Mitochondrial Genome Sequence and Expression Profiling for the Legume Pod Borer *Maruca vitrata* (Lepidoptera: Crambidae)

Venu M. Margam¹, Brad S. Coates², Richard L. Hellmich², Tolulope Agunbiade³, Manfredo J. Seufferheld⁴, Weilin Sun³, Malick N. Ba⁵, Antoine Sanon⁵, Clementine L. Binso-Dabire⁶, Ibrahim Baoua⁶, Mohammad F. Ishiyaku⁷, Fernando G. Covas⁸, Ramasamy Srinivasan⁹, Joel Armstrong¹⁰, Larry L. Murdock¹, Barry R. Pittendrigh³*⁸

¹ Department of Entomology, Purdue University, West Lafayette, Indiana, United States of America, ² United States Department of Agriculture – Agricultural Research Service, Corn Insect and Crop Genetics Research Unit, Genetics Laboratory, Iowa State University, Ames, Iowa, United States of America, ³ Department of Entomology, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, ⁴ Department of Crop Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America, ⁵ Station de Kamboinsé, Institut de l’Environnement et de Recherches Agricoles (INERA), Ouagadougou, Burkina Faso, ⁶ Institut National de Recherche Agronomique du Niger, Maradi, Niger, ⁷ Department of Plant Science, Institute for Agricultural Research, Ahmadu Bello University, Zaria, Nigeria, ⁸ University of Puerto Rico, Mayaguez, Puerto Rico, United States of America, ⁹ AVRDC-The World Vegetable Center, Tainan, Taiwan, ¹⁰ Entomology, The Commonwealth of Scientific and Industrial Research Organization, Black Mountain, Australian Capital Territory, Australia

**Abstract**

We report the assembly of the 14,054 bp near complete sequencing of the mitochondrial genome of the legume pod borer (*Maruca vitrata* (Lepidoptera: Crambidae), which we subsequently used to estimate divergence and relationships within the lepidopteran lineage. The arrangement and orientation of the 13 protein-coding, 2 rRNA, and 19 tRNA genes sequenced was typical of insect mitochondrial DNA sequences described to date. The sequence contained a high A+T content of 80.1% and a bias for the use of codons with A or T nucleotides in the 3rd position. Transcript mapping with midgut and salivary gland ESTs for mitochondrial genome annotation showed that translation from protein-coding genes initiates and terminates at standard mitochondrial codons, except for the *cox1* gene, which may start from an arginine CGA codon. The genomic copy of *cox1* terminates at a T nucleotide, and a proposed polyadenylation mechanism for completion of the TAA stop codon was confirmed by comparisons to EST data. EST contig data further showed that mature *M. vitrata* mitochondrial transcripts are monocistronic, except for bicistronic transcripts for overlapping genes *nd4*/*nd4L* and *nd6/*cytb*, and a tricistronic transcript for *atp8/*atp6/*cox1*. This processing of polycistronic mitochondrial transcripts adheres to the tRNA punctuated cleavage mechanism, whereby mature transcripts are cleaved only at intervening tRNA gene sequences. In contrast, the tricistronic *atp8/*atp6/*cox1* in *Drosophila* is present as separate *atp8*/*atp6* and *cox1* transcripts despite the lack of an intervening tRNA. Our results indicate that mitochondrial processing mechanisms vary between arthropod species, and that it is crucial to use transcriptional information to obtain full annotation of mitochondrial genomes.

**Citation:** Margam VM, Coates BS, Hellmich RL, Agunbiade T, Seufferheld MJ, et al. (2011) Mitochondrial Genome Sequence and Expression Profiling for the Legume Pod Borer *Maruca vitrata* (Lepidoptera: Crambidae). PLoS ONE 6(2): e16444. doi:10.1371/journal.pone.0016444

**Editor:** Guy Smagghe, Ghent University, Belgium

**Received** October 18, 2010; **Accepted** December 20, 2010; **Published** February 2, 2011

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

**Funding:** This research has been made possible through support provided to the Dry Grains Pulses Collaborative Research Support Program (CRSP) by the Bureau for Economic Growth, Agriculture, and Trade, U.S. Agency for International Development, under the terms of Grant No. EDH-A-00-07-00005. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Agency for International Development or the U.S. government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: pittendr@illinois.edu

**Introduction**

The mitochondrial genome encodes genes involved in oxidative phosphorylation and a unique translation system of 2 rRNAs and 22 tRNAs used for synthesis of the 13 inclusive protein coding genes (PCGs). Within animal species the mitochondrial genome shows a lack of introns and little intergenic space, the exception being the AT nucleotide rich displacement loop (D-loop) which encodes the origin of replication and promoters for the translation of both the heavy (H) and the light (L) strands (heavy strand promoter, HSP; light strand promoter, LSP) [1]. Thus, both H and L strands of the mitochondrial genome are transcriptionally active, and produce polycistronic RNA transcripts (cistrons) [2,3,4]. The tRNA regions form characteristic cloverleaf secondary structure within the initial polycistronic transcripts, and they are recognized by cleavage mechanisms that result in processed transcripts that encode one or more PCGs [5]. This mode of gene expression is reminiscent of the bacterial cistronic mode of gene expression from which eukaryotic mitochondria are believed to have been derived. As a consequence of gene order in the mitochondrial genome, the tRNA-punctuated mode of transcript processing can result in mature transcripts that encode more than one PCG [3], where bicistronic transcripts are predicted for mitochondrial PCG sequences (cds) that overlap within the genomic sequence, which includes *atp8*/*atp6* and *nd4L*/*nd4L* within insect mitochondria.
Mitochondrial genomes have been sequenced for insect species within the order Lepidoptera: the first ones sequenced were from the silk moth *Bombyx mori* [6] and the corn borer *Ostrinia* spp. [7]. Lepidopteran mitochondrial genomes have a relatively conserved gene order and orientation [8], which is shared with that of *Drosophila* species [9]. Despite the generation of full mitochondrial genome sequences, the corresponding annotation of gene coding regions in Lepidoptera has largely relied upon comparison to *Drosophila* gene models. There are discrepancies in mitochondrial gene boundaries in Lepidoptera, among them (1) the ambiguity in the translation start site of cytochrome c oxidase subunit I (coxI) at either a proposed arginine codon (CCR) or a TTAG site [6], and (2) the assumed polyadenylation following a T nucleotide at the terminus of coxII that would complete a truncated stop codon [9]. Expressed sequence tags (ESTs) are a source of gene expression information, and are typically inclusive of mitochondrial-derived transcripts. The use of EST data to assemble mitochondrial-derived transcripts has proven valuable in characterization of gene boundaries, polycistronic transcripts, and differential transcript processing among tissues, as well as for the quantification of mitochondrial transcript stability [10,11]. Despite the utility of ESTs in the annotation of mitochondrial genomes, these data are rarely incorporated into annotation efforts.

In this paper, we describe the nearly complete mitochondrial DNA sequence for the legume pod borer (LPB), *Maruca vitrata* Fabriscius (Lepidoptera: Pyraloidea: Grambidae). This insect species is a crop pest found throughout tropical and subtropical regions of the world. Larval stages of LPB feed upon leaves, flowers and pods of leguminous plants [12,13,14,15] and cause significant yield loss to legume crops cultivated in Southeast Asia [16,17,18], South Asia [19,20,21], and Central America and South Americas [22,23]. Its most significant impact is in sub-Saharan Africa [24,25,26,27], where between 20-80% yield reductions are incurred [25]. Accordingly, *M. vitrata* has been identified as a major emerging threat to legume production in developing and under-developed nations. The mitochondrial genome information provided here complements the mitochondrial coxDNA markers developed by Margam et al. [2010] [28], by providing additional sequence data to assist in defining population structure, and gene flow, which will hopefully lead to sustainable and economically viable methods of pest management. With the exception of *Drosophila melanogaster*, this study provides the first instance where (i) mitochondrial genome annotations have been validated by transcribed sequence data, and (ii) predictions of processed mitochondrial transcripts (citron) have been used for both structural and functional annotation of genes. This added value information suggests that available EST information is a valuable resource for mitochondrial genome annotations.

**Materials and Methods**

**Sequence and annotation of the *M. vitrata* mitochondrial genome**

Genomic DNA was isolated from a *M. vitrata* specimen (from Burkina Faso; BUR38) [28] using a DNeasy animal tissue kit following the manufacturer's instructions (Qiagen, Valencia, CA). A total of 100 ng genomic DNA was used as the template in 50 l PCR volumes that also contained 10 pmol primers, 5 l 10× PCR buffer, 0.4 l Taq polymerase (New England Biolabs, Ipswich, MA), and 1.2 l 10mM dNTP. Polymerase chain reaction (PCR) was carried out in an Eppendorf Mastercycler thermocycler (Eppendorf, Hamburg, Germany) using the thermal cycling profile of 95°C denaturation for 2 min, followed by 33 cycles of 95°C for 30 s, 52°C for 45 s, 72°C for 1 min, as well as a final cycle of 72°C for 8 min. The PCR products were cleaned to remove any residual primers and nucleotides using Qiaquick PCR purification kits (Qiagen, Valencia, CA) following the manufacturer’s protocols.

The PCR products were cycle sequenced using 1 l of purified template, 1 l of 10× BigDye® terminator cycle sequencing reaction mix (ABI), and 1 pmol forward or reverse primer (Table 1), and 6 l of ddH2O using the conditions: 95°C for 2 min, followed by 95 cycles of 95°C for 10 s, 50°C for 5 s, and 60°C for 4 min. Cycle-sequencing products were precipitated using an ethanol-sodium acetate procedure. Precipitates were dissolved in 35 l of ddH2O. From this solution, 15 l was separated on an ABI 3500 Sequencer (Applied Biosystems, Foster City, CA). Sequences from PCR fragments obtained from BUR38 and PR08 were assembled independently using the NextGENe software (Softgenetics®, State College, PA), and the assemblies were exported in FASTA format.

The FASTA-formatted BUR38 mitochondrial sequence was imported into a local database using BioEdit [29], and was queried with *O. nubilalis* protein and tRNA sequences from GenBank ID: AF442957.1. Protein cds were defined based on invertebrate mitochondrial translation rule using the Virtual Ribosome (http://www.cbs.dtu.dk/services/VirtualRibosome/) [30]. The tRNA gene boundaries and folded structures were defined using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) [31] with the following parameters: (i) search mode – organellar; (ii) search type – cove only; (iii) cove cut off score <15; (iv) translation code used – invertebrate mitochondrial; (v) maximum intron length = 40; and (vi) pseudogene checking – disabled. The mitochondrial sequences were imported into the Artemis visualization and annotation tool [32], positional information of sequence features was added, and the annotated data exported in GenBank.gb] format.

**Phylogenetics**

The FASTA-formatted mitochondrial genome assembly for *M. vitrata* was used in a query in the National Center for Biotechnology Information (NCB) non-redundant nucleotide database (nr) with the blastn search algorithm [33]. The resulting hits to full mitochondrial genome sequences for species of Lepidoptera were downloaded in FASTA format. The FASTA formatted mitochondrial genome sequences were imported into the MEGA 4.0 software package sequence alignment application [34] and a multiple sequence alignment was performed with the ClustalW algorithm using default parameters (gap opening penalty 15, gap extension penalty 6.66, weight matrix IUB, and transition weight of 0.5). The nucleotide diversity (p), base composition, transition to transversion bias (R; [35]), disparity in substitution among nucleotide sites (Disparity Index), and test substitution homogeneity [36] of the aligned sequence was analyzed by MEGA 4.0 [34].

Maximum likelihood analysis of derived amino acid sequences, to infer the phylogenetic relationships among mitochondrial genome sequences, was performed using the PHYLIP software package [37]. Concatenated protein sequences from each species of Lepidoptera that has a complete mitochondrial genome within GenBank (current December 7, 2010) were imported into the MEGA 4.0 sequence alignment suite, aligned using the ClustalW algorithm (default parameters using the PAM protein weight matrix), and all gaps deleted manually. One thousand bootstrap pseudo-datasets were constructed using the program seqboot, which was used as input into ProML that used a Hiddlen Markov Model method to infer evolutionary rates among residue positions [38] with the Jones, Talyor, and Thornton [39] probability matrix. An unrooted majority rule consensus tree was generated using Consense, and viewed using TreeView 1.6.6 [40].

---

**Figure 1**: Unrooted majority rule consensus tree (1000 bootstrap replicates) for mitochondrial genomes from selected species of Lepidoptera (O. nubilalis) and other insects (M. domestica, G. mellonella, D. melanogaster, T. castaneum, M. domestica). The tree topology was inferred using the Maximum Likelihood method [34] with the ClustalW algorithm using the default parameters (gap opening penalty 15, gap extension penalty 6.66, weight matrix IUB, and transition weight of 0.5). The nucleotide diversity (p), base composition, transition to transversion bias (R; [35]), disparity in substitution among nucleotide sites (Disparity Index), and test substitution homogeneity[36] of the aligned sequence was analyzed by MEGA 4.0 [34].
454 EST sequencing and mitochondrial transcript mapping

A total of 60 LPB larvae comprising of 3rd, 4th and 5th instars (20 each from Taiwan, Puerto Rico and Australia) were dissected to obtain the midgut and the salivary-gland tissues. Total RNA was extracted from these tissues using a TRizol® Reagent protocol (Invitrogen, Carlsbad, CA) according to the manufacturer’s instructions. RNA was quantified on a NanoDrop 2000 (Thermo Scientific, Wilmington, DE). Carlsbad, CA) according to the manufacturer’s instructions. RNA was from these tissues using a TRIzol from Taiwan, Puerto Rico and Australia) were dissected to obtain the midgut and the salivary-gland tissues. Total RNA was extracted using a TRIzol.

Table 1. Oligonucleotide primers used for amplification of overlapping fragments used to reconstruct the M. vitrata mitochondrial genome sequence (Fig. 2).

| Primer pair | Forward | Reverse |
|-------------|---------|---------|
| ND2        | TCTGATTTGGWGTGTYGAAATGGG | GTWCCAAATCTTTATGTATTGTGGAGG |
| COI        | TGGCTTTAGTACGATGAAATCCAA | CTAACTATTTATTTCTGAGG |
| COI3p      | TTAATCGATTCTATTTTAGGAGCTAAA | TTGCACCAAAAATTTCCAGACA |
| ND3a       | TCCCTTAATTTCTTTTCTGGTTATGAGG | GGAATTTAAACCATTTAATTTAAGTTAGC |
| ND3b       | TTTTTTCTTTTAGGATTTTACAGG | CCTTACTTTATATTATTTATTTACCTT |
| ND5        | GCTAACTGGGAACTCCATCCTCTT | TTATTATTAGTGGAGTTTATTAGTTT |
| Cytb-a     | CCAATTGGAAAAAACAACCCTTA | TCGGAGGACTTTACTCCGG |
| Cytb-b     | TCTTTTGGAAATTCAATGAAATTTAG | TCGGAGGACTTTACTCCGG |
| ND1        | TCATATGNAGACGGAAGTTCC | CCTGTCACTTAGTTATTTTTATCTT |
| gap1       | AGGGGGTTTATCCTCCTATTA | AATATAGCGTCAGGTTACAAATA |
| gap2       | TTACACCTGCAAGAAACTCA | CGTAAATGAAGTTAAGCAATAAAA |
| gap3       | GCAGAAGCTGCTTCTGAACCAT | AGATGGGGCTAACAAATTGAAAA |
| gap4       | CAGCCCAATATAATGAGGGCAAC | TCTGCAAAATATCTCGCATCA |
| gap5       | ATTGAAAGTCTCGCTGTAGAT | CGACGAAGGATAGTTGGTC |
| gap6       | GACCCAAATATTCTCCTGCGAA | GAGAAGTTTATACGGGTTAGGAA |
| gap7       | AACACGTTGAAAATACTAATAACAAAA | CATGTTAAATATTACAGGAATTT |
| gap8       | GCTCCCTCCACAATCTGAGGAA | TTCTTCTATTAGCAGTATCA |
| gap9       | TTAGGCTTTATTTGTGGGTTGGA | GGCGGCTCAGTTGGGATT |
| gap10      | CAACATCGAGGTCGCGAAC | TGACTCGTAAAAATGATGATA |
| gap11      | TATGCTACCTTTGTACAGTCA | TGATCTCTGTGATACAGGTTT |

doi:10.1371/journal.pone.0016444.t001

Scaffolding of M. vitrata ESTs and transcript mapping

The EST reads were scaffolded against the M. vitrata mitochondrial genome assembly using the NextGENe software (Softgenetics®, State College, PA; parameters used were – matching base number >= 8 and matching base percent >= 70), and contiguous overlapping sequences exported as consensus contigs in FASTA format. The coding region[s] of contig sequences were annotated manually via blastx search of the M. vitrata mitochondrial gene model developed previously, and derived peptide sequence were predicted using the Virtual Ribosome v.1.1 [30].

Scaffolding of M. vitrata ESTs and transcript mapping

The EST reads were scaffolded against the M. vitrata mitochondrial genome assembly using the NextGENe software (Softgenetics®, State College, PA; parameters used were – matching base number >= 8 and matching base percent >= 70), and contiguous overlapping sequences exported as consensus contigs in FASTA format. The coding region[s] of contig sequences were annotated manually via blastx search of the M. vitrata mitochondrial gene model developed previously, and derived peptide sequence were predicted using the Virtual Ribosome v.1.1 [30].
Results and Discussion

Sequence and annotation of the *M. vitrata* mitochondrial genome

Nineteen overlapping PCR products generated using the oligonucleotide primer pairs shown in Table 1 yielded a total of 15,385 bp of sequence data. The data were assembled into a 14,054 bp partial sequence of the *M. vitrata* mitochondrial genome that was submitted to GenBank (HZ751150). This sequence lacks only the displacement loop (D-loop) that encodes the origin of replication or heavy strand promoter (HSP) and light strand promoter (LSP). The *M. vitrata* mitochondrial genome showed a
Table 2. Summary of *Maruca vitrata* predicted tRNAs.

| Gene | Strand | 5' start | 3' stop | Length (bp) | Start | Stop |
|------|--------|----------|---------|-------------|-------|------|
| Nd2  | H      | 1        | 999     | 999         | TTA   | TAA  |
| trnA ^T^HUG | H  | 1007     | 1074    | 68          |       |      |
| trnC ^G^UGY | H  | 1068     | 1131    | 64          |       |      |
| trnT ^H^HAY | L  | 1199     | 1137    | 61          |       |      |
| coxl  | H      | 1207     | 2379    | 1533        | CGA   | TAA  |
| trnH ^H^HUR | H  | 2735     | 2801    | 67          |       |      |
| coxII | H      | 2802     | 3486    | 685         | ATG   | T    |
| trnL ^H^HAR | H  | 3484     | 3554    | 71          |       |      |
| trnG ^G^GAY | L  | 3557     | 3624    | 68          |       |      |
| atpB  | H      | 3624     | 3785    | 162         | ATT   | TAA  |
| atp6  | H      | 3779     | 4453    | 675         | ATG   | TAA  |
| coxI  | H      | 4461     | 5249    | 789         | ATG   | TAA  |
| trnG ^G^GGN | H | 5252     | 5317    | 66          |       |      |
| Nd3  | H      | 5318     | 5671    | 354         | ATT   | TAA  |
| trnG ^G^GNC | H | 5673     | 5738    | 66          |       |      |
| trnA ^A^AGCN | H | 5737     | 5802    | 66          |       |      |
| trnA ^A^AGY | H  | 5801     | 5866    | 66          |       |      |
| trnA ^A^AGN | H  | 5869     | 5936    | 68          |       |      |
| trnG ^G^GAR | H  | 5937     | 6001    | 65          |       |      |
| trnG ^G^HUY | L  | 6069     | 6004    | 64          |       |      |
| Nd 5 | L      | 7790     | 6250    | 1529        | ATT   | TAA  |
| trnH ^H^HAY | L | 7872     | 7806    | 65          |       |      |
| Nd 4 | L      | 9213     | 7872    | 1340        | ATG   | TAA  |
| Nd 4L | L   | 9491     | 9213    | 277         | ATG   | TAA  |
| trnH ^H^ACN | H  | 9465     | 9532    | 68          |       |      |
| trnH ^H^HCCN | L | 9591     | 9531    | 59          |       |      |
| Nd 6 | L      | 9611     | 10040   | 430         | ATT   | TAA  |
| cyrb  | H      | 10377    | 11364   | 988         |       |      |
| trnG ^G^HCN | H  | 11367    | 11432   | 66          |       |      |
| Nd 1  | L      | 12378    | 11449   | 928         | TTA   | TAG  |
| trnH ^H^ACN | L  | 12455    | 12388   | 66          |       |      |
| trnG ^G^GUN | L  | 13780    | 12476   | 1303        |       |      |
| trnA ^A^AGUN | L | 13852    | 13804   | 47          |       |      |
| trnS  | L      | 13863    | 14139   | 277         |       |      |

The tRNAs are presented in the order they occur on the mtgenome along with their beginning and ending nucleotide positions. The anticodon region detected and the corresponding amino acid the predicted tRNA transfer are presented. The orientation is indicated with respect to the heavy (H) or light (L) strand is indicated.

doi:10.1371/journal.pone.0016444.t002

High A+T nucleotide content of 80.0%, typical of animal and insect mitochondrial genome sequences [42,43]. The A+T bias in *M. vitrata* is lower than that in any other mitochondrial genome sequence from Lepidoptera (Fig. 1), but this likely is the consequence of our partial sequence that does not contain the A-T rich D-loop. Although it does not include the entire mitochondrial genome, the *M. vitrata* assembly allowed for the identification and analysis of 2 tRNAs, 13 protein coding genes, and 19 of 22 tRNA genes (Fig. 2). The gene order and orientation seen in *M. vitrata* is typical for insect mitochondrial genomes (Table 2) [9], and is identical to that of the other Crambid species *O. nubilalis* and *O. furnacalis* [7]. Phylogenetic analyses indicated that the *M. vitrata* mitochondrial DNA grouped with a clade containing other species from the lepidopteran Family Crambidae, and clustered with other Families. A clear distinction between moth and butterfly species was apparent (Fig 1; alignment not shown). Thus, the use of the entire mitochondrial genome may clarify phylogenetic relationships that could not be determined using smaller sequence sampling methods [44].

Protein coding regions encompassed 10,266 of the 14,052 bp assembled sequence (75.5%) and showed an A+T content of 82.9%, which is slightly higher than that of the entire sequence, but both were consistent with the high A+T content of insect [9,43] and lepidopteran mitochondrial DNAs [45]. This AT-bias was reflected in the codon usage of the 14 PCGs (Table 3), where phenylalanine (F; UUU), isoleucine (I; AUU), leucine (L; UUA), methionine (M; AUU), and tyrosine (Y; UAU) were proportionately the most represented. The AT-bias in codon usage also was likewise observed in that 3218 of 3422 codons had third positions with either an A or T nucleotide (94.0%), and was significantly higher proportion than at the first codon positions (73%; \( \chi^2 = 4.69, \) \( P \)-value = 0.030) and second codon positions (69.8%; \( \chi^2 = 0.013 \)). This difference may reflect the selection for optimal codon usage [46] or codons that match the anti-codons of tRNAs [47].

The methionine start codons, UTG, were used by 6 of 13 protein coding genes (46.2%). In contrast, an atypical Ile codon (AUU) was used to initiate protein synthesis in the atp8, nd3, nd5, and nd6 genes, and Phe (UUR) was used in the nd1 and nd2 genes, and Arg (CGA) was used for the cox1 gene. The use of Ile for initiation of peptide synthesis has previously been reported in other insects [48] including Lepidoptera [6,7], which suggests that it is common in this insect Order. Additionally, the cox1 gene was first annotated with an Arg (CGA) initiation codon for protein synthesis in *Drosophila yakuba* [9], and the same codon was also subsequently shown for the Lepidopteran species *B. mori* [6], *O. nubilalis* and *O. furnacalis* [7], *Adoxophyes honmani* [49], *Corana raphealis* [50], *Antheraea pernyi* [45], *B. mandarina* [51], *Ochrogaster lunifer* [52], *Artogeia nulicula* [53], *Eriogyna pyretorum* [54], and *Hyphantria cunea* [55]. The nucleotide sequence, TTAG, located immediately upstream of the putative Arg CGR start codon of cox1 was proposed to serve in a non-standard initiation process [6]. Our alignment of the mitochondrial genome sequence from 19 species of Lepidoptera (Fig. 1; alignment not shown) indicated that this putative TTAG initiation was conserved among 15 of 27 (55.6%) coxl genes from full lepidopteran mitochondrial genomes (Fig. 1). The lack of absolute conservation of the TTAG suggests that there may be flexibility in the function of the sequence, or that it may not serve as an initiation codon. In any case, further study is required to determine the mechanism of coxl initiation, but transcript information provided by Stewart and Beckenback (2009) [55] and our data described below indicate that the Arg (CGR) indeed functions as the initiation codon.

The corresponding stop codons used by *M. vitrata* mitochondrial genes were predicted to be TAA in all instances except for coxII and ND1 where T and TAG motifs were observed, respectively (Table 2). Based on invertebrate mitochondrial genetics, we predict that the UAG codon serves as a stop codon, and it has only been observed in *M. vitrata*, *Ostrinia sp.* [7], *Parnassius bremesi* (GenBank ID: FJ871123.1), and *Parnassius maraho* (GenBank ID: FJ810212.1) mitochondrial genomes. The infrequent use of the TAG stop codon is likely a consequence of the high A+T bias at the third codon position [43]. The incomplete stop codon of coxII (T nucleotide only) was previously reported in insects [7,9], and for other mitochondrial genes in bivalves [56] and mammalian species.
Mitochondrial RNA processing occurs by the tRNA punctuation model [3], wherein the cloverleaf-like secondary structures of tRNAs are required for the cleavage of polycistronic transcripts into mRNAs. Maturation of the mitochondrial mRNAs occur by the addition of poly(A) tails, and results in completion of transcripts into mRNAs. Maturation of the mitochondrial mRNAs is required for the cleavage of polycistronic transcripts [3,57], and in M. vitrata it showed that three contigs had more than one gene (contig 4: atp8, atp6, and cytIII; contig 6: nd4 and ndHL, and contig 7: nd8 and cyt6; Fig 2; Table 4). The atp6 and atp8 genes, as well as the nd4 and ndHL genes, were located upon the same mature mitochondrial transcript within the Dipteran insects D. melanogaster and D. pseudoobscura [55,57,58] and the swine Sus scrofa [10]. According to the tRNA punctuated model, all M. vitrata mitochondrial polycistronic transcripts are cleaved at junctions with tRNAs, such that the 3’ end of cistrons are each interspersed by a tRNA gene [3,5]. By default, mitochondrial genes on the same cistron that are contiguous within the genome lack an intervening tRNA. This contiguous structural arrangement of the atp8/atp6/cytIII, nd4/ndHL, and nd8/cyt6 genes was observed in the M. vitrata mitochondrial genome, and was also predicted within the resulting mature transcripts (cistrons; Fig 2; Table 4).

### Table 3. Codon usage in the *Maruca vitrata* mitochondrial genome showing a high A-T nucleotide bias in the 3rd (wobble) position (94.0%).

| Codon | UUU(F) | UUC(F) | UUA(L) | UUG(L) | CUU(L) | CUC(L) | CUA(L) | CUG(L) | AUU(I) | AUC(I) | AUU(I) | AUA(M) | AUG(M) | GUU(V) | GUC(V) | GUA(V) | GUG(V) |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| f.value | 315.0 | 31.0   | 426.0  | 9.0    | 26.0   | 2.0    | 23.0   | 0.0    | 394.0  | 18.0   | 240.0  | 19.0   | 71.0   | 4.0    | 59.0   | 3.0    |
| t.value | 0.18   | 0.01   | 5.26   | 0.11   | 0.32   | 0.02   | 0.28   | 0.00   | 1.91   | 0.09   | 1.85   | 0.15   | 2.07   | 1.28   | 1.72   | 0.09   |
| f.value | 98.0   | 8.0    | 12.0   | 1.0    | 74.0   | 8.0    | 38.0   | 0.0    | 83.0   | 8.0    | 49.0   | 1.0    | 75.0   | 0.0    | 9.0    | 0.0    |
| t.value | 2.69   | 0.32   | 0.06   | 0.05   | 47.2   | 2.27   | 1.27   | 0.00   | 2.35   | 0.23   | 31.9   | 0.03   | 1.18   | 0.00   | 3.0    | 0.00   |
| f.value | 169.0  | 11.0   | 10.0   | 1.0    | 60.0   | 6.0    | 54.0   | 1.0    | 202.0  | 14.0   | 73.0   | 13.0   | 55.0   | 8.0    | 68.0   | 4.0    |
| t.value | 1.88   | 0.12   | 1.82   | 1.18   | 0.62   | 1.08   | 0.96   | 0.04   | 1.87   | 0.13   | 0.70   | 0.30   | 1.70   | 0.25   | 1.89   | 0.11   |
| f.value | 28.0   | 2.0    | 2.0    | 1.0    | 12.0   | 6.0    | 37.0   | 2.0    | 31.0   | 0.0    | 76.0   | 0.0    | 32.1   | 0.0    | 11.0   | 0.36   |
| t.value | 1.87   | 0.04   | 0.04   | 0.18   | 0.94   | 0.00   | 2.90   | 0.16   | 0.85   | 0.03   | 2.09   | 0.00   | 2.45   | 0.36   |

**Table 4. The quantitative expression profile of cistrons transcribed from *Maruca vitrata* mitochondrial genome.**

| Contig (cistron) | Genes within contig | Contig length | Total nt | Mean depth (nt) | Read no. | Proportion |
|------------------|---------------------|---------------|----------|-----------------|----------|------------|
| MvMtd_01         | nd2                 | 1,033         | 7,829    | 7.58            | 24       | 0.2        |
| MvMtd_02         | coxl                | 1,532         | 265,789  | 290.16          | 916      | 6.9        |
| MvMtd_03         | coxl                | 713           | 239,189  | 300.49          | 796      | 6.0        |
| MvMtd_04         | atp8, atp6, coxII   | 2,049         | 171,490  | 83.69           | 558      | 4.2        |
| MvMtd_05         | nd5                 | 1,736         | 27,891   | 16.07           | 81       | 0.6        |
| MvMtd_06         | nd4                 | 1,640         | 89,043   | 54.29           | 297      | 2.2        |
| MvMtd_07         | nd6, cytb           | 1,768         | 117,208  | 66.29           | 359      | 2.7        |
| MvMtd_08         | nd1                 | 1,022         | 86,645   | 84.48           | 278      | 2.1        |
| MvMtd_09         | rml (16s rDNA)      | 1,408         | 1,053,548| 748.29          | 4,288    | 32.1       |
| MvMtd_10         | rrs (12s rDNA)      | 456           | 2870     | 6.29            | 11       | 0.1        |

The number of raw reads (Read no.) that were scaffolded to each contig (cistron) was used to calculate the relative proportion of mitochondrial transcripts within the pooled *M. vitrata* midgut and salivary gland EST library.
The overlap of the gene coding sequence between those of *M. vitrata* *atp8* and *atp6* (7 bp), and *nd1* and *nd4L* (1 bp), would logically dictate that the mature transcripts remain intact in order to preserve their respective coding sequences. This scenario was also predicted due to the absence of an intervening tRNA that would be necessary for transcript cleavage [3,5]. To our knowledge, the polycistronic *atp8/atp6/coxIII* transcript identified from *M. vitrata* has yet to be reported in insects despite ancestral genome arrangements that likely would lead to this tricistronic transcript. *Drosophila melanogaster* showed polyadenylation at the 3′ terminus of bicistronic *atp8/atp6* and monocistronic *coxIII* mitochondrial transcripts [55], which suggests that variation in mitochondrial transcript cleavage events may exist among insect species. Indeed, differences occur between species of salamander, where a tricistronic *atp8/atp6/coxIII* transcript from *Ambystoma tigrinum* is processed into *atp8/atp6* and *coxIII* transcripts within *A. mexicanum* [59]. Our mitochondrial transcript mapping data are the first to be reported for a lepidopteran, so direct comparison with data from other species cannot readily be made. Undoubtedly the inclusion of mitochondrial transcript mapping data from additional species will allow us to understand the functional variation in mitochondrial transcript processing.

Only two out of the eight (25%) mature processed *M. vitrata* mitochondrial transcripts have possible poly(A) tails (coxII and *atp8/atp6/coxIII* cistrons; Supplemental Data S1), which is in contrast to the observation of 3′ polyadenylation for all cistronic transcripts in *Drosophila* [55]. Generation of initial mitochondrial cDNA by use of a poly(T) oligonucleotide would indicate that adenylation of forms of each cistron would be within the EST assemblies, if indeed present within the biological samples. The post-transcriptional 3′ addition of adenine nucleotides results in increased transcript stability in many systems, but it has been linked to increased degradation in plant mitochondria [60]. RNAi knockdown of the enzymatic machinery involved in polyadenylation was previously reported to produce no apparent change in mitochondrial transcript stability in humans [61,62]. Hence, the role of adenylation in mitochondrial transcript stability remains unknown [63]. Adenylation is required by the *coxIII* transcript for the completion of its termination codon, but the functional or protective role of poly(A) tails remains uncertain given the absence from most cistronic transcripts in *M. vitrata*.

**Scaffolding of *M. vitrata* ESTs and transcript mapping**

Since polycistronic transcripts are cleaved to generate mRNAs for each mitochondrial gene-derived cistron, the initial molar ratios are equal and subsequent variation in processed transcript quantities result from differences in mRNA stability [64]. The scaffolds comprised of short read EST data from *M. vitrata* were representative of full-length transcripts (Fig. 2), and appear not to be significantly influenced by reported over-representations of the 3′ end of fragments sometimes observed from 454 data [57]. Furthermore, the scaffolds aligned with our previous contig (cistron) assemblies that were representative of mature processed mitochondrial transcripts, and showed a read depth of 1 to 784 (Fig. 2; Table 4). Similar data from Next Generation Sequencing (NGS) technologies has been used to characterize gene expression [57], and applied to describe the sequence and relative quantities of processed mitochondrial-derived transcripts [10,11]. The quantification described in those studies used simple calculations of the proportion of raw sequence reads mapped to each transcript (cistron) as a relative measure of cellular transcript level, which was used as an estimation of gene expression [37,65].

A similar quantification of *M. vitrata* mitochondrial-derived transcript levels from scaffolded EST data indicated a ~321-fold variance in cistron level, where the *rnl* gene was present in the highest proportion (Table 4). The *rnl* transcript was also present in the highest proportion of *D. melanogaster* [11] and porcine cistrons [10], and the analogous large abundance of rrmL transcripts in human expression data was attributed to the presence of a second HSP upstream of the gene that terminates in proximity near the tRNA-Lys/Lys downstream of the 16 S rRNA [66,67,68]. The HSP and LSP sequences have diverged rapidly in mammalian systems [69], such that defining these functional regions is difficult by comparative sequence analyses [70]. The proportional range of cistron levels for those that encode for proteins involved in the same process showed ≤11.15-fold difference in cistron quantities for the NADH dehydrogenase complex (*nd1* vs. *nd2*), and ≤3.59-fold variance among cistrons for the cytochrome *c* oxidase (*coxII* vs. *coxIII*). The estimated expression levels of *coxI* and *coxII* were approximately equal in *M. vitrata* ESTs. An overall lack of correlation between transcript levels within a protein complex has been observed among species of *Drosophila* [11], and likely can be attributed to differences in individual transcript stabilities [64]. Furthermore, these estimates do not take into account any differences in protein stability or turnover in the cell that would influence cellular metabolic processes [71]. The expression data represent additional data for annotation, but comparison with ESTs derived from different *M. vitrata* tissues or growth stages are likely to provide valuable information regarding metabolic variation.

**Conclusions**

This is the first study of a species of Lepidoptera to provide functional annotation of a complete mitochondrial genome using gene expression data. These results will be valuable for future studies to provide comparisons for other species of Lepidoptera and to describe tissue-specific mitochondrial transcript processing pathways. This study provides a paradigm for the mating of EST and functional mitochondrial genome annotations that can be accomplished for any species for which both data sets are present.

**Supporting Information**

**Data S1** *Maruca vitrata* mitochondrial DNA assembled contigs. (TXT)

**Data S2** *Maruca vitrata* 454 raw reads corresponding to the mitochondrial genome assembly. (TXT)

**Acknowledgments**

We thank Susan Balfe, Department of Entomology, University of Illinois at Urbana-Champaign, for her technical assistance on this project.

**Author Contributions**

Conceived and designed the experiments: VMM BSC BRP. Performed the experiments: VMM BSC TA WS MNB. Analyzed the data: VMM BSC. Contributed reagents/materials/analysis tools: RLH MNB AS CLB IB MFI FGC RS JA LLM BRP. Wrote the paper: VMM BSC RLH TA MJS WS MNB AS CLB IB MFI FGC RS JA LLM BRP.
References

1. Kasamura H, Robberson DL, Vinograd J (1971) A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. Proc Natl Acad Sci USA 68: 385–388.
2. Fernandez-Silva P, Enriquez JA, Montoya J (2003) Replication and transcription of mammalian mitochondrial DNA. Exp Physiol 88: 41–56.
3. Ojala D, Montoya J, Attardi G (1981) tRNA punctuation model of RNA processing in human mitochondria. Nature 292: 470–474.
4. Raheja AI (1974) Report on the insect pests of grain legumes in northern Nigeria. 1st IITA grain legume improvement workshop 1973 International Institute of Tropical Agriculture, Ibadan, Nigeria. pp 295–299.
5. Sharma HC (1998) Bionomics, host plant resistance, and management of the maize earworm, Maruca vitrata Fabricius (Lepidoptera: Pyraloidea: Crambidae). Mol Biol Rep.
6. Yukuhiro K, Sezutsu H, Itoh M, Shimizu K, Banno Y (2002) Significant levels of sequence divergence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkworm, Bombyx mori, and its close relative, the domesticated silkworm, Bombyx mori. Mol Biol Evol 19: 1385–1389.
7. Goates BS, Sumerford DV, Hellmich RL, Lewis LC (2005) Partial mitochondrial genome sequences of Ostrinia nubilalis and Ostrinia furnacalis. Int J Biol Sci 1: 13–18.
8. Liao F, Wang L, Wu, Li YP, Zhao L, et al. (2010) The complete mitochondrial genome of the fall webworm, Hyphantria cunea (Lepidoptera: Arctiidae). Int J Biol Sci 6: 172–186.
9. Clary DO, Wolenholme DR (1985) The mitochondrial DNA molecule of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code. J Mol Evol 22: 252–271.
10. Scheibye-Ahing K, Cirera S, Gilchrist MJ, Fredholm M, Gorodkin J (2007) EST analysis on pig mitochondria reveal novel expression differences between domestic pig and wild boar adult tissues. BMC Genomics 8: 362.
11. Torres TT, Dolezal M, Schlotterer C, Ottenwalder B (2009) Expression profiling of Drosophila melanogaster mitochondria via deep mRNA sequencing. Nucleic Acids Res 37: 7509–7518.
12. Huang CC, Peng WK, Telkar NS (2003) Parasitoids and other natural enemies of Maruca vitrata feeding on Solanum cannabina in Taiwan. BioControl 48: 407–416.
13. Arockoson DY, Tamo M, Clouter C, Broder J (2006) Larval parasitoids occurring on Maruca vitrata Fabrissius (Lepidoptera: Pyralidae) in Benin, West Africa. Agric Ecosyst Environ 113: 320–325.
14. Sharma HC, Saxena KB, Bhatwar VR (1999) The legume bean borer, Maruca vitrata: bionomics and management. Information bulletin 55 International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India.
15. Singh SR, van Emden HF (1979) Insect pests of grain legumes. Annu Rev Entomol 24: 375–393.
16. Chinh NT, Dzung DT, Long TD, Tam HM, Ramakrishna A, et al. (2000) Occurrence on Maruca vitrata feeding on Dendrogramma pyretorum (Lepidoptera: Saturniidae). Int J Biol Sci 5: 351–365.
17. Xia X (1996) Maximizing transcription efficiency causes odorus usage bias. Genetics 144: 1309–1320.
18. Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. Genetics 137: 97–117.
19. Zardoya R, Meyer A (2001) On the origin of and phylogenetic relationships among living amphibians. Proc Natl Acad Sci U S A 98: 7380–7383.
20. Liu Y, Li Y, Pan M, Dai F, Zhu X, et al. (2006) The complete mitochondrial genome of the Chinese oak silkmoth, Antheraea pernyi (Lepidoptera: Saturniidae). Acta Biochim Biophys Sin (Shanghai) 40: 693–703.
21. Lee ES, Shin KS, Kim MS, Park H, Cho S, et al. (2006) The mitochondrial genome of the smaller tea tortrix Adoxophyes honmae (Lepidoptera: Tortricidae). Gene 376: 52–57.
22. Kim I, Lee EM, Seol KY, Park YW, Lee YB, et al. (2006) The mitochondrial genome of the Korean hairstreak, Coreana raphaides (Lepidoptera: Lycaenidae). Jpn J Mol Biol 15: 217–226.
23. Pan M, Yu Q, Xia Y, Dai F, Lai Y, et al. (2008) Characterization of mitochondrial genome of Chinese wild mulberry silkworm, Bombyx mori. (Lepidoptera: Bombycidae): Sci China C Life Sci 51: 693–701.
24. Salvato P, Samonato M, Battini A, Negriolo E (2008) The complete mitochondrial genome of the bag-shepherd moth Ochrogaster lunifer (Lepidoptera: Notodontidae). BMC Genomics 9: 331.
25. Hong G, Jiang S, Yu M, Yang Y, Li F, et al. (2009) The complete nucleotide sequence of the mitochondrial genome of the cabbage butterfly, Aricia melete (Lepidoptera: Pieridae). Acta Biochim Biophys Sin (Shanghai) 41: 446–455.
26. Jiang S, Hong G, Yu M, Li N, Yang Y, Lai Y, Wei Z (2009) Characterization of the complete mitochondrial genome of the giant silk moth, Eriogyna pyretorum (Lepidoptera: Tortricidae). Gene 455: 49–57.
27. Jiang WP, Li JL, Zheng RL, Wang GL (2010) Analysis of complete mitochondrial genome of Cristaria platura. Yi Chuan 32: 135–162.
28. Torres TT, Metta M, Ottenwalder B, Schlotterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Res 18: 172–177.
29. Berthier F, Renaud M, Alaziari S, Durand R (1980) RNA mapping on Drosophila melanogaster DNA precursors and template strands. Nucleic Acids Res 14: 4519–4533.
30. Wernersson R (2006) Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res 34: W535–W538.
31. Lowe TM, Eddy SR (1997) tRNAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 953–964.
32. Rutherford K, Parkhill J, Crook J, Hornsell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
34. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGAS: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
35. Tamura K, Nei M, Shimizu K, Suyama M (2005) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci 101: 11030–11035.
36. Kumar S, Gadaugas SR (2001) Dnaparts Index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics 158: 1211–1227.
37. Felsenstein J (1985) PHYLIP:Phylogeny Inference Package (Version 3.2). Cladistics 3: 164–166.
38. Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51: 79–94.
39. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 273–282.
40. Page RDM (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357–358.
41. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
42. Boone JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27: 1767–1780.
61. Piwowarski J, Grzechnik P, Dziembowski A, Dmochowska A, Minczuk M, et al. (2003) Human polynucleotide phosphorylase, hPNPase, is localized in mitochondria. J Mol Biol 329: 853–857.

62. Tomecki R, Dmochowska A, Gewartowski K, Dziembowski A, Stepień PP (2004) Identification of a novel human nuclear-encoded mitochondrial poly(A) polymerase. Nucleic Acids Res 32: 6001–6014.

63. Bobrowicz AJ, N. Lightowlers RN, Chrzanowska-Lightowlers Z (2008) Polyadenylation and degradation of mRNA in mammalian mitochondria: a missing link? Biochem Soc Trans 36: 517–519.

64. Piechota J, Tomecki R, Gewartowski K, Szczesny R, Dmochowska A, et al. (2006) Differential stability of mitochondrial mRNA in HeLa cells. Acta Biochim Pol 53: 157–163.

65. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.

66. Christianson TW, Clayton DA (1986) In vitro transcription of human mitochondrial DNA: accurate termination requires a region of DNA sequence that can function bidirectionally. Proc Natl Acad Sci U S A 83: 6277–6281.

67. Montoya J, Christianson T, Levens D, Rabinowitz M, Attardi G (1982) Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. Proc Natl Acad Sci U S A 79: 7195–7199.

68. Montoya J, Gaines GL, Attardi G (1983) The pattern of transcription of the human mitochondrial rRNA genes reveals two overlapping transcription units. Cell 34: 151–159.

69. Fisher RP, Parisi MA, Clayton DA (1989) Fisher RP, Parisi MA, Clayton DA. 1989. Flexible recognition of rapidly evolving promoter sequences by mitochondrial transcription factor I. Gene Devel 3: 2202–2217.

70. Bravo JP, Felipe J, Zanatta DB, Silva JLC, Fernandez MA (2008) Sequence analysis of the mitochondrial DNA control region in the sugarcane borer Diatraea saccharalis (Lepidoptera: Crambidae). Braz Arch Biol Technol 51: 671–677.

71. Pratt JM, Petty J, Ribas-Garcia I, Robertson DH, Gaskell SJ, Oliver SG, Beynon RJ (2002) Dynamics of protein turnover, a missing dimension in proteomics. Mol Cell Proteomics 1: 579–591.