Journal of Visualized Experiments

OLego: de novo discovery of novel exon junctions and small exons from RNA-Seq data

--Manuscript Draft--

Manuscript Number:	JoVE52631R3
Full Title:	OLego: de novo discovery of novel exon junctions and small exons from RNA-Seq data
Article Type:	Invited Methods Article - JoVE Produced Video
Keywords:	Bioinformatics; Genomics; next generation sequencing (NGS); Alignment; Mapping; RNA-Seq; exon junction detection; splicing; small exons; micro-exons
Manuscript Classifications:	5.5.393.751: Sequence Alignment; 5.5.393.760: Sequence Analysis; 5.5.393.760.319: High-Throughput Nucleotide Sequencing; 7.5.355.315.700.700: RNA Splicing; 7.5.355.315.700.700.100: Alternative Splicing; 8.1.158.273.180: Computational Biology; 8.1.158.273.180.350: Genomics
Corresponding Author:	Chaolin Zhang Columbia University New York, New York UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author E-Mail:	cz2294@columbia.edu
Corresponding Author's Institution:	Columbia University
Corresponding Author's Secondary Institution:	
First Author:	Jie Wu
First Author Secondary Information:	
Other Authors:	Jie Wu
	Huijuan Feng
Order of Authors Secondary Information:	
Abstract:	OLego is a software tool specifically designed for de novo mapping of spliced RNA-Seq reads. It uses a seed-and-extend scheme to identify exon junctions and small exons from the reads. One major difference between OLego and other splice mappers is its ability to achieve high sensitivity by using strategic and efficient searches with very small seeds (12~15 nt for mammalian genomes). Meanwhile, to improve accuracy and to resolve ambiguous mapping at exon junctions, OLego uses a built-in regression model to score and rank exon junctions by considering both splice-site strength and intron size. To ensure efficient alignments of massive RNA-Seq datasets with this high-resolution search, we do not rely on external mappers but use Burrows-Wheeler transform (BWT) and full-text minute-space (FM)-index in multiple steps to efficiently map seeds, locate exon junctions and identify small exons. OLego is implemented in C++ supporting fully multi-threaded execution, and allows fast processing of large-scale data. In this article, we describe the workflow of OLego and present a standard protocol to run OLego for RNA-Seq alignment. Its performance was compared with other mainstream tools using both simulated and real experimental data to demonstrate its sensitivity, accuracy and efficiency.
Author Comments:	
Additional Information:	
Question	**Response**
If this article needs to be "in-press" by a certain date to satisfy grant requirements, please indicate the date below and explain in your cover letter.	

If this article needs to be filmed by a certain date to due to author/equipment/lab availability, please indicate the date below and explain in your cover letter.	

TITLE:
OLego: *de novo* discovery of novel exon junctions and small exons from RNA-Seq data

AUTHORS:
Wu, Jie
Cold Spring Harbor Laboratory
Cold Spring harbor, NY
wuj@cshl.edu

Feng, Huijuan
Department of Systems Biology
Department of Biochemistry and Molecular Biophysics
Center for Motor Neuron Biology and Disease
Columbia University
New York, NY
hf2304@columbia.edu

Zhang, Chaolin
Department of Systems Biology
Department of Biochemistry and Molecular Biophysics
Center for Motor Neuron Biology and Disease
Columbia University
New York, NY
cz2294@columbia.edu

CORRESPONDING AUTHOR:
Zhang, Chaolin
cz2294@columbia.edu

KEYWORDS:
Bioinformatics, genomics, next generation sequencing (NGS), alignment, mapping, RNA-Seq, exon junction detection, splicing, small exons, micro-exons

SHORT ABSTRACT:
Here we present a protocol to use OLego to perform fast, accurate and sensitive alignment of spliced RNA-Seq reads. OLego is optimized for *de novo* detection of exon junctions and small exons by using small seeds with very efficient algorithm implementations.

LONG ABSTRACT:
OLego is a software tool specifically designed for *de novo* mapping of spliced RNA-Seq reads. It uses a seed-and-extend scheme to identify exon junctions and small exons from the reads. One major difference between OLego and other splice mappers is its ability to achieve high sensitivity by using strategic and efficient searches with very small seeds (12~15 nt for mammalian genomes). Meanwhile, to improve accuracy and to resolve ambiguous mapping at exon junctions, OLego uses a built-in regression model to score and rank exon junctions by considering both splice-site strength and intron size. To ensure efficient alignments of massive

RNA-Seq datasets with this high-resolution search, we do not rely on external mappers but use Burrows–Wheeler transform (BWT) and full-text minute-space (FM)-index in multiple steps to efficiently map seeds, locate exon junctions and identify small exons. OLego is implemented in C++ supporting fully multi-threaded execution, and allows fast processing of large-scale data. In this article, we describe the workflow of OLego and present a standard protocol to run OLego for RNA-Seq alignment. Its performance was compared with other mainstream tools using both simulated and real experimental data to demonstrate its sensitivity, accuracy and efficiency.

INTRODUCTION:

Deep RNA sequencing (RNA-Seq) is a recently developed approach that uses the next-generation sequencing (NGS) technologies to profile the transcriptome at unprecedented depth and resolution, which has greatly facilitated the discovery and quantification of transcripts or transcript variants resulted from alternative splicing[1]. The first step to analyze RNA-Seq data is alignment of many millions of short reads to the reference genome to locate exon body reads and exon junction reads. The challenges of this step are two folds: First, short reads increase the possibility of having multiple hits on the genome by chance, and also make it difficult to identify exon junctions due to small overlaps in the exons across the junction. Second, the large amount of data requires optimizations in the algorithm to ensure high efficiency of alignment without sacrificing accuracy and sensitivity.

De novo discovery of exon junctions in RNA-Seq data has become an important resource to annotate gene structures and study gene expression regulation including alternative splicing. To overcome the aforementioned challenges, a number of algorithms have been developed in the recent years to specifically align spliced reads to exon junctions[2-8]. To locate exon junction reads, a common strategy used by most of the programs is the seed-and-extend method, although details may vary. In this method, each read is first split into two or more segments (seeds), which are then aligned to the genome independently without allowing splits. A search for an exon junction is then performed between two adjacent seeds that are mapped to two genomic loci separated by certain distance consistent with a candidate intron (double-anchor search). For a seed near the end of the read, the alignment of the single seed can also be used as an anchor to guide a search for a potential exon junction using the unaligned portion at the end of the read at a higher resolution (single-anchor search) (Figure 1).

Many of these software tools require an external mapper (e.g., Bowtie[9]) for seed mapping, which prevented the use of short seeds in practice, because of formidable requirements on storage space for temporary files and computation time. This caveat largely limits the mapping resolution of these tools, and hence their sensitivity to detect exon junctions and small exons. To overcome this limitation, we recently designed a novel algorithm and program named OLego for fast and sensitive mapping of RNA-Seq reads using very small seeds[10].

The overview of OLego is briefly described in the following workflow (Figure 1):

OLego first attempts to map unspliced reads to the genome. Reads which cannot be mapped within a certain number of mismatches or small insertions/deletions (indels) are processed in the following junction searching steps. Each read is segmented into multiple seeds. Small seed is used to improve sensitivity (default: 15 nt with 1-nt overlap, which is optimized for mammalian-

sized genomes). In the next step, each seed of a read is mapped independently to the reference genome by querying the indexed genome. Their hits are then collected, sorted, and clustered into one or more potential alignments of the read. Afterwards, the hits in each potential alignment are further grouped into candidate exons. A candidate exon is defined when one or more seeds are mapped to a continuous genomic region while two consecutive candidate exons are separated by a large gap representing potential introns. Alignment of each candidate exon is then extended to identify approximate exon boundary positions. Then double-anchor and single-anchor junction searches are performed between each pair of consecutive candidate exons or at the ends of each alignment if unaligned sequence segments are present at the end of the read. In addition, potential small exons are searched when there are unaligned gaps between two candidate exons. Finally, exon junctions and exons are connected to obtain the complete alignment of the read. Multiple candidate alignments, if any, are ranked and filtered by considering the number of mismatches, splice site strength and intron sizes.

Since each read is aligned independently, multiple threads can be easily enabled for parallelization. For more details of the algorithm, please refer to the original publication[10]. Additional online documentation and software updates are available at http://zhanglab.c2b2.columbia.edu/index.php/OLego.

[Figure 1]

PROTOCOL:

1. Installation and preparation to run OLego

1.1) Installation

Note: The major components of OLego (olego and olegoindex) were written in C++ and can be installed and run on mainstream Unix-based systems (Linux or Mac OS X) with the GCC compiler installed. Installation of Perl and R is required by additional scripts in the software package for offline construction of the regression models that score the strength of exon junctions (models for human and mouse have been provided in the package), and post processing of read alignment, such as merging paired-end reads and identification of unique exon junctions.

1.1.1) Download the most recent stable version of the OLego source code or executable binaries from SourceForge (http://sourceforge.net/projects/ngs-olego/files/. Alternatively, use git to retrieve the most updated version of the source code from the repository:

git clone git://git.code.sf.net/p/ngs-olego/code olego

Note: Skip to 1.2 if the binary executable files are downloaded in this step.

1.1.2) Decompress the package and change the current directory to OLego folder:

tar zxvf olego.src.v1.x.x.tgz
cd olego/

1.1.3) Compile the source code. Two executable files (olego and olegoindex) will be created in the folder.

make

Note: For 32-bit systems, edit the "Makefile" to remove the "-m64" tag before compiling the code.

1.2) Prepare for the indexed reference genome

1.2.1) Download the FASTA genome sequence from the UCSC genome browser.(e.g., http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz). Here the mouse genome is used as an example in this protocol.

1.2.2) Decompress the gz file and concatenate the files into a single FASTA file (e.g. mm10.fa). It is recommended that random chromosomes and haplotype sequences are typically excluded from the analysis.

```
tar zxvf chromFa.tar.gz
rm chr*random.fa chrUn*fa
cat *.fa >mm10.fa
```

1.2.3) Run olegoindex to build the index for the reference genome:

```
olegoindex mm10.fa –p mm10
```

Note: There will be eight files (mm10.pac, mm10.ann, mm10.amb, mm10.rpac, mm10.bwt, mm10.rbwt, mm10.sa and mm10.rsa) generated.

1.3) Other required files

1.3.1) Download the pre-built exon junction database from the OLego website "http://zhanglab.c2b2.columbia.edu/index.php/OLego". Decompress this file.

```
gunzip mm10.intron.hmr.bed.gz
```

1.3.1.1) Alternatively, create a custom junction database from transcript annotations (BED format file) using the Perl script included in the package (bed2junc.pl).

```
bed2junc.pl transcripts.bed transcripts.junc.bed
```

Note: It is strongly recommended that a database of annotated exon junctions should be provided to OLego (option -j) for more sensitive mapping.

1.3.2) Find the logistic regression models for mouse and human exon junctions included in the

sub folder “models”. Use the script included in the package to build models for other species. Type “perl regression_model_gen/OLego_regression.pl” in the olego folder to see the help message.

2. Align the reads using OLego

2.1) Align paired-end reads separately:

olego -v -r models/mm.cfg -j mm10.intron.hmr.bed -o r1.sam mm10 r1.fq
olego -v -r models/mm.cfg -j mm10.intron.hmr.bed -o r2.sam mm10 r2.fq

2.1.1) Use option “-t” to specify the number of CPU cores to be used in the alignment.

2.1.2) Use “-M” to specify the maximum number of mismatches or indels allowed. The default will be determined based on the read length (Supplementary Table 1).

2.1.3) Use “-w” to control the seed size.

Note: The seeds will be evenly distributed on the read from the beginning to the end with a maximum overlap defined by --word-max-overlap. The default is 15 nt seed with 1-nt overlap, which is recommended for relatively long reads (e.g., >70 nt). A smaller seed (e.g., 12-14 nt) should be used for shorter reads in combination with different seed overlap size (e.g., 0-2 nt) to maximize the number of seeds that covers the read.

2.1.4) For strand-specific RNA-Seq libraries, check option “–-strand-mode”.

2.1.5) Use “-r” to specify the logistic regression model prepared in step 1.3.2.

2.1.6) For other options, refer to the online documentation.

2.2) Merge the mapping results from paired-end reads according to their distance and orientations to resolve ambiguities:

mergePEsam.pl r1.sam r2.sam merge.sam

2.2.1) Use “-d” to specify the maximum distance between the two ends on the reference. This script requires the two ends mapped to different strands by default, use “--ss” (requiring mapping to the same strand) or “--ns” (ignoring strand information) to change the behavior.

2.3) Optionally, re-align reads with the exon junctions identified from the data. Some reads supporting novel exon junctions that are failed during the first-pass alignment can be recovered because of more extensive searches for reads mapped to known exon junctions.

2.3.1) Convert the SAM alignment output file to BED format:

sam2bed.pl merge.sam merge.bed

2.3.1.1) Use "--uniq" to keep only reads aligned to unique loci (i.e., single hits). Use "--use-RNA-strand" to extract the RNA strand instead of the read strand for junction reads when the orientation can be inferred from the splice sites ("+" is assigned to exonic reads) .

2.3.2) Extract the exon junctions from the BED file:

bed2junc.pl merge.bed merge.junc.bed

2.3.3) Remap with this new exon junction annotation file:

olego -v -r mm.cfg -j merge.junc.bed --non-denovo -o r1.remap.sam mm10.fa r1.fq
olego -v -r mm.cfg -j merge.junc.bed --non-denovo -o r2.remap.sam mm10.fa r2.fq

2.3.4) See Section 2.2 to merge the paired-end reads.

3 Post-processing and downstream analysis

3.1) Sort and convert the merge.sam into BAM format files. SAMtools[11] is required for this step:

samtools view -uSh merge.sam | samtools sort - merge.sort

3.2) Perform downstream analysis using other tools. Examples include quantification of gene expression (e.g., Cufflinks/Cuffdiff[12] and HTSeq-count[13]/DESeq[14]/EdgeR[15]), alternative splicing (e.g., Quantas: http://zhanglab.c2b2.columbia.edu/index.php/Quantas) and *de novo* transcriptome assembly (e.g., Trinity[16] and Velvet[17]).

REPRESENTATIVE RESULTS:

We previously assessed OLego (v1.0.0) using simulated data (10 million paired-end reads of 100 or 150 nt) and compared with other spliced alignment tools[10] (TopHat v1.4.0 , ref.[3], MapSplice v1.15.2, ref.[4] and PASSion v1.2.1, ref.[7]). For exon junction discovery, OLego had the lowest false negative rate (FNR) (6.8%~8.2%), which almost halved the FNRs from TopHat (12.8%~15.4%) and PASSion (14.8%~15.5%). Meanwhile, OLego achieved a high accuracy with high positive predictive values (PPVs) (97.7%~98.1%), which is comparable to TopHat v1.4.0 and better than MapSplice and PASSion. In addition, we demonstrated that OLego is very sensitive to discover small or micro-exons. For exons with a length between 9 and 39 nt, OLego achieved a much lower FNR (9.2%~13.4%), compared to TopHat v1.4.0 (31.5%~36.1%), MapSplice (19.4%~24.1%), and PASSion (21.3%~32.3%). OLego is particularly sensitive in detecting micro-exons smaller than 15 nt. While OLego discovered >75% of the exons in this range in the simulation test, the other programs discovered substantially lower percentages (6.1% ~ 47.4%).

The high sensitivity of exon-junction and small-exon discovery is attributed to the high-resolution searches using small seeds and optimization in various steps. Nevertheless, OLego is also among the fastest programs in terms of the mapping speed, especially when multiple

threading is enabled. The high efficiency of alignment originates from the internal querying algorithm (BWT and FM-index [18]) used in multiple mapping steps, the multi-threading implementation and several heuristics fine-tuned to limit the query complexity. In comparison with TopHat v1.4.0 , which was the fastest among the other three algorithms we tested, OLego achieved a comparable speed when using single CPU core, and 2-fold faster with 16 CPU cores (please see our initial publication[10] for more details).

The use of small seeds is very computationally expensive. To find a fine balance between mapping speed and sensitivity, we have made additional optimization and improvements since the initial release of OLego. These improvements include the use of overlapping seeds starting from v1.1.2, such that a similar sensitivity can be achieved with slightly longer seeds (e.g., 15-nt seeds with 1 nt overlap vs. 14-nt seeds without overlap), while the mapping speed is greatly improved.

Since our initial publication, several other programs (e.g., STAR[19]) were published and new versions have been released for some of the programs we previously compared. Therefore, we carried out additional comparisons of OLego v1.1.5 with TopHat v2.0.11[20] and STAR v2.3.0e using the same simulated datasets[10]. Default parameters are used for both programs, except that intron sizes were limited to the range of 20~500,000 nt and the "--microexon-search" option was enabled for TopHat2. The primary alignments from each program were then extracted to evaluate the performance in *de novo* discovery of exon junctions and small exons.

Compared to the results from OLego v1.0.0, the new version (v1.1.5) achieved higher sensitivity with lower FNRs on exon-junction discovery (5.7%~7.2%). The sensitivity for small exon discovery also improved. For example, the FNRs for exons of 9~39 nt decreased from 9.2%~13.4% to 8%~11.1%. At the same time, the PPVs for both exon junction and small exon discovery are maintained at the same level (94.2% ~ 98.1%). In addition, the speed has been improved significantly because of optimizations implemented in the newer version. The computing time decreased from 0.8h to 0.51h to map 10 million 2×100-nt reads, and from 1.4h to 1.07h to map 10 million 2×150-nt reads (Table 1). The accuracy and sensitivity of TopHat2 are comparable to TopHat v1.4.0 . STAR has the highest sensitivity for overall junction discovery at the expense of slightly lower PPV. In addition, its FNRs on discovery of small exons are higher than OLego, especially for exons with a length between 9 and 15 nt (~70% compared to ~20%). In terms of speed STAR is 4-6 fold faster than OLego, and ~10 fold faster than TopHat2. This is not surprising because STAR uses uncompressed suffix arrays to index the reference genome. Consequently, STAR has a large memory footprint (>24G for mouse genome), which is 6 times larger than TopHat2 and OLego (<4G in general).

[Table 1]

In our previous comparison, we also ran OLego on an RNA-Seq dataset from mouse retina[21] and identified 1,665 micro-exons between 9 nt and 27 nt, among which 630 were novel. We selected 15 potential novel cassette exons and performed RT-PCR validations. All these micro-exons were successfully validated and the inclusion ratios observed in RNA-Seq data and RT-PCR were highly correlated (Pearson correlation coefficient R=0.85)[10].

In this new test, we ran OLego v1.1.5, TopHat v2.0.11 and STAR v2.3.0e on the same dataset and extracted the micro-exons discovered by each program. The new version of OLego found even more micro-exons compared to the old version (1,792 vs. 1,665). Compared to OLego, TopHat2 found substantially fewer (713) exons, including 566 exons annotated in the inclusive gene models (combined from 11 sources[10,21]) and 431 exons annotated in RefSeq (these numbers are 1,067 and 731 for OLego). OLego also identified more novel micro-exons of high confidence (i.e., cassette exons flanked by annotated exons and supported by exon-junction reads[10], 452 vs. 70). STAR identified more micro-exons in the range of 9~39nt (2,249) overall. However, only 898 of them are in the inclusive gene models, and only 320 un-annotated exons fell into the high-confidence novel exon category (which together accounts for 1,218 exons or 54% of total, as compared to 1,519 exons or 85% of total for OLego). These results indicate that STAR has a higher false positive rate and lower sensitivity compared to OLego in discovery of small and micro-exons (Table 2).

[Table 2]

We then looked into the 15 novel micro-exons successfully validated in our previous study[10]. Among these, STAR and TopHat2 identified 14 and 9 exons (both flanking exon junctions were identified with at least one supporting read), respectively (Supplementary Table 2). In addition, we estimated the inclusion ratios of these exons based on alignments by each program and correlated them with the inclusion ratios estimated from RT-PCR validations (Figure 2). We found that TopHat2 and STAR substantially underestimated the inclusion ratios of a portion of these micro-exons, because many reads supporting the inclusion isoforms failed in alignment (Supplementary Table 2).

[Figure 2]

[Supplementary Table 1]

[Supplementary Table 2]

Figure 1: Workflow of OLego. Exonic alignment is attempted first. Exon-junction alignment is then performed for unmapped reads through the following steps: seeding, seed mapping and hit clustering, candidate-exon identification and extension, single- and double- anchor searches. Finally, exons and exon-junctions are connected to identify the optimal complete alignment for each read.

Figure 2: Comparison of OLego, TopHat2 and STAR on micro-exon discovery and quantification. A set of 15 micro-exons validated by RT-PCR was used for this analysis. (a) An IGV screenshot shows a micro-exon in *Ykt6* (15nt) identified by OLego but missed by TopHat2 and STAR. (b) Correlations between inclusion ratios estimated from RNA-Seq data based on alignment by each program and RT-PCR validations.

Table 1: Comparison of OLego, TopHat2 and STAR using simulated RNA-Seq data. Two simulated datasets with different read lengths (10 million paired-end reads, 100 nt or 150 nt) were used for comparison. Sensitivity and accuracy of OLego (v1.1.5), TopHat (v2.0.11) and

STAR (v2.3.0e) on both exon junction discovery and small exon discovery were evaluated. Mapping speed is based on 16 Intel Xeon CPU cores (2.0 GHz).

Table 2. Numbers of micro-exons discovered in RNA-Seq data by different programs.

Supplementary Table 1. Default maximum number of mismatches or indels allowed for different read lengths.

Supplementary Table 2. Micro-exon discovery in real data. The numbers of exon-junction reads mapped to the 15 RT-PCR-validated novel micro-exons by different programs in RNA-Seq data from mouse retina are shown. The inclusion ratios of these exons were computed using these junction reads and compared to the inclusion ratios estimated from RT-PCR analysis.

DISCUSSION:

OLego was specifically designed to align spliced RNA-Seq reads with high sensitivity using small seeds. As demonstrated by both simulated tests and real data analysis, this tool allows us to discover exon junctions and small exons *de novo* at high accuracy and sensitivity. In addition, despite the more exhaustive search at high resolution, OLego achieves high efficiency of read mapping, making it suitable for novel exon and exon-junction discovery in large-scale RNA-Seq data.

A unique feature of OLego is its ability to detect small and micro-exons which are generally more difficult to identify and hence frequently missed in previous studies and databases. These exons are particularly interesting, because many of them are alternatively spliced and under tissue-specific regulation[22,23]. In our comparisons, OLego identified substantially more small exons of high confidence from both simulated and real data. By analyzing RNA-Seq data obtained from various conditions, we will be able to obtain novel insights into these exons, which are excellent models to study mechanisms and function of alternative splicing regulation, given the very limited information encoded in the exon sequences.

To achieve the best sensitivity of this algorithm, we recommend users to specify a suitable combination of seed size and overlapping length (see Section 2.1.3) based on read length, and provide a comprehensive exon junction database (Section 1.3.1) during mapping. A remapping step can further rescue more reads mapped to novel junctions. To make the program even more sensitive to micro-exons, the minimum exon length can be decreased to as small as 6 nt using option "--min-exon" (the default is 9 nt). The accuracy of exon junction discovery can also be further improved by applying more stringent filters based on the numbers of supporting reads and anchor sizes.

To find a balance between sensitivity, accuracy and speed, we limit the *de novo* exon junction search to canonical (GT/AG) splice sites, because they account for about 99% of the mammalian splice sites. In addition, by default, we require prefect matches during certain steps in which querying of short sequences is involved (e.g., seed mapping, single-anchor search and small exon identification) to avoid very extensive search. However, these heuristics have relatively minimal effects on sensitivity in practice, according to the results in both simulated and real data tests.

The protocol presented in this paper can be readily combined with other tools for downstream analysis. OLego outputs the alignments in the standard SAM format which can be easily manipulated by SAMtools (e.g., converted to BAM files). Therefore, alignment results can serve as input for various tools for a range of different types of analysis, such as transcriptome reconstruction, differential gene expression and alternative splicing.

ACKNOWLEDGMENTS:
The work was supported by grants from National Institutes of Health (NIH) [R00GM95713 to C.Z.] and Simons Foundation Autism Research Initiative (297990).

DISCLOSURES:
The authors have no conflict of interests to disclose.

REFERENCES:
1 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63, doi: 10.1038/nrg2484 (2009).
2 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5**, 621-628, doi: 10.1038/nmeth.1226 (2008).
3 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi: 10.1093/bioinformatics/btp120 (2009).
4 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178, doi: 10.1093/nar/gkq622 (2010).
5 Au, K. F., Jiang, H., Lin, L., Xing, Y. & Wong, W. H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* **38**, 4570-4578, doi: 10.1093/nar/gkq211 (2010).
6 Huang, S. *et al.* SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. *Front Genet* **2**, doi: 10.3389/fgene.2011.00046 (2011).
7 Zhang, Y. *et al.* PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics* **28**, 479-486, doi: 10.1093/bioinformatics/btr712 (2012).
8 Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881, doi: 10.1093/bioinformatics/btq057 (2010).
9 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi: 10.1186/gb-2009-10-3-r25 (2009).
10 Wu, J., Anczukow, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* **41**, 5149-5163, doi: 10.1093/nar/gkt216 (2013).
11 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi: 10.1093/bioinformatics/btp352 (2009).
12 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi: 10.1038/nbt.1621 (2010).
13 Anders, S., Pyl, P. T. & Huber, W. *HTSeq: A Python framework to work with high-throughput sequencing data*. (2014).

14 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi: 10.1186/gb-2010-11-10-r106 (2010).
15 McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288-4297, doi: 10.1093/nar/gks042 (2012).
16 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi: 10.1038/nbt.1883 (2011).
17 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829, doi: 10.1101/gr.074492.107 (2008).
18 Ferragina, P. & Manzini, G. Opportunistic data structures with applications. *Proc FOCS 2000*, 390-398 (2000).
19 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi: 10.1093/bioinformatics/bts635 (2013).
20 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi: 10.1186/gb-2013-14-4-r36 (2013).
21 Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518-2528, doi: 10.1093/bioinformatics/btr427 (2011).
22 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523, doi: 10.1016/j.cell.2014.11.035 (2014).
23 Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res* **25**, 1-13, doi: 10.1101/gr.181990.114 (2015).

Figure 1
Click here to download Figure: Figures 1.pdf

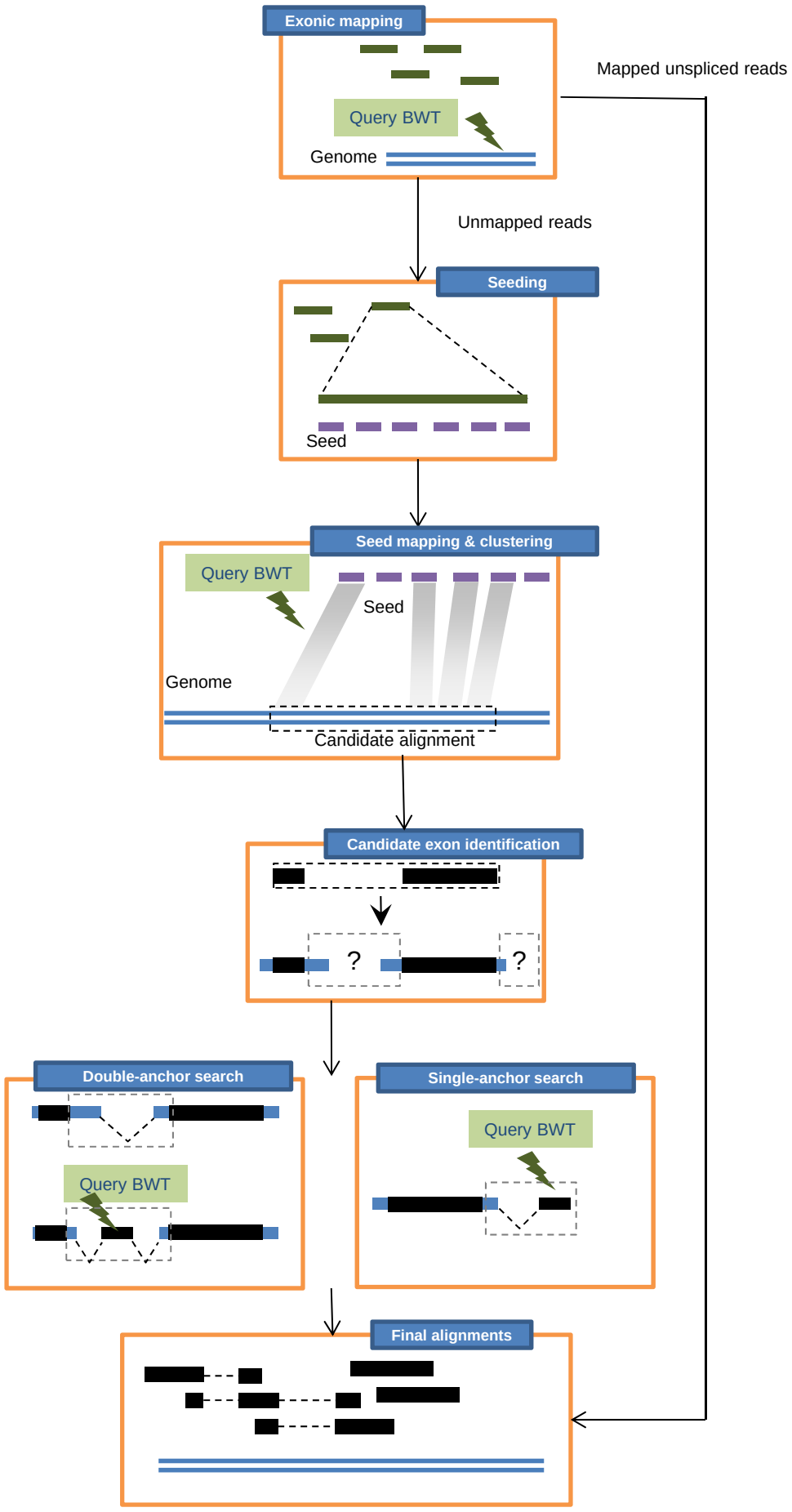

Figure 2
Click here to download Figure: Figures 2.pdf

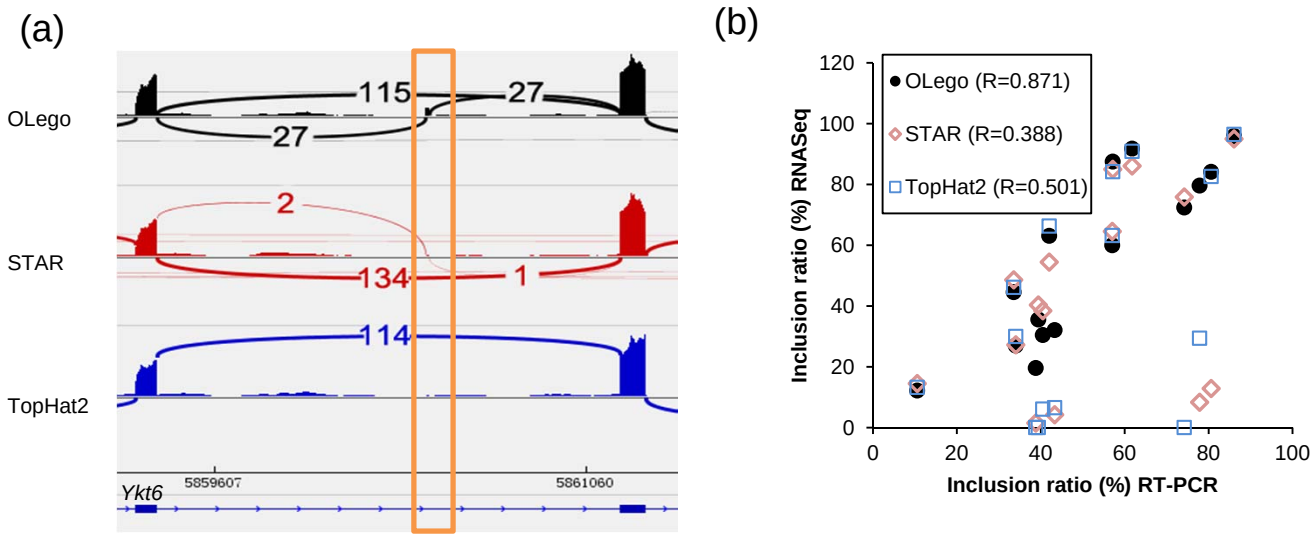

Table 1
Click here to download Table: table 1.xlsx

	100-nt reads (178,449 junctions)		
	Olego v1.1.5	**TopHat v2.0.11**	**STAR v2.3.0e**
Total junctions identifed	168,901	154,135	170,751
Found true junctions	165,632	150,575	166,770
Missed true junctions	12,816	27,873	11,678
PPV (junctions)	0.981	0.977	0.977
FNR (junctions)	0.072	0.156	0.065
PPV (exons:9~39nt)	0.973	0.917	0.963
FNR (exons: 9~39nt)	0.111	0.368	0.132
PPV (exons:9~15nt)	0.942	0.508	0.893
FNR (exons: 9~15nt)	0.225	0.918	0.712
Time (hour, 16 cores)	0.51	1.31	0.13

150-nt reads (189,106 junctions)		
Olego v1.1.5	**TopHat v2.0.11**	**STAR v2.3.0e**
181,871	169,974	184,691
178,245	166,003	180,516
10,861	23,103	8,590
0.98	0.977	0.977
0.057	0.122	0.045
0.968	0.883	0.959
0.08	0.31	0.093
0.969	0.25	0.934
0.204	0.894	0.706
1.07	1.72	0.17

Table 2

Click here to download Table: table 2.xlsx

	OLego v1.1.5	TopHat v2.0.11	STAR v2.3.0e
Total	1,792	713	2,249
Annotated in the inclusive gene models	1,067	566	898
Annotated in RefSeq	731	431	642
High-confidence novel exons	452	70	320
Other novel exons	273	77	1,031

Table S1

Click here to download Table: Supp table 1.xlsx

Supplementary Table 1. Default maximum number of mismatches or indels allowed for different read le

Read Length (nt)	Max difference
17	1
20	2
45	3
73	4
104	5
137	6
172	7
208	8
244	9

ngths

Table S2
Click here to download Table: Supp table 2.xlsx

Supplementary Table 2. Micro-exon discovery in real data

Gene symbol	Exon length (nt)	Upstream exon pos	Exon pos
Madd	9	chr2:91013270-91013350	chr2:91012484-91012492
Cadps	15	chr14:13282236-13282383	chr14:13274366-13274380
Amph	9	chr13:19192441-19192523	chr13:19193771-19193779
Ank3	12	chr10:69390443-69390566	chr10:69392195-69392206
Doc2b	27	chr11:75599595-75599674	chr11:75599019-75599045
Ykt6	15	chr11:5859309-5859382	chr11:5860430-5860444
Heg1	24	chr16:33720735-33721088	chr16:33722365-33722388
Mll3	23	chr5:24810496-24810559	chr5:24809915-24809937
Fermt2	21	chr14:46084399-46084620	chr14:46082390-46082410
Rab3gap1	21	chr1:129835623-129835725	chr1:129838171-129838191
Elmod1	24	chr9:53772207-53772275	chr9:53771749-53771772
Kif21a	12	chr15:90779437-90779475	chr15:90778755-90778766
Cd9912	18	chrX:68682222-68682341	chrX:68678590-68678607
Ubr5	27	chr15:37900017-37900102	chr15:37898682-37898708
Rims2	12	chr15:39123648-39123858	chr15:39137638-39137649

		Olego		
Downstream exon pos	**Strand**	**In reads**	**Ex reads**	**In% (RNA-seq)**
chr2:91010842-91010904	-	350	66	84.13
chr14:13273355-13273429	-	64	116	35.56
chr13:19194741-19194879	+	51	108	32.08
chr10:69395157-69395236	+	43	11	79.63
chr11:75595123-75595197	-	47	2	95.92
chr11:5861191-5861291	+	28	115	19.58
chr16:33725511-33725729	+	45	56	44.55
chr5:24808524-24808756	-	53	31	63.10
chr14:46081791-46081915	-	49	7	87.50
chr1:129838951-129839155	+	45	30	60.00
chr9:53769130-53769180	-	61	165	26.99
chr15:90774196-90774332	-	56	5	91.80
chrX:68677199-68677264	-	28	64	30.43
chr15:37898093-37898260	-	34	245	12.19
chr15:39176856-39177166	+	21	8	72.41

STAR			TopHat2		
In reads	**Ex reads**	**In% (RNA-seq)**	**In reads**	**Ex reads**	**In% (RNA-seq)**
9	61	12.86	352	74	82.63
75	111	40.32	0	131	0.00
5	113	4.24	8	114	6.56
1	11	8.33	5	12	29.41
57	3	95.00	55	2	96.49
2	134	1.47	0	114	0.00
53	56	48.62	48	56	46.15
55	46	54.46	61	31	66.30
51	9	85.00	48	9	84.21
71	39	64.55	55	32	63.22
58	155	27.23	95	222	29.97
31	5	86.11	70	7	90.91
25	40	38.46	4	62	6.06
36	212	14.52	41	269	13.23
22	7	75.86	0	9	0.00

In% (RT-PCR)
80.61
39.44
43.30
77.84
86.05
38.79
33.56
42.02
57.14
57.04
34.03
61.78
40.49
10.57
74.20

Excel Spreadsheet- Table of Materials/Equipment

Click here to download Excel Spreadsheet- Table of Materials/Equipment: JoVE_Materials.xls

Name of Material/ Equipment	Company	Catalog Number	Comments/Description

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: OLego: de novo discovery of novel exon junctions and small exons from RNA-Seq data

Author(s): Jie Wu, Huijuan Feng and Chaolin Zhang

Item 1 (check one box): The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via: [x] Standard Access [ ] Open Access

Item 2 (check one box):

[x] The Author is NOT a United States government employee.

[ ] The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

[ ] The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. Defined Terms. As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments;* "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. Background. The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. Grant of Rights in Article. In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. Retention of Rights in Article. Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. Grant of Rights in Video – Standard Access. This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. Grant of Rights in Video – Open Access. This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. Government Employees. If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. Likeness, Privacy, Personality. The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. Author Warranties. The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. JoVE Discretion. If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. Indemnification. The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in *JoVE* or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. Fees. To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. Transfer, Governing Law. This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name:	Chaolin Zhang
Department:	Department of Systems Biology, Department of Biochemistry and Molecular Biophysics
Institution:	Center for Motor Neuron Biology and Disease, Columbia University
Article Title:	OLego: de novo discovery of novel exon junctions and small exons from RNA-Seq data
Signature:	 Date:

Please submit a signed and dated copy of this license by one of the following three methods:

1) Upload a scanned copy of the document as a pfd on the JoVE submission site;
2) Fax the document to +1.866.381.2236;
3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

Rebuttal Comments

Click here to download Rebuttal Comments: response.docx

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.

We have proofread the manuscript carefully.

2. Please revise the tables to be in the .xls or .xlsx format instead of a pdf.

We have now provided these tables in excel format.

3. Please remove the section headers in the Introduction regarding the workflow.

We have removed the section headers as directed and reorganize this section.

4. Please revise the scriptwriter's guide. Please note that only protocol sections should be presented in the scriptwriter's guide. Similarly, each step in the protocol must have a visual accompanying the step and its action.

This document is meant to guide the scriptwriter through visualization of the protocol, which typically contains the most problematic part of the manuscript to film.

We have rewritten the guide and included much more details.