



Pergamon

Insect Biochemistry and Molecular Biology 29 (1999) 643–651

*Insect  
Biochemistry  
and  
Molecular  
Biology*

www.elsevier.com/locate/ibmb

# Vertebrate exon trapping methods: implications for transcript mapping with mosquito DNA

S.E. Brown <sup>a</sup>, S.H. Wood <sup>a, b</sup>, D.L. Knudson <sup>a,\*</sup>

<sup>a</sup> Department of Bioagricultural Sciences and Pest Management, College of Agricultural Sciences, Colorado State University, Fort Collins, CO 80523, USA

<sup>b</sup> Present address: Child Development Unit, B140, The Children's Hospital, 1056 E. Nineteenth Avenue, Denver, CO 80218, USA

Received 6 October 1998; received in revised form 9 April 1999; accepted 14 April 1999

## Abstract

Exon trapping methods have played an important role in the development of transcript maps. In one in vivo vertebrate method, exons in a genomic DNA clone are transcribed, and they are recovered without any a priori information on the nature of the expressed transcript. The only requirement is that the genomic DNA clone contains exons separated by intervening introns that are removed by splicing during mRNA transcription and that the splice donor and acceptor site sequences follow those used by vertebrates. It is not known whether invertebrate splice donor and acceptor sites from genes that contain short introns will be processed correctly using an in vivo vertebrate exon trapping method. In this report, an analysis of mosquito splice sites using software designed to identify exons in genomic DNA sequence suggested that the vertebrate exon trapping method could recognize mosquito introns and exons. When a mosquito genomic DNA clone containing the D7 gene was tested experimentally, this method failed to recognize and process small introns (<63 bp) faithfully. In spite of this failure, exons and exon fragments were recovered. The implications of these findings and their application to map-based positional cloning in mosquito genomics is discussed. © 1999 Elsevier Science Ltd. All rights reserved.

**Keywords:** Bioinformatics; D7 gene; Exon trapping; Intron; RNA splicing; Transcript mapping

## 1. Introduction

Filarial and malarial parasites are responsible for devastating infections of humans. Since these parasites are vectored in nature by mosquitoes, the elucidation of the molecular mechanism(s) that underlie parasite vector competence has been a goal that we have shared with others. Our research efforts in collaboration with others have focused on map-based positional cloning of genes responsible for parasite vector competence in *Aedes aegypti*. A fluorescence in situ hybridization (FISH) physical map of the genome has been constructed. FISH probes, consisting of recombinant cosmid and bacterial artificial chromosome (BAC) clones, have been mapped with an average spacing of ~1 Mb along the three metaphase chromosomes (Brown et al., 1995; Knudson et al., 1996; Ferguson et al., 1996; Brown and Knudson, 1997,

S.E. Brown et al, in preparation). The genetic linkage map of *Aedes aegypti* (Severson et al., 1993; Severson et al., 1995a) has been correlated with the FISH physical map (S.E. Brown, D.W. Severson, and D.L. Knudson, in preparation). In addition, regions of the genome that control parasite vector competence have been identified, and their contribution to the phenotype has been recognized as quantitative (Severson et al., 1994, 1994; Thathy et al., 1994; Beerntsen et al., 1994; Severson et al., 1995b). Contig construction across these quantitative trait loci has begun (S.E. Brown, D.W. Severson, and D.L. Knudson, unpublished), and as such, the efforts have progressed to the point where strategies for transcript mapping of mosquito genomic sequences need to be evaluated. Several strategies for isolating transcribed sequences from large genomic regions could be applicable to mosquito sequences. We chose first to evaluate the in vivo exon trapping method (also known as exon amplification). This method isolates internal and terminal exons from genomic clones, such as, cosmids and BACs (Auch and Reth, 1990; Duyk et al., 1990; Buckler

\* Corresponding author. Fax: +1-970-491-3862.  
E-mail address: dknudson@lamar.colostate.edu (D.L. Knudson)

et al., 1991; Church et al., 1994; Nehls et al., 1994; Datson et al., 1996; Krizman, 1996). This transcript mapping method has been applied widely in human genetic research, and it has been a primary tool in gene identification because several 100 kb of a recombinant DNA clone can be screened in a single transfection.

Can an in vivo vertebrate exon trapping method be used with invertebrate genes? It is not known whether invertebrate splice signal sequences will be processed by an in vivo vertebrate exon trapping method. *Drosophila* genes exhibit both long and very short introns, with more than fifty percent of the characterized introns being <80 bp (Mount et al., 1992). Long *Drosophila* introns (>100 bp) have been shown to be processed correctly by in vitro mammalian cell extracts (Reed and Maniatis, 1985; Siebel and Rio, 1990; Mount et al., 1992; Guo et al., 1993). Short *Drosophila* introns, however, are not processed by mammalian extracts in vitro (Guo et al., 1993; Talerico and Berget, 1994), suggesting that ~70 bp (Kennedy and Berget, 1997) is the minimal intron length that may be processed by vertebrate cells (Mount et al., 1992).

In this report, the genomic DNA sequence of the mosquito D7 gene which contains short introns was analyzed using software that is designed to identify exons in genomic DNA sequence, and the success of these software methods is reported. The in vivo vertebrate exon trapping method was also used to determine whether exons of the mosquito genomic clone containing the D7 gene would be processed faithfully in vivo. The application of in vivo vertebrate exon trapping methods to map-based positional cloning in mosquito genomics is discussed in light of the findings reported herein.

## 2. Methods

### 2.1. Cell line

The SV40 transformed African green monkey kidney cell line (COS-7) was supplied by the American Type Culture Collection as their designation, ATCC CRL-1651. The cell line was maintained in 150 cm<sup>2</sup> flasks and grown in Leibovitz's L-15 medium (GIBCO-BRL Life Technologies, Inc.) with 100 Units/ml of penicillin and 100 µg/ml streptomycin (GIBCO-BRL Life Technologies, Inc.) and 10% fetal bovine serum. The cell cultures were split 1:20 once a week.

### 2.2. Bacteria and DNA

Bacteria were grown and transformed following standard methods (Ausubel et al., 1987) or supplier's recommendations. Restriction endonuclease analysis followed standard protocols (Ausubel et al., 1987).

Plasmids were purified using QIAGEN-tip 20 columns following the manufacturer's protocol (Qiagen, Inc.).

The purified plasmids served as double-stranded templates in cycle sequencing reactions following standard protocols (Perkin-Elmer, Inc.). Sequence data for both strands was generated using an ABI Prism 377 DNA sequencer (Perkin-Elmer, Inc.). ABI trace files were converted and imported into ACEMBLY for sequence assembly (Thierry-Mieg, D. and J. Thierry-Mieg 1995. ACEMBLY. Documentation, code and data available by anonymous ftp from ftp.crbm.cnrs-mop.fr).

Dr. Anthony A. James (University of California, Irvine) provided purified DNA from a lambda clone containing the MQSD7AA genomic sequence (GenBank, M33156). A 3819 bp *EcoRI* fragment containing the complete MQSD7AA sequence was isolated from low-melting point agarose (Maruniak et al., 1984), and it was subcloned into pUC19. *E. coli* host cells DH5α were transformed and a recombinant clone p2411 containing MQSD7AA was identified by *EcoRI* restriction endonuclease digestion and separation of digest fragments by agarose gel electrophoresis.

### 2.3. Exon trapping

The exon trapping (also known as exon amplification) protocols and vectors were essentially those of Buckler and colleagues (Buckler et al., 1991) with later modifications (Church et al., 1994). A partial kit and detailed methods are available commercially (GIBCO-BRL Life Technologies, Inc., Instruction Manual — Exon Trapping System, Cat. No. 18449-017, also at URL <http://www2.lifetech.com/techonline.html>).

Plasmid p2411 was digested with *EcoRI* and the 3819 bp MQSD7AA *EcoRI* fragment containing the D7 gene was isolated using a low-melting point agarose gel separation. MQSD7AA was subcloned into the *EcoRI* digested pSPL3 "splicing" vector. pSPL3 contained an origin of replication, an ampicillin resistance gene, which was a selectable marker for growth in *E. coli*, and a SV40 promoter for replication and transcription in SV40-transformed vertebrate cells. In addition, pSPL3 contains HIV-1 *tat* gene, an intron, and a multiple cloning site that is flanked by HIV-1 *tat* splice donor and acceptor sites. In addition to the multiple cloning site, the pSPL3 vector contained a complete and half *BstXI* site so that products resulting from cryptic splicing events would be reduced when digested with *BstXI*. *E. coli* (MAX Efficiency™ HB101, GIBCO cat#18295-014) host cells were transformed, and a single colony was isolated that contained a pSPL3/MQSD7AA plasmid. The colony was amplified by growth in broth culture and the plasmid DNA was purified, and the DNA was electroporated into COS-7 cells using a ElectroCell Manipulator ECM600 (BTX, Inc.) following the manufacturer's guidelines. COS-7 was an SV40-derivative,

transformed African green monkey kidney cell line that permits a high level of replication and transcription to occur from the pSPL3 SV40 promoter. After electroporation, the cells were incubated for 18 to 24 h and the total RNA was purified. cDNA was synthesized from the total RNA isolated and its synthesis was catalyzed by SuperScript II RNase H-Reverse Transcriptase (GIBCO-BRL Life Technologies, Inc.). The resulting cDNA synthesis products were used as templates for a “primary PCR” amplification using primers complementary to pSPL3 and that flank the HIV-1 *tat* splice donor and acceptor sequences. The primary PCR amplification products were digested with *Bst*XI to reduce vector–vector splicing and cryptic splice event products. dUMP-tailed primers that annealed 3′ to the initial primers were used for a secondary PCR amplification. These amplification products resulted in CUA-tailed fragments, and these tailed fragments were cloned into a linear pAMP10 vector using uracil-DNA glycosylase (UDG), where UDG removed U in the CUA-tailed fragments, which disrupted base-pairing and exposed 3′ overhangs for annealing to the pAMP10 vector. The pAMP10 vector was a derivative of the pSPORT1 and contains a multiple cloning site, SP6 and T7 RNA polymerase promoters, *lac* promoter and repressor sequence for blue–white screening, the  $\beta$ -lactamase gene for ampicillin resistance, and the phase f1 intergenic region. Competent bacterial cells were transformed and recombinant colonies were isolated. These colonies were screened using colony PCR with the secondary amplification primers. The size of the colony PCR product was determined by agarose gel electrophoresis. If the PCR product was >177 bp, then mosquito DNA was likely “trapped” and these colonies were amplified by growth in broth culture. The pAMP10 plasmid containing trapped DNA was purified and they were sequenced in one pass. The sequence data was compared with the MQSD7AA sequence. If mosquito DNA from the D7 gene was present, then the same plasmid was sequenced from the opposite strand to yield sequence data from both strands.

## 2.4. Bioinformatics

A variety of exon-identification computational tools, such as GRAIL (and its derivatives, e.g., GRAIL2, XGRAIL) and GENEFINDER, were used (Table 1). These programs were used either locally as a UNIX version or via a WWW INTERNET interface. Detailed information on these programs and their use may be found at URL (<http://igs-server.cnrs-mrs.fr/banbury/>). Human or vertebrate features were used for all gene or exon predictions unless noted otherwise. The GENEFINDER package is integrated into the current mosquito databases, such as MsqDB or AaeDB (<http://klab.agsci.colostate.edu>) as an integral part of the database software engine (ACeDB, Durbin, R. and Thi-

erry–Meig, J.). GENEFINDER output of MQSD7AA is presented in the upper panel of Fig. 2. SEQUIN sequence preparation package was used for GenBank submissions (<http://www.ncbi.nlm.nih.gov>).

## 3. Results and discussion

When 19 mosquito genomic sequences containing introns were analyzed using GRAIL or GENEFINDER, most splice sites were accurately identified and rated by the software as a good to excellent splice site match (data not shown). Given the previous splicing studies with *Drosophila* introns in mammalian cell extracts in vitro (Reed and Maniatis, 1985; Siebel and Rio, 1990; Mount et al., 1992; Guo et al., 1993), it seemed likely that mosquito genes with long introns would be spliced properly using vertebrate in vivo exon trapping methods. If 78 mosquito genes that contain introns are examined for intron length, 57% of mosquito introns are less than 80 bp and 63% are less than 100 bp (data not shown). With the relative abundance of short mosquito introns being similar to *Drosophila* (Mount et al., 1992), would a mosquito gene that contained short introns be processed faithfully by mammalian in vivo exon trapping methods?

### 3.1. Bioinformatics of the D7 encoding genomic sequence

To test this question, the 3819 bp mosquito genomic clone MQSD7AA (GenBank M33156), which encodes the D7 gene that contains five exons and four short introns (<63 bp) was examined by 10 gene or exon prediction software programs. The D7 gene is well characterized because the D7 mRNA (MQSD7AB) had been isolated and also sequenced (GenBank M33157) confirming the spliced nature of the gene and identifying its splice signals (James et al., 1991). Clearly, this mosquito genomic sequence would provide a stringent test of the ability of the mammalian exon trapping methods to process short introns. When the results of these programs are taken together, the five D7 exons are identified (Table 1). All of the programs found at least two D7 exons, and the GENEID program found all five D7 exons. The GENEID program, however, also compares the query sequence with GenBank in the analysis, and thus, it is not surprising that GENEID identified the correct five D7 exons. Only the GENEID and MORGAN correctly identified the small 44 bp exon 1, exon 4 was missed by four of the programs, and three programs selected a possible exon upstream from the recognized D7 exons (Table 1).

Table 1

Exon prediction software analyses of the MQSD7AA genomic clone

| SOFTWARE   | Non-D7     | D7 Exon Coordinates <sup>a</sup> |                      |                      |                      |                      |
|------------|------------|----------------------------------|----------------------|----------------------|----------------------|----------------------|
|            | Exon       | Exon 1<br>2377..2421             | Exon 2<br>2481..2696 | Exon 3<br>2757..2975 | Exon 4<br>3038..3232 | Exon 5<br>3288..3578 |
| FGENEH     |            |                                  | 2481..2654           |                      |                      | +                    |
| GENEID     |            | +                                | +                    | +                    | +                    | +                    |
| GENIE      |            |                                  | 2490..2612           | +                    |                      |                      |
| GENSCAN    | 1311..1427 |                                  | +                    | +                    | +                    | +                    |
| GENVIEW    | 1311..1427 |                                  |                      | +                    |                      | 3288..3558           |
| GRAIL2     | 1311..1427 |                                  | 2481..2654           | +                    | +                    | 3288..3567           |
| MORGAN     |            | +                                | +                    | +                    |                      | +                    |
| MZEF       |            |                                  | 2481..2654           | +                    | +                    | +                    |
| PROCRUSTES |            |                                  | +                    | +                    | 3086..3197           | 3295..3575           |
| VEIL       |            |                                  | 2514..2654           | +                    | +                    | +                    |

<sup>a</sup> Exons indicated by a + are exact matches with D7 exons and those exons that are not exact matches have their coordinates indicated relative to the MQSD7AA sequence (start..stop). Please note the D7 gene actually ends at 3639, just beyond the polyA addition signal, but exon 5 as delineated here is used throughout. These programs were used either locally as a UNIX version or via a WWW INTERNET interface. Information and URL links to these programs may be found at <http://igs-server.cnrs-mrs.fr/banbury/>. Human or vertebrate features were used for gene or exon predictions and the program scores were (program, score;): fgeneh, none; GeneID, 6.91; Genie, 4.7; GenScan, 89.02 (exon scores totaled); GenView, none; Grail2, good to excellent; Morgan, 818.72/21.44 (Score/Normalized score); MZEF, none; Procrustes, 818; Veil, 1.0001–52800701. The GENEID program also compares the query sequence with GenBank in the analysis, and thus, it is not surprising that GENEID identified the correct five D7 exons.

### 3.2. Exon trapping of the D7 encoding genomic sequence

Briefly, the MQSD7AA *Eco*RI fragment representing the genomic sequence that contained the D7 gene was experimentally tested using an in vivo vertebrate exon trapping method. Since the secondary PCR amplification products would provide the first indication whether the vertebrate system could trap D7 exons, the secondary PCR products and appropriate controls were examined by agarose gel electrophoresis (Fig. 1a). In the second lane from the left, the pSPL3 control contained an expected, prominent 177 bp band that represented a portion of the pSPL3 vector containing the HIV splice donor and acceptor sequence without any trapped DNA. When the pSPL3 was digested with *Bst*XI as an internal control seen the third lane (labelled, pSPL3+*Bst*XI), the 177 bp band was diminished as expected. In the fourth lane (labelled, exon control+*Bst*XI), the positive control that contained an exon was digested with *Bst*XI and it resulted in the anticipated prominent ~244 bp band indicating the presence of a 67 bp exon (67=244–177). Visual inspection of the *Bst*XI digested MQSD7AA secondary PCR amplification products suggested that DNA had been trapped because at least three bands above 177 bp were seen (Fig. 1a, fifth lane from the left).

Since there was an indication that DNA had been trapped, these MQSD7AA secondary PCR products were cloned into the specialized, pAMP10 plasmid and *E. coli* host cells were transformed. Although pAMP10 transformants, which contained the MQSD7AA sequence, could have been identified by hybridization of

bacterial colony lifts using MQSD7AA as a probe, we chose to select and size colonies at random. Sixteen random clones are shown in Fig. 1b and six different size classes of trapped sequences were present. Four clones were selected for sequence determination and they represented the larger size classes (Fig. 1b). The resulting sequence data indicated that D7 sequences were present in all four exon trap plasmids, et2415, et2416, et2417, and et2430.

### 3.3. Analysis of trapped products

An alignment of the exon trap sequences with MQSD7AA genomic sequence revealed the trapped regions from the MQSD7AA genomic clone were derived from D7 exon 2, 3, and 5 with a 258 bp sequence upstream from the known D7 gene (Fig. 2, lower panel). D7 exon 1 should not be trapped because this exon trapping method will not trap the initial exon and its sequence was not seen. In theory, D7 exon 4 could have been recovered, but it was not seen. Whether sequence determination of additional clones would have yielded exon 4 sequence is not known. GENEFINDER and other software programs (Table 1) indicated that alternative vertebrate splice donor and acceptor sites were present in the MQSD7AA sequence (Fig. 2, upper panel). These exon trap sequences indicated that these alternative sites were functional in the COS-7 cells because 5' truncated forms of D7 exon 3 and 3' truncated forms of D7 exon 2 and 5 were observed in this alignment (Fig. 2). Only exon 3 was recovered intact in two of the four exon trap plasmids. GENEFINDER predicts the presence of

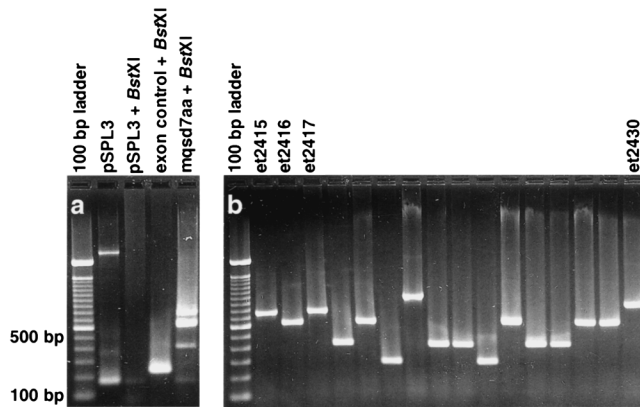


Fig. 1. Gel electrophoresis of secondary PCR products from the exon trap method. (a) a digital image of a 2% agarose gel electrophoresis of exon trap secondary PCR products is represented. A 100 bp DNA ladder molecular size standard (GIBCO-BRL cat #15628-019) was included in the left lane of the gel. The second through fourth lanes from the left were controls and the fifth lane contained the experimental DNA, MQSD7AA. The pSPL3 control contained an expected, prominent 177 bp band that represented a portion of the pSPL3 vector containing the HIV splice donor and acceptor sequence without trapped DNA (lane labeled, pSPL3). When the pSPL3 was digested with *Bst*XI, the 177 bp band was diminished as expected (lane labeled, pSPL3+BstXI). The exon containing positive control, which was digested with *Bst*XI displayed the anticipated prominent band at ~244 bp indicating the presence of the positive control 67 bp exon (67=244–177) (lane labeled, exon control+BstXI). Visual inspection of the *Bst*XI digested MQSD7AA secondary PCR products suggested that DNA had been trapped because at least three bands above 177 bp were seen (lane labeled, MQSD7AA+BstXI). (b) a digital image of a 2% agarose gel electrophoresis of 16 random exon trap clones is depicted and there are six different size classes discernible when compared to the 100 bp DNA ladder molecular size standard. Four larger clones were selected for sequence determination, and the sequences have been deposited in the GenBank database under accession numbers, AF027814–AF027817 for et2415, et2416, et2417, and et2430, respectively.

alternative splice sites in exon 2 and 5, which may help to explain their truncation (Fig. 2, top panel). Neither exon 3 nor exon 4 exhibited alternative splice sites in the GENEFINDER analysis (Fig. 2, top panel). Although D7 exon 1 is small, the previous D7 characterization data (James et al., 1991) indicated that it was the initial exon. This interpretation would suggest that the upstream 258 bp trapped sequence found at coordinates 1170..1427 resulted from the vertebrate exon trap method using alternative splice sites found upstream in the genomic sequence (Fig. 2, upper panel). Three software programs also selected similar coordinates at 1311..1427 as an exon upstream from the D7 coding sequence (Table 1). There is no experimental evidence from mosquitoes that indicates that this region is found in transcripts.

From the perspective of the intron, these sequence data indicate that none of the short introns were processed faithfully using vertebrate exon trapping methods. The introns were all effectively “lengthened,” resulting in the truncated exons as mentioned above. For example, the shortest intron length that was processed was 93 bp

for intron 2 in et2417 (Fig. 2). Given these findings, exon 4 was probably not recovered because it did not contain alternative vertebrate splice sites which prevented the effective lengthening of intron 3.

The D7 gene introns are A+T rich ranging from 60% to 73% with intron 2 being the lowest at 60% (Table 2). The *Drosophila* CTAAT branchpoint consensus sequence (Mount et al., 1992) was not seen, and nor was any other consensus branchpoint sequence discernable (Mount, 1996). In short introns of *Drosophila*, polypyrimidine tracts of at least 8 nucleotides in length are located between the 5′ splice site and branchpoint (Mount et al., 1992; Kennedy and Berget, 1997). Polypyrimidine tracts of 12 or more nucleotides in length are also located downstream of the branchpoint in short introns (Mount et al., 1992; Guo et al., 1993; Talerico and Berget, 1994). These two tract locations facilitate processing of short *Drosophila* introns in invertebrate cell extracts in vitro (Guo et al., 1993; Guo and Mount, 1995; Kennedy and Berget, 1997). The D7 gene introns do not exhibit polypyrimidine tracts and could be characterized as pyrimidine-poor, and those tracts that are present are predominately located downstream in the intron (Table 2). Rare, short vertebrate introns exhibit the trinucleotide GGG between the 5′ splice site and branchpoint and its presence is important for their efficient processing (McCullough and Berget, 1997). In the absence of alternative splice sites, the 5′ truncation of exon 3 may be due to the vertebrate exon trapping method elongating intron 2 and may similarly explain why exon 4 was apparently skipped.

These spliced intron sequences followed the GT..AG rule for group II introns in all cases (Table 3) (Mount, 1996). Similarly, the bases used in the splice donor and acceptor site from the exon trap plasmids were HR/GTMHGN and DHAG/S, which yields a simple consensus sequence of TG/GTAAGT and TAAG/G (Table 3). While this consensus was functional and consistent with the vertebrate consensus sequence, there were two minor exceptions (Table 3). For example, the –2 position of the donor site used T more frequently than A, and the –3 position in the acceptor site sequence used an A more often than C.

In an effort to determine a consensus sequence for mosquito splice sites, 1780 sequences in MsqDB (<http://klab.agsci.colostate.edu>) were scanned, and a base count was done around the splice sites of 156 introns found. The simple consensus sequences for the donor and acceptor splice site were AG/GTAAGT and GCAG/A respectively (Tables 4 and 5). Since the –2 position of the donor site was equally likely to have T as an A, and an A in the +1 position of the acceptor site was twice as likely as a G, a general consensus of the splice donor and acceptor sites would be WG/GTAAGT and NCAG/R. In addition, group II introns with GT..AG ends predominated in the 156 introns scanned. These

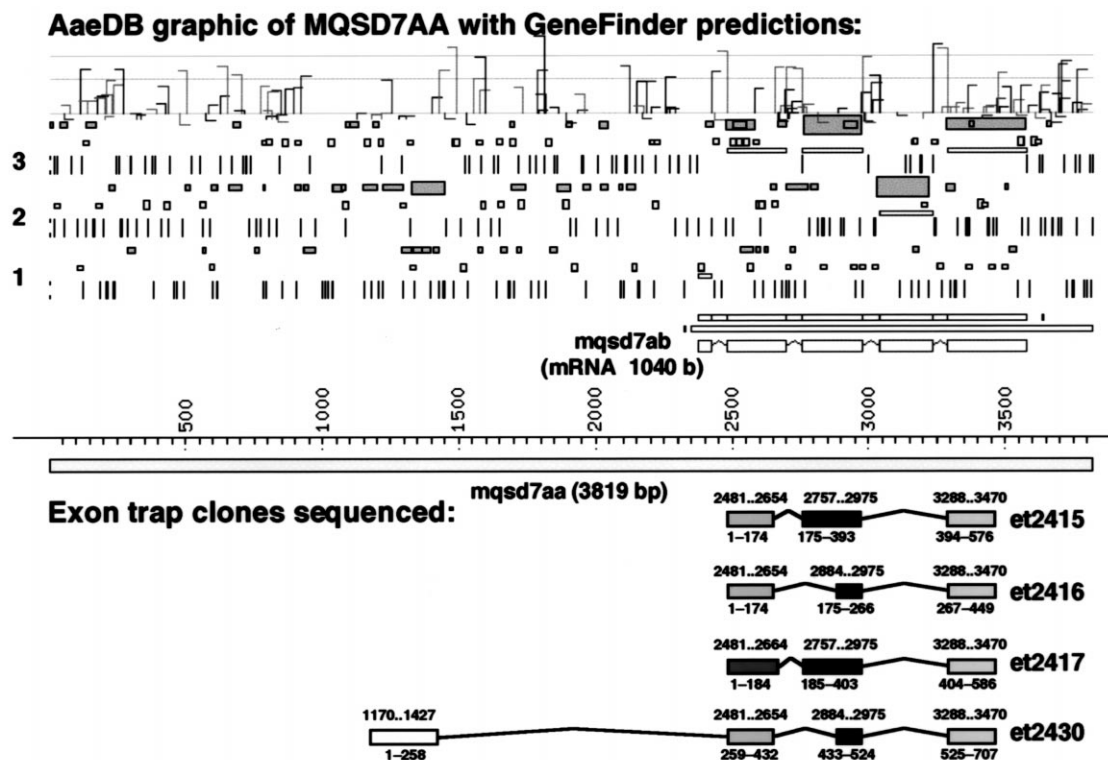


Fig. 2. Organization of the MQSD7AA genomic sequence, the D7 gene and exons trapped. A GENEFINDER graphic representing the organization of the MQSD7AA genomic sequence is depicted in the upper panel, which is labeled, AaeDB graphic of MQSD7AA with GENEFINDER predictions. The GENEFINDER graphic was produced using the AaeDB (*Aedes aegypti* DataBase, <http://klab.agsci.colostate.edu>), and it reveals the D7 gene organization relative to the MQSD7AA genomic clone. Elements of the graphic are from top to bottom: GENEFINDER 5' and 3' splice site predictions represented as left or right bars respectively on the height line which reflects the relative quality of the splice site prediction with higher being a better splice site. In the next portion of the upper panel, three open reading frames (ORF) of the sequence are shown and each frame is numbered on the left. The start codons are indicated as vertical bars, the stop codons are denoted as boxes immediately above, and potential ORFs are represented above the stop codons as boxes. Below the reading frames in the upper panel, three horizontal bars on the right depict annotated GenBank features (GenBank M33156). For example, the bottom series of boxes that is labeled on the left as MQSD7AB represents the mRNA with exon/intron boundaries indicated. The organization of the exon trap clones is represented in the lower panel, which is labeled, Exon trap clones sequenced. Based upon sequence alignment, the exon trap clones are depicted as elements aligned to the MQSD7AA genomic sequence. Numbers above the exon are in MQSD7AA coordinates and those below represent the exon trap coordinates. This alignment of the exon trap sequences with MQSD7AA genomic sequence revealed the trapped regions from the MQSD7AA genomic clone were derived from D7 exon 2, 3, and 5 with a 258 bp sequence upstream from the known D7 gene.

Table 2

Alignment of the intron sequences from the MQSD7AA genomic sequence<sup>a</sup>

|                  |                                                                             |
|------------------|-----------------------------------------------------------------------------|
| Intron 1 (59 bp) | GTAAGATGTTCTAGTGCATCCAACCTCTTC-ATTCTTA-ACAAATACTCCACCT-TTTAAG               |
| Intron 2 (60 bp) | .....T..G..GG..T..AT-...GGG..GGAA-..GGA..ATTTC..TCG..T..AT..G..GT..         |
| Intron 3 (62 bp) | .....TGT.....GT..T-..T..G..T..GAAACA...GTG..G..T..TG..C..TATT..AT..GC...C.. |
| Intron 4 (55 bp) | .....AAAA..AA..AAT-...ATG..AAACATC..AG-...TCG..C..T..T-...C..               |

<sup>a</sup> CLUSTALX (1.4b) multiple sequence alignment of MQSD7AA intron sequences where hyphen (-) represents a gap in the sequence and . represents an identical base when compared to intron 1.

consensus sequences are apparently so similar to the vertebrate consensus sequences that one would anticipate that most mosquito exons would be trapped using vertebrate exon trapping methods. A compilation of short introns (<87 bp) versus long introns (>117 bp) splice site sequences, which were taken from a subset of sequences where multiple exons were present, resulted in different splice site consensus sequences as compared

to the overall consensus sequences (Tables 4 and 5). The simple consensus for the donor and acceptor site for short introns, for example, was AG/ATAGGA and AGAR/A while the simple long intron consensus sequences were TG/AAAAATC and CGCG/V for the donor and acceptor sites respectively. These analyses suggest that caution is warranted in describing consensus sequences.

Table 3

A summary of the splice donor and acceptor sites used in exons trapped from the MQSD7AA genomic clone<sup>a</sup>

| Sequence                    | D7 Exon  | 5'-Splice Donor Site | 3'-Splice Acceptor Site |
|-----------------------------|----------|----------------------|-------------------------|
| Vertebrate Exon Trap Method |          | AG/GTRAGT            | NCAG/G                  |
| MQSD7AB                     | 1        | T./...A..A           | -                       |
|                             | 2        | .T/...A...           | TA.../.                 |
|                             | 3        | T./...A...           | GT.../.                 |
|                             | 4        | .T/...A..A           | T.../.                  |
|                             | 5        | -                    | T.../.                  |
| et2415                      | 2        | T./...A..A           | TA.../.                 |
|                             | 3        | T./...A...           | GT.../.                 |
|                             | 5        | .../...AC.G          | T.../.                  |
| et2416                      | 2        | T./...A..A           | TA.../.                 |
|                             | 3        | T./...A...           | AA.../C                 |
|                             | 5        | .../...AC.G          | T.../.                  |
| et2417                      | 2        | .../...CT..          | TA.../.                 |
|                             | 3        | T./...A...           | GT.../.                 |
|                             | 5        | .../...AC.G          | T.../.                  |
| et2430                      | upstream | CA/...A..C           | A.../C                  |
|                             | 2        | T./...A..A           | TA.../.                 |
|                             | 3        | T./...A...           | AA.../C                 |
|                             | 5        | .../...AC.G          | T.../.                  |

<sup>a</sup> N=any base, R=A or G, /represents the site, and . represents an identical base when compared to the vertebrate consensus.

Table 4

Mosquito 5' donor site simple consensus sequences overall and short versus long intron consensus sequences<sup>a</sup>

|                                                                             |           |           |           |           |                       |          |          |          |          |          |    |
|-----------------------------------------------------------------------------|-----------|-----------|-----------|-----------|-----------------------|----------|----------|----------|----------|----------|----|
| 5' donor sites                                                              |           |           |           |           |                       |          |          |          |          |          |    |
| Base                                                                        | -4        | -3        | -2        | -1        | 1                     | 2        | 3        | 4        | 5        | 6        | 7  |
| 5' donor site base count of 156 mosquito introns:                           |           |           |           |           |                       |          |          |          |          |          |    |
| A                                                                           | <b>25</b> | <b>64</b> | <b>54</b> | <b>28</b> | 39                    | 30       | 96       | 82       | 19       | 44       | 47 |
| C                                                                           | <b>36</b> | <b>33</b> | <b>20</b> | <b>10</b> | 12                    | 8        | 7        | 15       | 8        | 39       | 22 |
| G                                                                           | <b>53</b> | <b>31</b> | <b>25</b> | <b>92</b> | 89                    | 13       | 30       | 30       | 96       | 9        | 39 |
| T                                                                           | <b>34</b> | <b>20</b> | <b>49</b> | <b>18</b> | 8                     | 97       | 15       | 21       | 25       | 56       | 40 |
| Consensus:                                                                  |           |           | <b>A</b>  | <b>G</b>  | <i>G</i>              | <i>T</i> | <i>A</i> | <i>A</i> | <i>G</i> | <i>T</i> |    |
| 5' donor site base count of 22 sequences containing long mosquito introns:  |           |           |           |           |                       |          |          |          |          |          |    |
| A                                                                           | <b>1</b>  | <b>15</b> | <b>4</b>  | <b>0</b>  | 20                    | 19       | 16       | 15       | 0        | 1        | 15 |
| C                                                                           | <b>4</b>  | <b>5</b>  | <b>1</b>  | <b>0</b>  | 0                     | 1        | 4        | 2        | 0        | 20       | 1  |
| G                                                                           | <b>16</b> | <b>0</b>  | <b>1</b>  | <b>21</b> | 0                     | 0        | 1        | 0        | 5        | 0        | 0  |
| T                                                                           | <b>0</b>  | <b>1</b>  | <b>15</b> | <b>0</b>  | 1                     | 1        | 0        | 4        | 16       | 0        | 5  |
| Consensus:                                                                  |           |           | <b>T</b>  | <b>G</b>  | <i>A</i>              | <i>A</i> | <i>A</i> | <i>A</i> | <i>T</i> | <i>C</i> |    |
| 5' donor site base count of 42 sequences containing short mosquito introns: |           |           |           |           |                       |          |          |          |          |          |    |
| A                                                                           | <b>12</b> | <b>14</b> | <b>14</b> | <b>12</b> | 16                    | 8        | 16       | 8        | 11       | 16       | 11 |
| C                                                                           | <b>9</b>  | <b>7</b>  | <b>6</b>  | <b>4</b>  | 7                     | 4        | 2        | 8        | 7        | 13       | 10 |
| G                                                                           | <b>8</b>  | <b>10</b> | <b>10</b> | <b>17</b> | 12                    | 12       | 14       | 13       | 16       | 2        | 10 |
| T                                                                           | <b>11</b> | <b>9</b>  | <b>10</b> | <b>7</b>  | 5                     | 16       | 8        | 11       | 6        | 9        | 9  |
| Consensus:                                                                  |           |           | <b>A</b>  | <b>G</b>  | <i>A</i>              | <i>T</i> | <i>A</i> | <i>G</i> | <i>G</i> | <i>A</i> |    |
| Vertebrate 5' donor site consensus:                                         |           |           | <b>A</b>  | <b>G</b>  | —intron—><br><i>G</i> | <i>T</i> | <i>R</i> | <i>A</i> | <i>G</i> | <i>T</i> |    |

<sup>a</sup> 1780 sequences were scanned in MsqDB (<http://klab.agsci.colostate.edu>) resulting in frequency data from 156 introns tabulated. A subset of 47 sequences was selected where there were at least two exons in each entry and these were separated into short (<87 bp) and long introns (>118 bp) based upon an examination of the intron length. Intron regions are italicised and exon regions are bold.

### 3.4. Implications for mosquito genomics

In spite of the inability of the mammalian exon trapping method to process short introns properly in vivo, genes with longer introns will likely yield bona fide exons, though some may be truncated at their 3' end due to the use of alternative splice sites downstream. Genes

with short introns are likely to yield truncated exons, but the sequence data will likely be unique and repetitive element sequence data will be minimized. Since exons or exon fragments will be recovered, this method will be useful for transcript mapping with mosquito DNA. For example, the complete gene or transcript could be identified using other approaches when one has exon or

Table 5

3' acceptor site simple consensus sequences overall and short versus long intron consensus sequences<sup>a</sup>

| 3' acceptor sites                                                              |    |    |    |           |          |          |          |           |           |           |           |
|--------------------------------------------------------------------------------|----|----|----|-----------|----------|----------|----------|-----------|-----------|-----------|-----------|
| Base                                                                           | −7 | −6 | −5 | −4        | −3       | −2       | −1       | 1         | 2         | 3         | 4         |
| 3' acceptor site base count of 156 mosquito introns:                           |    |    |    |           |          |          |          |           |           |           |           |
| A                                                                              | 21 | 27 | 16 | 38        | 18       | 92       | 15       | <b>63</b> | <b>42</b> | <b>45</b> | <b>41</b> |
| C                                                                              | 38 | 39 | 38 | 32        | 67       | 19       | 23       | <b>25</b> | <b>29</b> | <b>27</b> | <b>30</b> |
| G                                                                              | 46 | 29 | 21 | 43        | 23       | 18       | 96       | <b>38</b> | <b>22</b> | <b>36</b> | <b>39</b> |
| T                                                                              | 37 | 46 | 66 | 28        | 33       | 12       | 7        | <b>15</b> | <b>48</b> | <b>33</b> | <b>31</b> |
| Consensus:                                                                     |    |    |    | <i>G</i>  | <i>C</i> | <i>A</i> | <i>G</i> | <b>A</b>  |           |           |           |
| 3' acceptor site base count of 22 sequences containing long mosquito introns:  |    |    |    |           |          |          |          |           |           |           |           |
| A                                                                              | 3  | 3  | 1  | 5         | 4        | 1        | 2        | <b>6</b>  | <b>1</b>  | <b>7</b>  | <b>2</b>  |
| C                                                                              | 5  | 10 | 8  | 11        | 5        | 8        | 8        | <b>6</b>  | <b>12</b> | <b>5</b>  | <b>5</b>  |
| G                                                                              | 6  | 6  | 7  | 0         | 9        | 4        | 9        | <b>6</b>  | <b>3</b>  | <b>7</b>  | <b>8</b>  |
| T                                                                              | 5  | 0  | 3  | 3         | 1        | 6        | 0        | <b>1</b>  | <b>3</b>  | <b>0</b>  | <b>4</b>  |
| Consensus:                                                                     |    |    |    | <i>C</i>  | <i>G</i> | <i>C</i> | <i>G</i> | <b>V</b>  |           |           |           |
| 3' acceptor site base count of 42 sequences containing short mosquito introns: |    |    |    |           |          |          |          |           |           |           |           |
| A                                                                              | 9  | 7  | 6  | 12        | 9        | 13       | 11       | <b>13</b> | <b>8</b>  | <b>11</b> | <b>9</b>  |
| C                                                                              | 11 | 14 | 13 | 9         | 10       | 9        | 9        | <b>6</b>  | <b>7</b>  | <b>9</b>  | <b>14</b> |
| G                                                                              | 14 | 9  | 9  | 11        | 13       | 10       | 11       | <b>11</b> | <b>9</b>  | <b>8</b>  | <b>7</b>  |
| T                                                                              | 5  | 8  | 10 | 6         | 6        | 6        | 7        | <b>8</b>  | <b>14</b> | <b>10</b> | <b>8</b>  |
| Consensus:                                                                     |    |    |    | <i>A</i>  | <i>G</i> | <i>A</i> | <i>R</i> | <b>A</b>  |           |           |           |
| Vertebrate 3' acceptor site consensus:                                         |    |    |    | <—intron— |          |          |          |           |           |           |           |
| Consensus:                                                                     |    |    |    | <i>N</i>  | <i>C</i> | <i>A</i> | <i>G</i> | <b>G</b>  |           |           |           |

<sup>a</sup> 1780 sequences were scanned in MsqDB (<http://klab.agsci.colostate.edu>) resulting in frequency data from 156 introns tabulated. A subset of 47 sequences was selected where there were at least two exons in each entry and these were separated into short (<87 bp) and long introns (>118 bp) based upon an examination of the intron length. Intron regions are italicised and exon regions are bold.

exon fragment sequence information in hand. This exon trapping method, however, would be most useful in position-based transcript mapping. In addition, the method will also be extremely helpful in contig building where sequence tagged connector (STC) strategies are used. STC-based contig building relies upon making minimal overlaps based on sequence tagged sites (STS) derived from end sequence data on large insert clones such as cosmid and BAC clones. The end sequence data will be either unique STS or contain repetitive elements. If repetitive elements are present, then exon trapping of the genomic clone would help to identify other STSs located internal to the ends. From these data, new primers or probes could be designed that could be used to rescreen recombinant libraries.

In summary, the following conclusions may be drawn of these studies. Short mosquito introns are not processed faithfully using mammalian *in vivo* exon trapping methods. The mammalian mRNA processing machinery can use certain mosquito splice sites, but it will also use alternative splice sites in exons that are preceded by short introns. Mosquito exons and exon fragments may be recovered, but those exons that do not have alternative vertebrate splice sites and are preceded by short introns may not be represented. Mosquito genes without introns will be missed by these methods. Bioinformatic analyses of mosquito genomic sequence have sufficient predictive value in transcript identification that direct sequencing of a region of interest and subsequent bioinformatic analysis may be an efficient way to identify genes.

## Acknowledgements

This work was supported by the National Institutes of Health, National Institute of Allergy and Infectious Disease grant AI34337, in part by the MacArthur Foundation through their Research Network on the Biology of Parasite Vectors program, and in part by the Colorado Agricultural Experiment Station. Dr. Anthony A. James is acknowledged for kindly providing purified DNA from a lambda clone containing the MQSD7AA genomic sequence. Drs. A.A. James and David W. Severson are also thanked for their comments on an early draft of the manuscript. Dr. S.M. Mount is acknowledged for his helpful discussions on the manuscript. The laboratory support services of Peter Carlson and Shaun Deeney are also acknowledged. The ABI Prism 377 sequencing was performed at the Colorado State University Macromolecular Resource Facility.

## References

- Auch, D., Reth, M., 1990. Exon trap cloning: using PCR to rapidly detect and clone exons from genomic DNA fragments. *Nucleic Acids Res.* 18, 6743–6744.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K. (Eds.), 1987. *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley-Interscience, New York.
- Beerssen, B.T., Severson, D.W., Christensen, B.M., 1994. *Aedes aegypti*: characterization of a hemolymph polypeptide expressed during melanotic encapsulation of filarial worms. *Exp. Parasit.* 79, 312–321.

- Brown, S.E., Knudson, D.L., 1997. FISH landmarks for *Aedes aegypti* chromosomes. *Insect Mol. Biol.* 6, 197–202.
- Brown, S.E., Menninger, J., Difillipantonio, M., Beaty, B.J., Ward, D.C., Knudson, D.L., 1995. Towards a physical map of *Aedes aegypti*. *Insect Mol. Biol.* 4, 161–167.
- Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A., Housman, D.E., 1991. Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* 88, 4005–4009.
- Church, D.M., Stotler, C.J., Rutter, J.L., Murrell, J.R., Trofatter, J.A., Buckler, A.J., 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat. Genet.* 6, 98–105.
- Datson, N.A., Van De Vosse, E., Dauwerse, H.G., Bout, M., Van Ommen, G.J., Den Dunnen, J.T., 1996. Scanning for genes in large genomic regions: cosmid-based exon trapping of multiple exons in a single product. *Nucleic Acids Res.* 24, 1105–1111.
- Duyk, G.M., Kim, S., Myers, R.M., Cox, D.R., 1990. Exon trapping: A genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc. Natl. Acad. Sci. USA* 87, 8995–8999.
- Ferguson, M.L., Brown, S.E., Knudson, D.L., 1996. FISH digital imaging microscopy in mosquito genomics. *Parasitology Today* 12, 91–96.
- Guo, M., Lo, P.C., Mount, S.M., 1993. Species-specific signals for the splicing of a short *Drosophila* intron in vitro. *Mol. Cell. Biol.* 13, 1104–1118.
- Guo, M., Mount, S.M., 1995. Localization of sequences required for size-specific splicing of a small *Drosophila* intron in vitro. *J. Mol. Biol.* 253, 426–437.
- James, A.A., Blackmer, K., Marinotti, O., Ghosn, C.R., Racioppi, J.V., 1991. Isolation and characterization of the gene expressing the major salivary gland protein of the female mosquito, *Aedes aegypti*. *Molecular and Biochemical Parasitology* 44, 245–254.
- Kennedy, C.F., Berget, S.M., 1997. Pyrimidine tracts between the 5' splice site and branch point facilitate splicing and recognition of a small *Drosophila* intron. *Mol. Cell. Biol.* 17, 2774–2780.
- Knudson, D.L., Zheng, L., Gordon, S.W., Brown, S.E., Kafatos, F.C., 1996. Genome organization of vectors. In: Beaty, B.J., Marquardt, W.C. (Eds.), *The Biology of Disease Vectors*. University Press of Colorado, Niwot, pp. 175–214.
- Krizman, D.B., 1996. Gene identification by 3' terminal exon trapping. *Genet. Eng. (NY)* 18, 49–56.
- Maruniak, J.E., Brown, S.E., Knudson, D.L., 1984. Physical maps of SfMNPV baculovirus DNA and its genomic variants. *Virology* 136, 221–234.
- McCullough, A.J., Berget, S.M., 1997. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* 17, 4562–4571.
- Mount, S.M., 1996. AT-AC introns: an ATtACk on dogma [comment]. *Science* 271, 1690–1692.
- Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., Fields, C., 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20, 4255–4262.
- Nehls, M., Pfeifer, D., Boehm, T., 1994. Exon amplification from complete libraries of genomic DNA using a novel phage vector with automatic plasmid excision facility: application to the mouse neurofibromatosis-1 locus. *Oncogene* 9, 2169–2175.
- Reed, R., Maniatis, T., 1985. Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell* 41, 95–105.
- Severson, D.W., 1994. Applications of molecular marker analysis to mosquito vector competence. *Parasitology Today* 10, 336–340.
- Severson, D.W., Mori, A., Zhang, Y., Christensen, B.M., 1993. Linkage map for *Aedes aegypti* using restriction fragment length polymorphisms. *J. Heredity* 84, 241–247.
- Severson, D.W., Mori, A., Zhang, Y., Christensen, B.M., 1994. Chromosomal mapping of two loci affecting filarial worm susceptibility in *Aedes aegypti*. *Insect Mol. Biol.* 3, 67–72.
- Severson, D.W., Mori, A., Kassner, V.A., Christensen, B.M., 1995a. Comparative linkage maps for the mosquitoes, *Aedes albopictus* and *Ae. aegypti* based upon common RFLP loci. *Insect Mol. Biol.* 4, 41–45.
- Severson, D.W., Thathy, V., Mori, A., Zhang, Y., Christensen, B.M., 1995b. Restriction fragment length polymorphism mapping of quantitative trait loci for malaria parasite susceptibility in the mosquito *Aedes aegypti*. *Genetics* 139, 1711–1717.
- Siebel, C.W., Rio, D.C., 1990. Regulated splicing of the *Drosophila* P transposable element third intron in vitro: somatic repression. *Science* 248, 1200–1208.
- Talerico, M., Berget, S.M., 1994. Intron definition in splicing of small *Drosophila* introns. *Mol. Cell. Biol.* 14, 3434–3445.
- Thathy, V., Severson, D.W., Christensen, B.M., 1994. Reinterpretation of the genetics of susceptibility of *Aedes aegypti* to *Plasmodium gallinaceum*. *J. Parasitol.* 80, 705–712.