics*, 9(Suppl 10):O7, 2008.

[3] N. Donmez and M. Brudno. Hapsembler: an assembler for highly polymorphic genomes. In *Proceedings of the 15th Annual international conference on Research in computational molecular biology*, RECOMB'11, pages 38–52, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] J. Feng, W. Li, and T. Jiang. Inference of Isoforms from Short Sequence Reads. *Journal of Computational Biology*, 18(3):305–321, 2011.

[5] S. Filichkin, H. Priest, S. Givan, R. Shen, D. Bryant, S. Fox, W. Wong, and T. Mockler. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, 20(1):45, 2010.

[6] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigó, and M. Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 2012.

[7] M. Guttman, M. Garber, J. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, 2010.

[8] B. Harr and L. Turner. Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Molecular Ecology*, 19:228–239, 2010.

[9] E. Harrington and P. Bork. Sircah: a tool for the detection and visualization of alternative transcripts. *Bioinformatics*, 24(17):1959, 2008.

[10] G. Jean, A. Kahles, V. Sreedharan, F. Bona, and G. Rätsch. RNA-Seq Read Alignments with PALMapper. *Current Protocols in Bioinformatics*, 32:11.6.1–11.6.37, 2010.

[11] A. Kalsotra and T. A. Cooper. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics*, 12(10):715–729, 2011.

[12] J. D. Klein, S. Ossowski, K. Schneeberger, D. Weigel, and D. H. Huson. LOCAS – A Low Coverage Assembly Tool for Resequencing Projects. *PLoS ONE*, 6(8):e23455, 08 2011.

[13] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[14] L. LeGault and C. Dewey. Learning Probabilistic Splice Graphs from RNA-Seq data. *under review*, 2013.

[15] H. Li and R. Durbin. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, 2009.

[16] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851, 2008.

[17] J. J. Li, C.-R. Jiang, J. B. Brown, H. Huang, and P. J. Bickel. Sparse linear modeling of next-generation

mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, 108(50):19867–19872, 2011.

[18] W. Li, J. Feng, and T. Jiang. IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly. In V. Bafna and S. Sahinalp, editors, *Research in Computational Molecular Biology*, volume 6577 of *Lecture Notes in Computer Science*, pages 168–188. Springer Berlin / Heidelberg, 2011.

[19] W. Li and T. Jiang. Transcriptome assembly and isoform expression level estimation from biased rna-seq reads. *Bioinformatics*, 28(22):2914–2921, 2012.

[20] D. D. Licatalosi and R. B. Darnell. RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, 11(1):75–87, 2010.

[21] S. Mangul, A. Caciula, S. Al Seesi, D. Brinza, A. R. Banday, and R. Kanadia. An integer programming approach to novel transcript reconstruction from paired-end RNA-Seq reads. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 369–376. ACM, 2012.

[22] Y. Marquez, J. W. Brown, C. Simpson, A. Barta, and M. Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, 22(6):1184–1195, 2012.

[23] A. M. Mezlini, E. J. Smith, M. Fiume, O. Buske, G. L. Savich, S. Shah, S. Aparicio, D. Y. Chiang, A. Goldenberg, and M. Brudno. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome research*, 23(3):519–529, 2013.

[24] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5:621–628, 2008.

[25] M. Nicolae, S. Mangul, I. Măndoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms in Bioinformatics*, pages 202–214, 2010.

[26] A. Ramani, J. Calarco, Q. Pan, S. Mavandadi, Y. Wang, A. Nelson, L. Lee, Q. Morris, B. Blencowe, M. Zhen, et al. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome Research*, 21(2):342, 2011.

[27] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 2011.

[28] M. Rogers, J. Thomas, A. Reddy, and A. Ben-Hur. SpliceGrapher: Detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. *Genome Biology*, 13(R4), 2012.

[29] R. Ronen, C. Boucher, H. Chitsaz, and P. Pevzner. Sequel: improving the accuracy of genome assemblies. *Bioinformatics*, 28(12):i188–i196, 2012.

[30] P. A. Sharp. The centrality of RNA. *Cell*, 136(4):577–580, 2009.

[31] C. Trapnell, L. Pachter, and S. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 2009.

[32] C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[33] K. Wang, D. Singh, Z. Zeng, S. Coleman, Y. Huang, G. Savich, X. He, P. Mieczkowski, S. Grimm, C. Perou, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 2010.

[34] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[35] R. Xiao, P. Tang, B. Yang, J. Huang, Y. Zhou, C. Shao, H. Li, H. Sun, Y. Zhang, and X.-D. Fu. Nuclear Matrix Factor hnRNP U/SAF-A Exerts a Global Control of Alternative Splicing by Regulating U2 snRNP Maturation. *Molecular Cell*, 45(5):656–668, March 2012.

[36] M. Yassour, T. Kaplan, H. Fraser, J. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences*, 106(9):3264, 2009.