Insect Molecular Biology

Insect Molecular Biology (2017) **26**(3), 298–307 doi: 10.1111/imb.12294

Single molecule RNA sequencing uncovers *trans*-splicing and improves annotations in *Anopheles stephensi*

X. Jiang*†‡, A. B. Hall*‡, J. K. Biedler*† and Z. Tu*†‡

**Program in Genetics Bioinformatics and Computational Biology, Virginia Tech, Blacksburg, VA, USA; †Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA; and ‡Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*

Abstract

Single molecule real-time (SMRT) sequencing has recently been used to obtain full-length cDNA sequences that improve genome annotation and reveal RNA isoforms. Here, we used one such method called isoform sequencing from Pacific Biosciences (PacBio) to sequence a cDNA library from the Asian malaria mosquito *Anopheles stephensi*. More than 600 000 full-length cDNAs, referred to as reads of insert, were identified. Owing to the inherently high error rate of PacBio sequencing, we tested different approaches for error correction. We found that error correction using Illumina RNA sequencing (RNA-seq) generated more data than using the default SMRT pipeline. The full-length error-corrected PacBio reads greatly improved the gene annotation of *Anopheles stephensi*: 4867 gene models were updated and 1785 alternatively spliced isoforms were added to the annotation. In addition, six *trans*-splicing events, where exons from different primary transcripts were joined together, were identified in *An. stephensi*. All six *trans*-splicing events appear to be conserved in Culicidae, as they are also found in *Anopheles gambiae* and *Aedes aegypti*. The proteins encoded by *trans*-splicing events are also highly conserved and the orthologues of these proteins are *cis*-spliced in outgroup species, indicating that *trans*-splicing may arise as a mechanism to rescue genes that broke up during evolution.

First published online 9 February 2017.

Correspondence: Z. J. Tu, Department of Biochemistry, Virginia Tech, Blacksburg, VA 24061, USA; e-mail: jaketu@vt.edu

Keywords: SMRT sequencing, Iso-Seq, malaria, mosquito.

Introduction

Single molecule real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio) is a third-generation sequencing method that provides long reads that go well beyond kilobases. It has been used to facilitate assemblies and analyses of genomes of various species including a few insects (Jiang *et al.*, 2014; Berlin *et al.*, 2015; Hall *et al.*, 2016). The long reads offered by SMRT sequencing has also been used, in the form of isoform sequencing (Iso-Seq), to obtain full-length cDNA sequences that could improve genome annotation and reveal RNA isoforms (Sharon *et al.*, 2013; Abdel-Ghany *et al.*, 2016). It is for this purpose that we used Iso-Seq to sequence a cDNA library from the adult males of the Asian malaria mosquito, *Anopheles stephensi*.

In addition to significantly improving the annotation of the recently published *An. stephensi* genome (Jiang *et al.*, 2014), full-length or long-read cDNA sequencing revealed a very interesting evolutionary phenomenon, namely *trans*-splicing. RNA splicing is the process that forms a mature messenger RNA (mRNA) by joining exons and removing introns from the precursor mRNA (pre-mRNA). For majority of eukaryotic genes, splicing is mediated in *cis* by the spliceosome. The spliceosome brings the exons on both sides of an intron into close proximity and then cleaves the 5′ splice site and ligates the 5′ splice site to the branch point in the intron. This produces a lariat-structured RNA. The spliceosome then cleaves the 3′ splice site, ligates exons and releases the lariat. Intriguingly, splicing can also occur in *trans*, where exons from more than one separate pre-mRNA are joined. *Trans*-splicing is well studied in trypanosomes and nematodes, where a spliced leader RNA is spliced to the 5′ ends of the first exon on many pre-mRNAs (Douris *et al.*, 2010).

In higher eukaryotes, *trans*-splicing does not involve spliced leaders. *Trans*-splicing has been observed in fruit flies, rodents, humans and many other organisms (Caudevilla *et al*., 1998; Dorn *et al*., 2001; Horiuchi *et al*., 2003; Herai & Yamagishi, 2010; Lasda & Blumenthal, 2011; Shao *et al*., 2012). Based on the relationship of the two pre-mRNAs joined in *trans*-splicing, *trans*-splicing can be grouped into three categories: interallelic, intragenic and intergenic. A well-known example of interallelic *trans*-splicing is the *longitudinals lacking* (*lola*) gene, which is essential for development of the nervous system in *Drosophila* (Horiuchi *et al*., 2003). *Trans*-splicing of *lola* was inferred from interallelic complementation tests on lethal mutations in *lola* exons and verified by allelic Single nucleotide polymorphism makers in *Drosophila* hybrids. Later, a study utilizing RNA sequencing data from *Drosophila* hybrids identified more *trans*-splicing between homologous alleles, suggesting interallelic *trans*-splicing occurs commonly. Intragenic *trans*-splicing is the scenario where splicing occurs between two pre-mRNAs from the same genetic loci. The two pre-mRNAs can come from the same strand, and an example is the *Carnitine O-octanoyltransferase* gene in the rat liver where the exons are duplicated in the mRNA (Caudevilla *et al*., 1998). They can also come from the opposite strand, like the fruit fly *modifier of mdg4 (mod(mdg4))* genes (Dorn *et al*., 2001). Intergenic *trans*-splicing occurs when the pre-mRNAs come from different genes. These genes can be located at distant genomic loci on different chromosomes. For example, the *bursicon* gene in *Anopheles gambiae* is *trans*-spliced from three exons on chromosome arm 2L and one exon on chromosome arm 2R (Robertson *et al*., 2007).

The understanding of *trans*-splicing has been significantly improved by the advent of next generation sequencing technology. *Trans*-splicing events are generally identified by finding non-colinear transcripts, which are RNA-seq reads that fail to align to the corresponding DNA sequences in the reference genome in a linear pattern. Although this approach cannot detect interallelic and other *trans*-splicing events that generate colinear transcripts, a significant number of *trans*-splicing events have been detected (Davidson *et al*., 2015; Liu *et al*., 2015). For example, a recent study of eight insect species across five orders detected 1627 *trans*-splicing events (Kong *et al*., 2015). Some of the *trans*-splicing events are conserved across species, indicating that *trans*-splicing is not transcriptional noise and is likely to be functionally important (Kong *et al*., 2015). Moreover, the previous notion that fusion transcripts are the markers of tumour cells has been called into question, as several studies and the Encyclopedia of DNA Elements (ENCODE) project demonstrated that chimeric RNAs are common in normal tissues and cell lines (Gingeras, 2009).

The SMRT Iso-Seq technology from Pacific Biosciences has been applied to discover *trans*-splicing or fusion genes (Weirather *et al*., 2015). Iso-Seq can generate full-length transcript sequences from the polyA-tail to the 5′ end, providing isoform-level resolution of transcriptome data. Full-length or long-read cDNA sequences obtained by Iso-Seq provide significant advantages in the identification and characterization of *trans*-splicing events compared with short RNA-seq data obtained from Illumina sequencing, which can only provide information on the small segment around the *trans*-spliced site. Furthermore, the structure of full *trans*-spliced mRNA is hard to infer from short RNA-seq data, owing to the fact that the majority of reads generated from the *trans*-spliced mRNAs cannot be differentiated from the ones from the *cis*-spliced mRNA. This is not an issue for Iso-Seq data, which provide reads representing full-length transcripts.

In this research, we used both SMRT Iso-Seq data and Illumina RNA-seq data to detect *trans*-splicing events in the Asian malaria mosquito, *An. stephensi*. To eliminate false positive discoveries owing to PCR chimeras and transcriptional noise, only *trans*-splicing events supported by both data sets were used. In total, we identified six *trans*-splicing events in *An. stephensi*, all of which are also found and conserved in *Aedes aegypti*. The proteins encoded by the *trans*-spliced mRNAs are also highly conserved and their orthologues are colinearly transcribed in Culicidae outgroups. This finding indicates that the need to preserve the mRNA completeness and protein function of genes broken up during the course of evolution may be the driving force behind *trans*-splicing. As indicated earlier, we also used the Iso-Seq data to improve the *An. stephensi* genome annotation.

Results

Error correction with Illumina RNA-seq outperforms SMRT pipeline in both data accuracy and data quantity

Three Iso-Seq libraries with insert sizes of 1 − 2, 2 − 3 and 3 − 6 kb were sequenced with four PacBio SMRT cells each (Sequence Read Archive (SRA) accession: SRP081051). Each SMRT cell produced around 50 000 to 60 000 reads of insert. Full-length transcripts were defined by the presence of a 5′ primer, 3′ primer and a polyA tail in the reads of insert. Approximately 38, 31 and 9% of the reads of insert were identified as full-length, nonchimeric reads for the 1 − 2, 2 − 3 and 3 − 6 kb insert size libraries, respectively (Table 1). During the clustering process of the SMRT pipeline, on average two to three full-length, nonchimeric reads could be clustered as one consensus isoform for the 1 − 2 and 2 − 3 kb libraries. For the 3 − 6 kb size library, most consensus isoforms were only represented by a single full-length, nonchimeric read. Therefore, in order to obtain a sufficient number of long reads for analysis, larger insert sizes require a deeper sequencing depth. Less than 17%

Table 1. Pacific Biosciences single molecule real-time pipeline output metrics

Data type	1–2 kb	2–3 kb	3–6 kb
Number of reads of insert	248 903	210 594	202 440
Number of five prime reads	136 440	100 147	54 463
Number of three prime reads	146 668	109 776	65 571
Number of poly-A reads	142 375	106 746	61 281
Number of filtered short reads	13 209	8876	7165
Number of non-full-length reads	138 909	135 425	175 629
Number of full-length reads	96 785	66 293	19 646
Number of full-length nonchimeric reads (bp)	96 170	65 955	19 094
Average full-length nonchimeric read length	1388 (bp)	1948 (bp)	4357 (bp)
Number of consensus isoforms	35 248	30 793	17 405
Average consensus isoform read length (bp)	1465 (bp)	2044 (bp)	4376 (bp)
Number of polished high-quality isoforms	7414	6075	636
Number of polished low-quality isoforms	27 834	24 718	16 769

bp, base pairs.

of the total consensus isoforms were polished as high-quality isoforms by Quiver. This indicates that in order to obtain enough high accuracy data through the SMRT pipeline alone, the number of cells sequenced for each library should be higher than the Iso-Seq data used in this paper to provide enough coverage, particularly for long-insert libraries. All polished isoforms were used for the further analysis to avoid discarding useful information.

As an alternative, we used RNA-seq data to error correct the Iso-Seq data. Of 1321 million base pairs of reads of insert, 668 000 000 base pairs were corrected by Proovread (Hackl *et al.*, 2014) with a high level of accuracy (Table 2). The mean of the average quality scores of each read of insert improved from 14.73 to 36. 4 after correction, indicating a significant improvement in accuracy. Compared with polished high-quality isoforms from the SMRT pipeline, 30 times more base pairs were corrected, although the mean value of the median quality scores of high accuracy corrected reads was slightly lower. In addition, the mean value of median quality scores of high accuracy corrected reads was 37.85, equivalent to an accuracy above 99.98%. This result demonstrates that it is favourable to use high-quality short reads for correction of Iso-Seq reads. This is also more economical as the RNA-seq data needed for the analysis is significantly cheaper than sequencing additional Iso-Seq SMRT cells.

Proteins encoded by trans*-splicing are conserved*

490 *trans*-splicing events were detected based on RNA-seq processed by MapSplice (Wang *et al.*, 2010). 3359 *trans*-splicing events were found by the PacBio pbtranscript-tofu package. In both RNA-seq and Iso-Seq technology, PCR chimeras can cause a large number of false positive results. Therefore, we set a criterion that splice junctions must be supported by both data sets to be considered as valid. In the end, six pairs of splice junctions were identified. All these six *trans*-splicing events are interchromosomal (Table 3).

Trans-spliced mRNA 1 (Tm1) is one mRNA created from two *trans*-splicing events (Fig. 1A). The pre-mRNAs of Tm1 are located in chromosome elements 1, 2 and 3. This mRNA has five exons: two shared with gene *ASTEI07024*, one shared with gene *ASTEI02601* and one shared with the intron of gene *ASTEI04882*. The mRNA encodes a 475 amino acid (aa) peptide with two domains. The first domain is similar to microphthalmia/transcription factor E (MiT/TFE, IPR031867), which is shared with gene *ASTEI07024*. The second domain is similar to the basic helix-loop-helix domain (IPR011598), shared with gene *ASTEI02601*. *ASTEI07024* is a mosquito-specific gene. Alignment of peptide sequences of *ASTEI02601* to its *Drosophila* orthologue *FBgn0041164* revealed that the exon utilized by the Tm1 *trans*-splicing event contributes to amino acid sequences that do not exist in their *Drosophila* orthologues. This is

Table 2. Comparisons of processed isoform-sequencing data

		Polished isoforms	Polished high-quality isoforms	Polished low-quality isoforms	High-accuracy corrected reads	Complete corrected reads	Reads of insert
Mean quality score	Min.	0	0.01	0	16.01	0.11	0.98
	Max.	17.36	40	12.74	39.81	37.36	27.31
	Average	12.7	37.15	7.72	36.4	28.25	14.73
	Median	13.67	39.84	8.34	37.85	30.87	15.45
Length (base pairs)	Total (million)	190.6	23.8	166.8	668.6	1254.9	1321.4
	Min.	330	567	330	70	249	11
	Max.	25502	5673	25502	8098	31415	31532
	Average	2284.59	1686.06	2406.55	1284.73	2005.91	1996.32
	Median	1829	1630	1891	1145	1636	1641
	N25	4325	2005	4404	1939	4074	4214
	N50	2244	1742	2471	1444	2234	2342
	N75	1720	1416	1780	1046	1565	1624
	N90	1358	1193	1407	727	1181	1206
	N95	1213	1113	1244	613	919	928

Table 3. *Trans*-splicing sites in three species of Culicidae

Anopheles stephensi

	Donor				Acceptor			
	Contig	Position	Strand	Sequence around splice site*	Contig	Position	Strand	Sequence around splice site*
Tm1.1	stl-e1	8284604	–	ATCAAGAAGGATAATCATAACTGCA|GT	stl-e2	29650741	–	AC|TGTTGGTTGGGAGGGTGAACAACGG
Tm1.2	stl-e3	29650494	–	CGCAATGCTGCTGAAGCGTGTTGCG|GT	stl-e2	49700929	+	AG|GAGTTGCAAAATCAGGTTGATTTTC
Tm2	stl-e1	10533884	–	ATCTGTTCGATGATGATCGAAAGTT|GT	stl-e2	13176261	+	AG|TATCGCACACGGTACAAGATTTGGT
Tm3	stl-e4	31171572	+	CATCACTACTCCTGCCATCTGTGTC|GT	stl-e1	4936658	–	AG|GGCGTATTTATCTTCAACATCGTGC
Tm4	stl-e4	31171572	+	CATCACTACTCCTGCCATCTGTGTC|GT	stl-e1	4948239	+	AG|GGCGTATTTATCTTCAACATCGTGC
Tm5	stl-e1	14042259	+	AAAGCTGAAAGATGTCGTTGATCAG|GT	stl-e2	31153557	–	AG|CAAAGTTCCTTCGCCTGCTGGCTAC

Anopheles gambiae

	Donor				Acceptor			
	Contig	Position	Strand	Sequence around splice site*	Contig	Position	Strand	Sequence around splice site*
Tm1.1	X	14803808	–	ATCAAGAAGGACAATCACAACTGCA|GT	2L	40170764	+	AG|TTGAGCGGCGTCGTCGATTCAACAT
Tm1.2	2L	40171011	+	CGCAATGCTGCAGAAGCGTGTCGCG|GT	2R	56727316	+	AG|GAGTTGCAAAACCAAGTTGACTTTC
Tm2	X	4334405	–	ATCTGCTCGATGATGATCGAAAGTT|GT	2R	13216641	–	AG|TATCGCACACGGTACAGGATTTGGT
Tm3	3R	42647718	+	CATCACCACTCCTGCCATCTGTGTC|GT	X	8704053	–	AG|GGCGTATTCATCTTCAACATCGTGC
Tm4	3R	42647718	+	CATCACCACTCCTGCCATCTGTGTC|GT	X	8717291	+	AG|GGCGTATTCATCTTCAACATCGTGC
Tm5	X	1025131	+	AAAGCTGAAAGATGTCGTTGATCAG|GT	2R	27239887	+	AG|CAAAGTTCCTTCGCCTGCTGGCTAC

Aedes aegypti

	Donor				Acceptor			
	Contig	Position	Strand	Sequence around splice site*	Contig	Position	Strand	Sequence around splice site*
Tm1.1	supercont1.30	2525719	+	ATCAAGAAGGATAACCATAACTGCA|GT	supercont1.497	69990	–	AG|TCGAGAGGCGTCGTCGTTTCAATAT
	supercont1.322	325659	–	ATCAAGAAGGATAACCATAACTGCA|GT	supercont1.497	69990	–	AG|TCGAGAGGCGTCGTCGTTTCAATAT
Tm1.2	supercont1.497	69743	–	CATCATTCTGCAGCACAAACTGGCG|GT	supercont1.541	339301	–	AG|GAAATGCAAAAGCAGCTCGACTATT
Tm2	supercont1.75	2116294	–	GTTTGCTCGATGATGATCGAAAGTT|GT	supercont1.187	287309	–	AG|TATCCCACACCGTTCAGGATTTGGT
Tm3/Tm4	supercont1.496	104246	–	TATCACGACTCCGACGATCTGTGTT|GT	supercont1.715	64386	+	AG|GGCGTATTCATCTTCAACCTGGTGC
Tm5	supercont1.54	128027	–	GAAGCTGAAGGACGTAGTCGATCAG|GT	supercont1.179	490627	–	AG|CAAAGTTCCTTTGCCTGCTGGTTAC

*Splice junctions are indicated as '|'.

.25 bp upstream of donor site and 25 bp downstream of acceptor site are shown.

Tm1–5, *trans*-spliced mRNA 1–5.

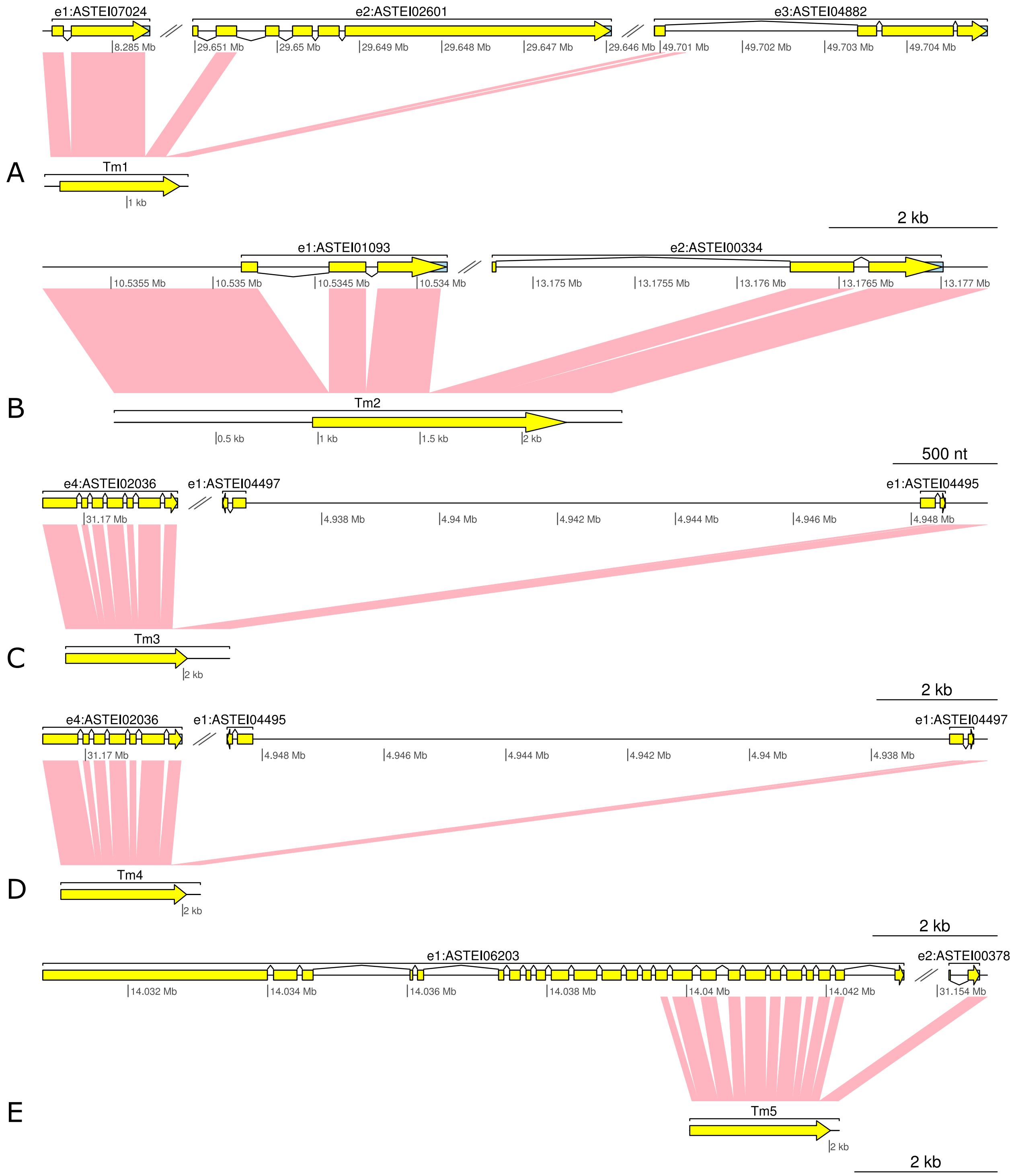

Figure 1. *Trans*-splicing events in *Anopheles stephensi*. In each panel, the top section stands for the genomic regions to which the *trans*-spliced mRNA aligns. Related gene annotation is also provided. The bottom section stands for the full-length mRNA sequence. The pink blocks in the middle represent matches between genomic sequence and mRNA. Yellow bars represent coding region. Abbreviations: Tm1–5, *trans*-spliced mRNA 1–5. [Colour figure can be viewed at wileyonlinelibrary.com]

also the case for the exon shared between Tm1 and *ASTEI04882* when we aligned the peptide sequence of ASTEI04882 to that of the orthologue FBgn0034176. No obvious *Drosophila* orthologue for the complete Tm1 protein has been observed. There is some similarity between the protein of FBgn0263112 and the peptide sequences coded by the first three exons, particularly the third exon of Tm1 (28.87% identify). Interestingly, the complete 474 aa peptide sequence coded by Tm1 is highly homologous to some dipteran outgroups. The examples include gene *XP_011304746* in *Fopius arisanus* (33.17% identity) and *XP_012252483* in *Athalia rosae* (31.53% identity). The mRNAs of these genes in the outgroup species are colinear to the genome, and thus likely to be *cis*-spliced.

The exons of Tm2 come from two *cis*-spliced genes *ASTEI01093* and *ASTEI00334* (Fig. 1B). Neither of these genes has orthologues outside of mosquitoes. The protein encoded by the *trans*-spliced Tm2 consists of 515 amino acids, which belongs to the neurotransmitter-gated ion-channel (IPR006201) family. This Tm2 protein is orthologous to the *Drosophila* gene *FBgn0037950* with high similarity (83.57% identity). This *trans*-spliced protein is conserved in Insecta. All of its nonmosquito orthologues appear to be *cis*-spliced.

The donor sites of the *trans*-splicing event of Tm3 and Tm4 are identical (Fig. 1C, D). The genes on the acceptor site of these two events are paralogous to each other. The paralogues are in close proximity but of different orientation, probably owing to a tandem duplication. The coding sequences of the two paralogues are identical, and consequently Tm3 and Tm4 encode identical proteins. The 3' untranslated region (UTR) is different in Tm3 and Tm4. *Trans*-splicing exists in both *ASTEI004497* and *ASTEI004495*, as supported by full-length transcripts covering the 3' UTR in both genes. The encoded protein is a neurotransmitter symporter (IPR000175). The *Drosophila* gene *FBgn0181657* is annotated as an orthologue to *ASTEI02036* but in fact, sequence alignment showed that this is only a partial match, and FBgn0181657 is more similar and aligns over its full length to the fusion protein of ASTEI02036 and ASTEI004497/ASTEI004495. This fusion protein is highly conserved across Insecta. Like Tm2, Tm3 and Tm4 orthologues outside mosquitoes are *cis*-spliced.

The longest read that we obtained from Tm5 is 2142 bp (Fig. 1E). This read probably represents an mRNA with an incomplete 5' end, because the start codon is missing. Nevertheless, this read covers the *trans*-splicing site between chromosome elements 1 and 2. Tm5 joins exons from ASTEI06203 and ASTEI00378. The protein Tm5 encodes is uncoordinated protein 13 (IPR027080). It is conserved across Insecta. *FBpp0300963* is annotated as an orthologue to *ASTEI06203*. Interestingly, the last 53-aa sequence encoded by *FBpp0300963* is 76% identical to the 54 aa encoded by the last exon of *ASTEI00378*, whereas it is only 7% identical to the sequence of amino acids encoded by the last exon of *ASTEI06203*. This implies that the fusion protein is ancestral and *trans*-splicing between exons of *ASTEI006203* and *ASTEI00378* is a way to keep the protein intact.

Trans-splicing is highly conserved in *Culicidae*

To investigate whether the above *trans*-splicing events are *An. stephensi* specific or are conserved, we checked the transcriptome data of *An. gambiae* and *Ae. aegypti*. We predicted *trans*-splicing sites using MapSplice (Wang *et al*., 2010) with RNA-seq data as described in the Experimental procedures.

All of the *trans*-splicing events in *An. stephensi* also exist in *An. gambiae* (Table 3). In addition, the chromosomal assignment of the orthologues involved in *trans*-splicing are the same and the sequences around the splice sites are highly similar between these two species. In *Ae. aegypti*, the supercontigs are not assigned to chromosomes and thus chromosomal position cannot be inferred. Based on supercontigs, the orthologues of the *trans*-splicing *An. stephensi* events are observed with a few differences. First, the *trans*-spliced gene may share the exons with a different *cis*-spliced gene. For example, the *ASTEI02601* orthologues *AAEL010693* and *AAEL010696* do not share exons with Tm1 in *Ae. aegypti*. Instead, the shared exon is in their neighbouring gene *AAEL010700*. Second, duplication events are different between Anophelinae and Culicinae: the *ASTEI07024* orthologues are duplicated and located on different supercontigs in *Ae. aegypti*, whereas the orthologue of gene *ASTEI004495*/*ASTEI004497* is a single gene, *AAEL012596*, in *Ae. aegypti*. Interestingly, *trans*-splicing was kept during duplications of these genes. In 17 of the 18 *trans*-splicing events in all three species, the 5' and 3' termini of the introns follow the GU-AG rule. The only exception is the first *trans*-splicing site of Tm1 in *An. stephensi*, where the acceptor site is AC instead of AG.

Improvement of genome annotation

Comparisons of the existing gene annotation of *An. stephensi* and the updated annotation by pasa (Program to Assemble Spliced Alignments) with error-corrected high-accuracy Iso-Seq data are provided in the link http://tu07.fralin.vt.edu/cgi-bin/PASA_r20140417/cgi-bin/status_report.cgi?db=ECRItr. The previous gene annotation is based on the gene annotation software Maker (Holt & Yandell, 2011), where protein homology based and *ab inito* prediction were applied. Transcriptomes were not used by this Maker annotation. In the previous annotation, 11 789 protein-coding genes were annotated. Each gene has only one isoform, which indicates that alternative splicing remains undetected. In addition, genes were mostly annotated with UTRs missing. The updated annotation enhanced the existing one by adding UTRs, identifying alternative spliced isoforms, and adjusting exon boundaries. In total, 3323 genes were updated with the addition of UTRs, 1785 genes were updated with alternatively spliced isoforms and 1923 genes were updated with exons adjusted or gene

Table 4. Annotation improvement in *Anopheles stephensi* using PASA

	No. of gene model updates	No. of alternative splice isoforms to add
EST assembly extends UTRs.	3323	0
EST assembly alters protein sequence, passes validation.	697	0
EST assembly properly stitched into gene structure.	1065	0
EST assembly stitched into gene model requires alternative splicing isoform.	0	1785
EST assembly found capable of merging multiple genes.	161	0
Totals (some models in multiple classes)	4867	1785

EST, expressed sequence tags; UTRs, untranslated regions.

merging. These structural changes of genes altered 1878 protein sequences (Table 4).

One example demonstrating the improvement in the gene annotation can be seen in the annotation of the gene *doublesex* (Suzuki *et al.*, 2001). *doublesex* is a gene essential for sexual dimorphism and it contains male-specific and female-specific isoforms. In our analysis, the Iso-Seq data was obtained from males only, so we would expect to observe only the male isoform. The gene *doublesex* in mosquitoes spans a region of 90 000 base pairs with another gene inserted in one of its introns. As a result, in the majority of Anophelinae, this gene is misannotated as two genes. After the annotation update by PASA (Fig. 2), the two parts of *doublesex*, *ASTEI07080* and *ASTEI07082*, were merged into one complete model. This model is the complete male isoform of *doublesex* as expected.

Discussion

Two separate gene breakup events lead to the formation of trans-spliced gene Tm1. The first one, which separated the third and fourth exons of Tm1, happened before the formation of Diptera. In Culicidae, *trans*-splicing was used to join these two separated exons, whereas this *trans*-splicing either did not happen or was later lost in *Drosophila*. The second gene breakup, which separated exon2 and exon3, happened only in Culicidae. In *Aedes*, the region transcribing the first pre-mRNA of Tm1 was duplicated and both copies maintained their ability to be *trans*-spliced. The breakup of the ancestor genes of the other *trans*-spliced mRNAs described in this paper happened after the formation of Diptera but before Culicidae. All their *Drosophila* orthologues remained as canonical genes that utilize *cis*-splicing, whereas the formation of the complete mRNAs in Anophelinae and *Aedes* relies on *trans*-splicing. The high conservation of the *trans*-splicing sites across three divergent mosquito species indicates a single origin for each *trans*-splicing event.

Although *trans*-splicing has been observed in many higher eukaryotes, its mechanism remains largely unclear. One well-known model is that *trans*-splicing happens through mutually complementary intron sequences. The bases of the introns of the two separate pre-mRNAs will pair with each other, bringing the two molecules together and thus promoting *trans*-splicing (Wally *et al.*, 2012). However, a recent study in *Drosophila* showed that two intronic RNA sequence motifs are critical and perhaps sufficient to initiate *trans*-splicing in the *mod* gene (Gao *et al.*, 2015). In both models, the nucleotide sequences of pre-mRNAs effect the conformation of the RNA-spliceosome complex and then influence splicing, which is essentially the same as *cis*-splicing. It is reasonable to assume that the splicing machinery is no different for *trans*-splicing. *Trans*-splicing is observed across a wide range of eukaryotes and probably exists in all eukaryotes, just like *cis*-splicing (Douris *et al.*, 2010). In addition, the only factor differentiating *cis*-splicing and *trans*-splicing is whether there is more than one pre-mRNA. As a process in three dimensions, splicing requires spatial proximity of splice sites (Hiller *et al.*, 2007; Warf & Berglund, 2010). It may not matter whether the two separate pre-mRNAs directly interact with each other through base pairing or through binding to the spliceosome using specific sequence motifs. As long as the splice sites are spatially close and accessible, *trans*-splicing reactions may occur just as *cis*-splicing.

Splicing greatly diversifies the proteome by promoting the formation of new genes through alternative splicing. Allelic *trans*-splicing creates new combinations of alleles in mRNA (Horiuchi *et al.*, 2003). Intragenic *trans*-splicing can generate new transcripts by exon reuse (Caudevilla *et al.*,

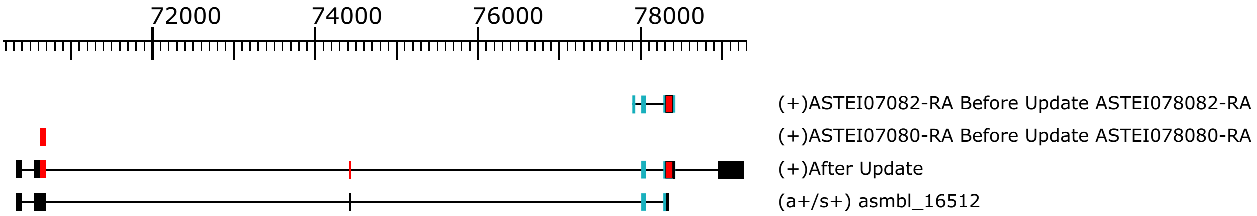
72000
74000
76000
78000
(+)ASTEI07082-RA Before Update ASTEI078082-RA
(+)ASTEI07080-RA Before Update ASTEI078080-RA
(+)After Update
(a+/s+) asmbl_16512

Figure 2. Updated annotation for the *doublesex* gene in *Anopheles stephensi*. The first row and second row below the genomic size ruler represent genes *ASTEI07082* and *ASTI07080*. The third row is an updated annotation from PASA, which merges the two genes. The fourth row is the evidence from isoform sequencing transcripts that supports the updated annotation. [Colour figure can be viewed at wileyonlinelibrary.com]

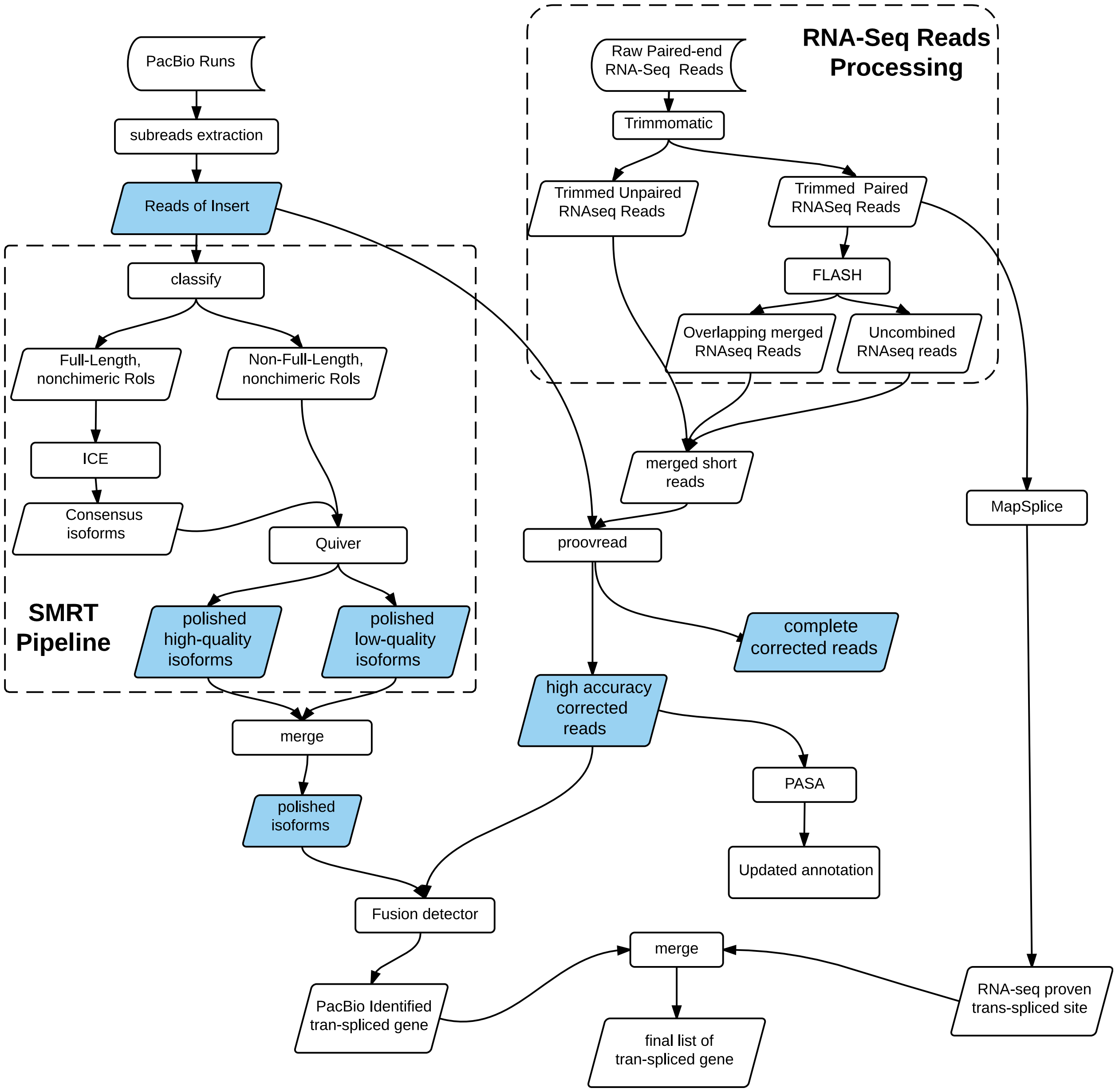
PacBio Runs
subreads extraction
Reads of Insert
classify
Full-Length, nonchimeric RoIs
Non-Full-Length, nonchimeric RoIs
ICE
Consensus isoforms
Quiver
SMRT Pipeline
polished high-quality isoforms
polished low-quality isoforms
merge
polished isoforms
Fusion detector
PacBio Identified tran-spliced gene
RNA-Seq Reads Processing
Raw Paired-end RNA-Seq Reads
Trimmomatic
Trimmed Unpaired RNAseq Reads
Trimmed Paired RNASeq Reads
FLASH
Overlapping merged RNAseq Reads
Uncombined RNAseq reads
merged short reads
proovread
complete corrected reads
high accuracy corrected reads
PASA
Updated annotation
MapSplice
merge
final list of tran-spliced gene
RNA-seq proven trans-spliced site

Figure 3. Data processing and analysis pipelines for both RNA-seq data and isoform sequencing (Iso-Seq) data. Processed Iso-Seq data that are highlighted in blue are compared in Table 2. Abbreviations: PacBio, Pacific Biosciences; SMRT, single molecule real-time; RoIs, Reads of Insert; RNA-Seq, RNA sequencing; ICE, isoform-level clustering. [Colour figure can be viewed at wileyonlinelibrary.com]

1998). Although intergenic *trans*-splicing in theory can produce new genes by joining exons from different genes, such a mechanism for novel gene formation does not appear to be favoured, as the intergenic *trans*-splicing events observed are largely involved in ancient gene rescue rather than new gene creation. Thus, intergenic *trans*-splicing events probably evolved from ancestral *cis*-splicing or allelic *trans*-splicing; the scenario where two random unrelated distant segments of the genome acquire the ability to be *trans*-spliced together should be rare if any. In addition, all of the proteins encoded by our *trans*-spliced events are highly conserved. It appears that strong purifying selection acts to maintain these *trans*-splicing events. If the protein is not essential, or other alternative strategies are adopted as in the case of Tm1 in *Drosophila*, intergenic *trans*-splicing may not occur or may be lost later during evolution. We can only speculate as to the possible reasons why intergenic *trans*-splicing appears to exist mainly as a gene rescue mechanism and not as a common facility to generate transcriptome/proteome diversity (Kong *et al*., 2015). First, unlike *cis*-splicing or the other two types of *trans*-splicing, it is difficult to ensure that two pre-mRNAs made from distant genomic loci share the same or overlapping temporal and spatial transcription (Gingeras, 2009). Second, upon transcription the physical distance or locations of the pre-mRNAs could hinder *trans*-splicing. Finally, it is hard for the two pre-mRNAs to coevolve when their DNA templates are shaped by potentially different evolutionary forces.

Experimental procedures

Library preparation and RNA sequencing

Fifteen 1–3-day-old *An. stephensi* (Indian type strain) adult male mosquitoes were homogenized in 900 μl RNA lysis buffer and total RNA was isolated using a *Quick-RNA* MiniPrep kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's protocol. Three hundred μl of the homogenate was used for total RNA isolation and total RNA was eluted with 30 μl H_2O. Two samples of 3.2 μl each (300 ng/μl) total RNA were subjected to PacBio sequencing at the Interdisciplinary Center for Biotechnology Research, University of Florida (Gainesville, FL, USA). The RNA was reverse transcribed using a SMRTer PCR cDNA synthesis kit (Clontech, Mountain View, California, USA) and amplified. Three sequencing libraries (1–2, 2–3, 3–6 kb) were prepared according to the PacBio Iso-Seq protocol. The sequencing was performed on the PacBio RS II using P4-C2 chemistry (Clontech, Mountain View, California, USA). Four SMRT cells were run from each of the three libraries. Illumina RNA-seq data used in this study are from Jiang *et al*. (2015).

SMRT pipeline analysis for Iso-Seq data

Analysis was performed using the PacBio SMRT-Analysis package v. 2.3 (http://www.pacb.com/devnet/). The three libraries were analysed separately. The Iso-Seq bioinformatics pipeline consists of two major modules: classify and cluster. Reads of insert were obtained by identifying the adapter separator and then merging subreads into consensus sequence reads. Reads of insert were then classified into full-length, non-artificial-concatemer reads and non-full-length reads. Full-length, non-artificial-concatemer reads from the same isoform were clustered using the isoform-level clustering algorithm and consensus isoforms were predicted. The consensus isoforms were then polished by Quiver utilizing the non-full-length reads. Default parameters were used when running Quiver, which means only consensus isoforms with more than 99% accuracy were binned into high-quality isoforms by Quiver. The low-quality isoforms binned by Quiver are generally the ones with low transcription level or low sequencing depth. Although less accurate, the low-quality isoforms also contain useful information and were used together with high-quality isoforms for our analysis.

Error correction of Iso-Seq data

The reads of insert of the three libraries were combined and subjected to error correction. RNA-seq data were processed as shown in Fig. 3 to achieve best performance for error correction. First, raw reads from RNA sequencing were trimmed with Trimmomatic with parameter '2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36'; (Bolger *et al*., 2014). The resulting trimmed paired reads were merged with FLASH with default parameters (Magoč & Salzberg, 2011). The merged reads along with reads that failed to merge and unpaired reads from Trimmomatic were combined into one fastq file. This fastq file was 26 Gb in size and used as short reads to correct the reads of insert with Proovread (Hackl *et al*., 2014). Proovread is a high-accuracy PacBio correction tool, which works via iterative alignment of short reads to produce consensus sequences. Proovread outputted high-accuracy PacBio reads with low-quality regions trimmed as well as complete corrected PacBio reads including poorly corrected regions. Only the high-accuracy trimmed Iso-Seq reads were used for our further analysis.

Fusion transcript detection with both Iso-Seq data and RNA-seq data

The *An. stephensi* Indian strain genome version 2 was downloaded from Vectorbase (Giraldo-Calderón *et al*., 2015). Based on our previous research (data not shown), we were able to include the majority of the *An. stephensi* Indian genome in five fasta sequences, with each of the sequences representing one chromosomal arm. The five fasta sequences were used as the genome sequence in the fusion transcript detection analysis. The high-accuracy trimmed Iso-Seq reads were aligned to the genome and then processed by the fusion_finder.py script of the PacBio pbtranscript-tofu package (https://github.com/PacificBiosciences/cDNA_primer.git). Proovread could potentially have removed *trans*-spliced transcripts when it removed PCR chimeric reads during correction. Therefore, the unpolished consensus isoforms were added to the above analysis to provide more sequencing information. MapSplice, a splice junction discovery software, was used to predict fusion genes in the RNA-seq data (Wang *et al*., 2010). With the '—fusion'; option, MapSplice performed canonical and semicanonical fusion junction detection after the RNA-seq reads were aligned to the genome. Only fusion junctions supported by both Iso-Seq and RNA-seq data were used. In total, six splice junctions were identified.

Genome annotation updates

Genome version 2 and annotations version 2.2 of the *An. stephensi* Indian strain were downloaded from Vectorbase (Giraldo-Calderón *et al*., 2015). pasa release r20140417 (http://pasapipeline.github.io/) was used to update the existing annotations using evidence generated from the high-accuracy trimmed Iso-Seq reads, and then to compare the updated annotation to existing gene structure annotations (Haas *et al*., 2003). As the high-accuracy trimmed Iso-Seq reads kept the transcribed orientation, option '–transcribed_is_aligned_orient'; was added when the PASA pipeline were launched.

Acknowledgements

This work was supported by National Institutes of Health grants AI121284, AI077680 and AI105575, and by the Virginia Agricultural Experimental Station.

References

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F. *et al*. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun* **7**: 11706. doi:10.1038/ncomms11706.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M. and Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**:623–630.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. doi:10.1093/bioinformatics/btu170.

Caudevilla, C., Serra, D., Miliar, A., Codony, C., Asins, G., Bach, M. *et al*. (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci USA* **95**: 12185–12190. doi:10.1073/pnas.95.21.12185.

Davidson, N.M., Majewski, I.J. and Oshlack, A. (2015) JAFFA: high sensitivity transcriptome-focused fusion gene detection. *Genome Med* **7**: 43. doi:10.1186/s13073-015-0167-x.

Dorn, R., Reuter, G. and Loewendorf, A. (2001) Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila*. *Proc Natl Acad Sci USA* **98**: 9724–9729. doi:10.1073/pnas.151268698.

Douris, V., Telford, M.J. and Averof, M. (2010) Evidence for multiple independent origins of trans-splicing in Metazoa. *Mol Biol Evol* **27**: 684–693. doi:10.1093/molbev/msp286.

Gao, J.L., Jie Fan, Y., Ye Wang, X., Zhang, Y., Pu, J., Li, L. *et al*. (2015) A conserved intronic U1 snRNP-binding sequence promotes trans-splicing in *Drosophila*. *Genes Dev* **29**: 760–771. doi:10.1101/gad.258863.115.

Gingeras, T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature* **461**: 206–211. doi:10.1038/nature08452.

Giraldo-Calderón, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Emrich, S., Collins, F. *et al*. (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* **43**: D707–D713. doi:10.1093/nar/gku1117.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Hannick, L.I. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666. doi:10.1093/nar/gkg770.

Hackl, T., Hedrich, R., Schultz, J. and Förster, F. (2014) Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics (Oxford, England)* **30**: 1–8. doi:10.1093/bioinformatics/btu392.

Hall, A.B., Papathanos, P.-A., Sharma, A., Cheng, C., Akbari, O.S., Assour, L. *et al*. (2016) Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *Proc Natl Acad Sci USA* **113**: 201525164. doi:10.1073/pnas.1525164113.

Herai, R.H. and Yamagishi, M.E.B. (2010) Detection of human interchromosomal trans-splicing in sequence databanks. *Brief Bioinform* **11**: 198–209. doi:10.1093/bib/bbp041.

Hiller, M., Zhang, Z., Backofen, R. and Stamm, S. (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genetics* **3**:2147–2155. doi:10.1371/journal.pgen.0030204.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform* **12**: 491. doi:10.1186/1471-2105-12-491.

Horiuchi, T., Giniger, E. and Aigaki, T. (2003) Alternative trans-splicing of constant and variable exons of a *Drosophila* axon guidance gene, Lola. *Genes Dev* **17**: 2496–2501. doi:10.1101/gad.1137303.

Jiang, X., Peery, A., Brantley Hall, A.B., Sharma, A., Chen, X.-G., Waterhouse, R.M. *et al*. (2014) Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. *Genome Biol* **15**: 459. doi:10.1186/s13059-014-0459-2.

Jiang, X., Biedler, J.K., Qi, Y., Hall, A.B. and Tu, Z. (2015) Complete dosage compensation in *Anopheles stephensi* and the evolution of sex-biased genes in mosquitoes. *Genome Biol Evol* **7**: 1914–1924. doi:10.1093/gbe/evv115.

Kong, Y., Zhou, H., Yu, Y., Chen, L., Hao, P. and Li, X. (2015) The evolutionary landscape of intergenic trans-splicing events in insects. *Nat Commun* **6**: 8734. doi:10.1038/ncomms9734.

Lasda, E.L. and Blumenthal, T. (2011) Trans-splicing. *Wiley Interdiscip Rev RNA* **2**: 417–434. doi:10.1002/wrna.71.

Liu, S., Tsai, W.-H., Ding, Y., Chen, R., Fang, Z., Huo, Z., Kim, S.H. *et al*. (2015) Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-Seq data. *Nucleic Acids Res* **44**: e47. doi:10.1093/nar/gkv1234.

Magoč, T. and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963. doi:10.1093/bioinformatics/btr507.

Robertson, H.M., Navik, J.A., Walden, K.K.O. and Honegger, H.W. (2007) The Bursicon gene in mosquitoes: an unusual example of mRNA trans-splicing. *Genetics* **176**: 1351–1353. doi:10.1534/genetics.107.070938.

Shao, W., Zhao, Q-y., Wang, X-y., Xu, X-y., Tang, Q., Li, M. *et al*. (2012) Alternative splicing and trans-splicing events revealed by analysis of the Bombyx *m*ori transcriptome. *RNA* **18**: 1395–1407. doi:10.1261/rna.029751.111.5.

Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013) A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009–1014. doi:10.1038/nbt.2705.

Suzuki, M.G., Ohbayashi, F., Mita, K. and Shimada, T. (2001) The mechanism of sex-specific splicing at the doublesex gene is different between *Drosophila melanogaster* and *Bombyx mori*. *Insect Biochem Mol Biol* **31**: 1201–1211. doi:10.1016/S0965-1748(01)00067-4.

Wally, V., Murauer, E.M. and Bauer, J.W. (2012) Spliceosome-mediated trans-splicing: the therapeutic cut and paste. *J Invest Dermatol* **132**: 1959–1966. doi:10.1038/jid.2012.101.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L. *et al*. (2010) MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. *Nucleic Acids Res* **38**: e178. doi:10.1093/nar/gkq622.

Warf, M.B. and Berglund, J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35**: 169–178. doi:10.1016/j.tibs.2009.10.004.

Weirather, J.L., Tootoonchi Afshar, P., Clark, T.A., Tseng, E., Powers, L.S., Underwood, J.G. *et al*. (2015) Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res* **43**: e116. doi:10.1093/nar/gkv562.