Chasing sympatric speciation: The relative importance and genetic basis of prezygotic isolation barriers in diverging populations of Spodoptera frugiperda

Hänniger, S.

Publication date
2015

Document Version
Final published version

7

ANNOTATION OF CIRCADIAN CLOCK GENES IN THE GENOME OF *SPODOPTERA FRUGIPERDA*

S. Hänniger[1*], P. Dumas[2*], D.G. Heckel[1] & A.T. Groot[1,2]

[1]Department of Entomology, Max Planck Institute for Chemical Ecology, Hans-Knöll Strasse 8, 07745 Jena, Germany; [2]Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

*Shared first authorship, both authors contributed equally

Part of *The* Spodoptera frugiperda *Whole Genome Sequencing Project*.
The Fall armyworm International Public Consortium (FAW-IPC) (in prep.)
http://www6.inra.fr/lepidodb/SfruDB

INTRODUCTION

All life on earth is subject to rhythmical changes of light and temperature as day and night alternate in a rhythm of roughly 24 hours. Consequently, organisms have evolved reliable internal clocks that are entrained by changing external factors like light and temperature. These circadian clocks (from Latin circa = approximately, dia = day) enable them to predict these changes and 'schedule' physiological as well as behavioral processes in beneficial time windows, e.g. to avoid heat stress. Hence, circadian clocks are involved in almost all physiological and behavioral processes in animals such as mating, feeding and cell-division (Dunlap 1999; Wijnen and Young 2006). They are composed of genes and their protein products that form interlocked transcriptional/translational feedback loops, which repeat a feedback cycle every approximately 24 hours (Edery 2000). Key genes involved in the circadian clock are well characterized in a number of model organisms and are conserved within kingdoms and particularly strongly conserved within the animal kingdom (reviewed e.g. in Rear and Allada 2012). For example the important signal sensor region PAS-B of the protein CLOCK shows 80-88% sequence similarity between the silkmoth *Antheraea pernyi*, the fruitfly *Drosophila melanogaster* and the mouse *Mus musculus*. Similarly the basic helix-loop-helix (bHLH) domain of the CLK protein, facilitating its DNA binding, shows 59-76% sequence similarity between these species (Chang et al. 2003).

Homologues of most of the known *Drosophila* clock genes are found in Lepidoptera (Sandrelli et al. 2008; Zhan et al. 2011). In general, the lepidopteran clockwork (like the clockwork of *Drosophila*) is proposed to consist of two feedback loops (Hardin 2005; Zhan et al. 2011). They are interlocked by both involving the genes (always named in lower case italic letters) *clock* (*clk*) and *cycle* (*cyc*) and their protein products (always named in upper case letters) CLOCK (CLK) and CYCLE (CYC). In the core transcriptional/translational feedback loop CLK:CYC heterodimers drive *timeless* (*tim*), *period* (*per*) and *cryptochrome 2* (*cry2*) transcription. TIM, PER and CRY2 form a complex that enters the nucleus, where CRY2 inhibits the transcription mediated by CLK:CYC, including transcription of *tim*, *per* and *cry2*. The light-dependent CRYPTOCHROME 1 (CRY1) in involved in TIM degradation, facilitating light-entrainment of the clock. The degradation of TIM and PER is signaled by SUPERNUMERARY LIMBS (SLIMB) and JETLAG (JET) and kinases and phosphatases like CASEIN KINASE II (CKII), DOUBLETIME (DBT) and PROTEIN PHOSPHATASE 2A (PP2A) are involved in posttranslational modifications of PER and TIM. In the modulatory feedback loop, VRILLE (VRI) inhibits *clk* transcription and PAR DOMAIN PROTEIN 1 (PDP1) promotes *clk* transcription. Both *vri* and *pdp1* transcription are driven by the CLK:CYC heterodimer. The amplitude of the clock is modified by CLOCKWORKORANGE (CWO) (Hardin 2005; Zhan et al. 2011). From this list of circadian clock genes (Table 1), in total nine critical genes involved in the core

feedback loop (*clk*, *cyc*, *per*, *tim*, *cry2*), the modulatory feedback loop (*vri*, *pdp1*), the light entrainment (*cry1*) and in the posttranslational modification (*dbt*) were chosen for annotation in as a starting set.

The annotation of the circadian clock genes is part of the *Spodoptera frugiperda* whole genome sequencing project of The Fall Armyworm International Public Consortium (FAW-IPC) (in prep.). The project aims to sequence and assemble the whole genome of both the corn-strain and rice-strain of *S. frugiperda* and annotate the genomes by identifying genetic elements (genes and transposable elements) in the genome and adding relevant biological information (name and function) to these elements. The database SfruDB, which provides the genome sequences, transcriptional data and an annotation interface (WebApollo) to the consortium is hosted by the French National Institute for Agricultural Research (INRA) and can be found at http://www6.inra.fr/lepidodb/SfruDB.

Table 1. List of clock genes with abbreviations, their part in the clockwork and their annotation status.

Gene name	Abbrev.	Clockwork part	Annotation status
Clock	clk	core feedback loop	yes
Cycle	cyc	core feedback loop	yes
Timeless	tim	core feedback loop	yes
Period	per	core feedback loop	yes
Cryptochrome 2	cry2	core feedback loop	yes
Cryptochrome 1	cry1	photic entrainment	yes
Supernumerary limbs	slimb	TIM and PER degradation	not yet
Jetlag	jet	TIM and PER degradation	not yet
Casein kinase II	ckII	posttranslational modification	not yet
Doubletime	dbt	posttranslational modification	yes
Protein phosphatase 2A	pp2a	posttranslational modification	not yet
Vrille	vri	modulatory feedback loop	yes
PAR domain protein 1	pdp1	modulatory feedback loop	yes
Clockworkorange	cwo	amplitude modification	not yet

METHODS

To annotate the clock genes, we conducted the following steps. First, in the GenBank database of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), DBT, TIM, CRY1, CRY2, CLK and PER amino acid sequences were identified in the closely related species *Spodoptera exigua*. As for PDP1 and CYC, no sequence was available from *S. exigua*, but PDP1 was identified in *Drosophila melanogaster* and CYC in *Danaus plexippus*. For VRI, the *S. frugiperda* DNA sequence obtained in our laboratory was used.

Secondly, for CLK, PER, CYC, VRI and PDP1, homologs were BLAST searched in an RNAseq assembly from larval midguts of both strains using tblastn and the program SEQtools (Rasmussen 2002) and using the protein sequences obtained from NCBI. The used RNAseq assemblies are now available on SfruDB WebApollo: A1: corn strain midguts (larvae fed on maize), B1: corn strain midguts (larvae fed on pinto bean diet), C1: rice strain midguts (larvae fed on maize), D1: rice strain midguts (larvae fed on pinto bean diet). The cDNA sequences of *clk*, *per*, *cyc*, *vri* and *pdp1* that were obtained from the RNAseq assemblies were then used for the next step. For DBT, TIM, CRY1 and CRY2 we used the protein sequences of *S. exigua* that were obtained from the GenBank database of NCBI in the first step.

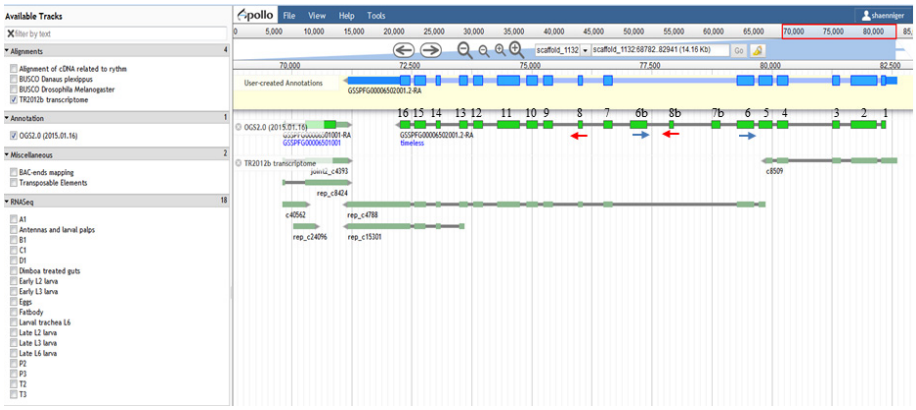

FIGURE 1. Screenshot of the annotation of timeless in the WebApollo annotation platform of the SfruDB. Available resources can be chosen on the left panel. The User-created Annotations (UCA) shows own annotations and that of other users. Below the UCA field appear the resources chosen on the left panel. They can be dragged and dropped onto the UCA field. In this example, a previous annotation (OGS2.0) was corrected (3 exons deleted) and the UTRs were added based on the TR2012b transcriptome (see results and discussion for details). Numbers over OGS2.0 correspond to exon numbers. Arrows under OGS2.0 indicate the approximate position of primers (see results and discussion for details).

In the third step, we BLAST searched *Spodoptera frugiperda* homologs of the nine chosen clock genes in the corn-strain variant of the genome in the SfruDB database using tblastn (for protein query sequences) or blastn (for cDNA query sequences). In the annotation platform WebApollo, we annotated the corresponding transcripts, if present, and corrected the exon-intron structure based on homology between the *S. frugiperda* and *S. exigua* sequences. We further corrected the gene structure, including 5' and 3' UTRs, based on transcriptome data available in WebApollo (RNAseq and TR2012b). Figure 1 shows an example for *timeless*.

In the fourth step, we carefully named the alleles and parts of all genes, if present.

As a final check, we retrieved the created protein sequences from the annotated genes in WebApollo and performed blastp against insects on the NCBI blast server to confirm homology in other lepidopteran insects.

RESULTS AND DISCUSSION

In the corn-strain variant of the *Spodoptera frugiperda* genome in SfruDB, we found the homologs of all nine chosen clock genes: c*lock*, *cycle*, *timeless*, *period*, *cryptochrome 2*, *cryptochrome 1*, *double-time*, *vrille* and *pdp1*. The genes have not been annotated in the rice-strain variant yet, as it only became available for annotation when this thesis was submitted.

The exon-intron structures of the annotated genes as well as the distribution on different scaffolds are summarized in Table 2 and shown in Figure 2 below.

Most of the annotated circadian clock genes, i.e. *clk*, *cyc*, *per*, *pdp1* and *cry1*, are located on several scaffolds (see Table 2). Since the genome assembly is still very fragmented, i.e. consists of some large and several thousand small scaffolds that are not connected, this information is useful to merge scaffolds or at least determine the right order and orientation of the scaffolds and will be used by the bioinformatics experts developing and improving the assembly in the near future.

For *clk* the 3' UTR and the first exon could not be annotated, because a homologue was missing in the genome assembly. The annotated parts of the gene spread over five scaffolds and consist of 12 exons, thus five scaffolds could be arranged in the right order or possibly even merged. For *cyc*, all parts could be annotated, thirteen exons were distributed over three scaffolds. The 25 exons and the 3' UTR of *per* were distributed over 7 scaffolds. Exon 12 is missing in the genome assembly and could thus not be annotated. Also the 5' UTR could not be annotated. All parts of *pdp1* could be annotated. The seven exons were distributed over four scaffolds. The twelve exons and the 3' UTR of *cry1* were annotated on three different scaffolds. The 5' UTR could not be annotated.

As evident from Table 2, many UTRs could only be annotated partially or not at all. This is because these UTRs are (partially) located on a separate scaffold and the Web Apollo interface does not allow the annotation of a sequence as UTR when it is not connected to a coding sequence. This extra information could be taken into account for merging or arranging the scaffolds, equivalent to the information from separated exons. In case of *vri* we can contribute one missing piece of sequence information to the genome assembly. The coding region of *vri*, consisting of only one exon, was annotated on one scaffold. However, the upstream part of the 5' UTR is located on a separate scaffold. The 5' UTR is split by a large (>7,000 bp) intron that is not fully present in the genome assembly. Because of this large gap the two scaffolds cannot be merged. We could obtain the sequence of this intron by

sequencing two BAC clones (AUA0AAA25YL06 and AUA0AAA20YH15) whose ends were mapping to the scaffolds containing *vri*. This sequence information will be useful to merge the two scaffolds and close a large gap in the genome assembly.

TABLE 2. Summary of annotated circadian clock genes. Numbers in brackets include the exons that could not be annotated and their unknown scaffolds.

Gene name	Symbol	Nr of exons	Nr of scaffolds	5' UTR annotated?	3' UTR annotated?
Clock	CLK	12 (13)	5 (6)	no	yes
Cycle	CYC	13	3	partially	yes
Timeless	TIM	16	1	yes	yes
Period	PER	24 (25)	7 (8)	partially	partially
Cryptochrome 2	CRY2	9	1	no	partially
Vrille	VRI	1	1	partially	yes
PAR domain protein 1	PDP1	7	3	yes	yes
Double-time	DBT	8	1	yes	no
Cryptochrome 1	CRY1	12	3	partially	yes

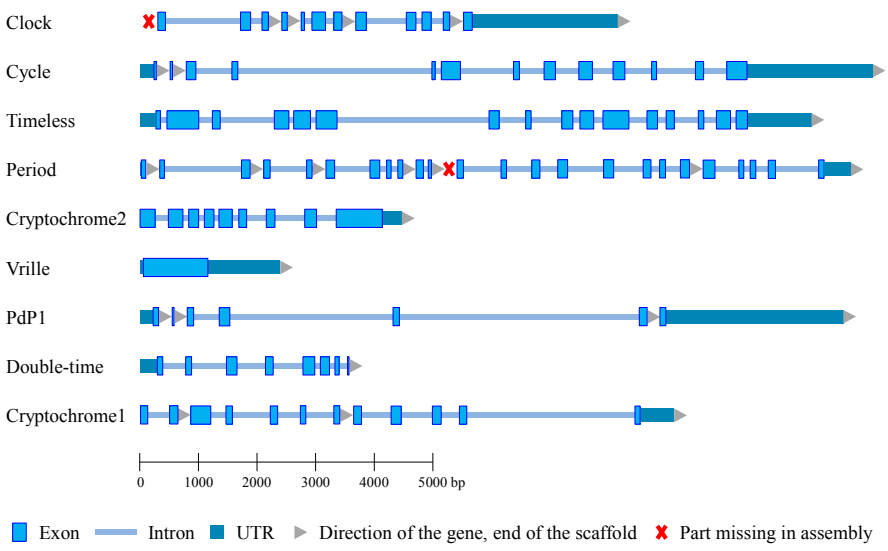

FIGURE 2. Exon-intron structure of the clock genes annotated in the whole genome assembly of *S. frugiperda*. The 3' end of a gene part that is located on one scaffold is indicated by a grey arrow head. Introns that are spanning scaffolds are not depicted. Exons that could not be annotated are indicated by a red cross.

A special case occurred when annotating *tim*. All 16 exons and the UTRs of the gene are located on one scaffold (see Figure 1). The sequences of exons 6, 7 and 8 are replicated in concert on this scaffold as follows: 6-7-8-6-7-8, resulting in 19 exons instead of 16 (see OGS2.0 in Figure 1, numbers above the OGS2.0 correspond to exon numbers, duplicated exons are named 6b, 7b and 8b). The sequences of these exons are only present once in e.g. the *S. exigua tim* gene. The transcriptomic and RNAseq data available in Web Apollo indicate that the replication is a mistake in the genome assembly rather than a genuine duplication of this part of the gene. This can be seen in the missing transcriptomic information in TR2012b for the exons 6b, 7b and 8b of OGS2.0 in Figure 1. To confirm a mistake in the genome assembly and rule out a duplication in the *tim* gene of *S. frugiperda*, we used a forward primer in exon 8 (red arrows in Figure 1) and a reverse primer in exon 6 (blue errors in Figure 1) for a PCR, which would only amplify a sequence if (a second) exon 6 (6b) would follow an exon 8 (8b). The PCR did not amplify a product. Thus, the exons 6b, 7b and 8b were removed from the annotation, leaving *tim* with the expected 16 exons and a sequence homologous to that of *S. exigua tim*.

In conclusion, the annotation of the nine clock genes in SfruDB has resulted in the elucidation of the exon-intron structure of these genes in the corn-strain of *S. frugiperda*. Through this annotation, three sequences of three scaffolds each can be arranged in the right order or could possibly be merged, as well as one sequence of five scaffolds and one sequence of seven scaffolds. The arrangement of the scaffolds, and with this the overall genome annotation of *S. frugiperda*, would benefit from the possibility to annotate UTRs that are located on separate scaffolds and not attached to a coding sequence of a gene.

REFERENCES

CHANG, D. C., H. G. MCWATTERS, J. A. WILLIAMS, A. L. GOTTER, J. D. LEVINE, S. M. REPPERT. 2003. Constructing a feedback loop with circadian clock molecules from the silkmoth, *Antheraea pernyi*. Journal of Biological Chemistry 278:38149-38158.

DUNLAP, J. C. 1999. Molecular bases for circadian clocks. Cell 96:271-290.

EDERY, I. 2000. Circadian rhythms in a nutshell. Physiological Genomics 3:59-74.

HARDIN, P. E. 2005. The circadian timekeeping system of *Drosophila*. Current Biology 15:R714-R722.

LEAR, B. C. AND R. ALLADA. 2012. Circadian rhythms. Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, Chichester.

RASMUSSEN, S. W. 2002. SEQtools, a software package for analysis of nucleotide and protein sequences.

SANDRELLI, F., R. COSTA, C. P. KYRIACOU, E. ROSATO. 2008. Comparative analysis of circadian clock genes in insects. Insect Molecular Biology 17:447-463.

WIJNEN, H. AND M. W. YOUNG. 2006. Interplay of circadian clocks and metabolic rhythms. Pp. 409-448. Annual Review of Genetics. Annual Reviews, Palo Alto.

ZHAN, S., C. MERLIN, J. L. BOORE, S. M. REPPERT. 2011. The monarch butterfly genome yields insights into long-distance migration. Cell 147:1171-1185.