Annotated expressed sequence tags for studies of the regulation of reproductive modes in aphids

Denis Tagu, N. Prunier-Leterme, F. Legeai, J.-P Gauthier, A Duclert, B. Sabater-Munoz, J Bonhomme, Jean-Christophe Simon

ELSEVIER

Insect Biochemistry and Molecular Biology 34 (2004) 809–822

Insect Biochemistry and Molecular Biology

www.elsevier.com/locate/ibmb

Annotated expressed sequence tags for studies of the regulation of reproductive modes in aphids

D. Tagu[a,*], N. Prunier-Leterme[a], F. Legeai[b], J.-P. Gauthier[a], A. Duclert[b], B. Sabater-Muñoz[a], J. Bonhomme[a], J.-C. Simon[a]

[a] *INRA Rennes, UMR INRA-AGROCAMPUS BiO3P, BP 35327, Le Rheu F-35653, cedex, France*
[b] *INRA, URGI—Genoplante Info, Infobiogen, 523 place des Terrasses, F-91000, Evry, France*

Received 15 March 2004; received in revised form 23 April 2004; accepted 17 May 2004

Abstract

The damaging effect of aphids to crops is largely determined by the spectacular rate of increase of populational expansion due to their parthenogenetic generations. Despite this, the molecular processes triggering the transition between the parthenogenetic and sexual phases between their annual life cycle have received little attention. Here, we describe a collection of genes from the cereal aphid *Rhopalosiphum padi* expressed during the switch from parthenogenetic to sexual reproduction. After cDNA cloning and sequencing, 726 expressed sequence tags (EST) were annotated. The *R. padi* EST collection contained a substantial number (139) of bacterial endosymbiont sequences. The majority of *R. padi* cDNAs encoded either unknown proteins (56%) or housekeeping polypeptides (38%). The large proportion of sequences without similarities in the databases is related to both their small size and their high GC content, corresponding probably to the presence of 5′-unstranslated regions. Fifteen genes involved in developmental and differentiation events were identified by similarity to known genes. Some of these may be useful candidates for markers of the early steps of sexual differentiation.

Keywords: Aphid; cDNA; EST; Development; Reproduction

1. Introduction

Aphids are plant-sucking insects which cause serious damage on most cultivated and ornamental plants world-wide. They affect plant growth either directly by depletion of sucrose and amino-acids content, or indirectly, especially by plant virus transmission. The impact of aphid on crops is largely determined by their high rates of multiplication and dispersal conferred by both (1) a peculiar mode of reproduction which is cyclically parthenogenetic (i.e. alternance of many parthenogenetic generations and a single sexual generation within the annual life cycle) and (2) an amazing phenotypic plasticity—called polyphenism. Aphid polyphenism allows one single genotype (a clone) to adapt to rapidly changing environmental conditions by expressing multiple and often morphologically distinct phenotypes. Up to now, most effort has focused on the ecological factors and the physiological changes responsible for this aphid polyphenism, but its molecular basis remains largely unexplored. Changes in day length can induce the switch from parthenogenetic to sexual reproduction within an aphid colony (Dixon, 1998) and the length of the dark phase is one of the key factors determining the reproductive outcome (reviewed by Hardie and Nunes, 2001). Perception by aphids of photoperiodic changes is independent of eyes and localised illuminations of the head capsule indicated that the central dorsal region was most sensitive to the photoperiodic variations. It is probable that proteins involved in photoperception (e.g. opsins) and phototransduction (e.g. arrestins) are localised in that brain region but nothing is known about these putative photoreceptors and phototransducers in aphids. Early

[*] Corresponding author. Tel.: +33-223-48-51-65; fax: +33-223-48-51-50.
E-mail address: tagu@rennes.inra.fr (D. Tagu).

0965-1748/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2004.05.001

experiments performed by microcautery disruption of cells indicated a group of clock-neurosecretory cells in the protocerebrum which were involved in the photoperiodic response of aphids. The first aphid gene regulated by reproductive polyphenism has been recently identified; it corresponds to a putative amino acid transporter in GABAergic neurons which could play a role in the generation and modulation of circadian rhythmicity (Ramos et al., 2003). Triggering the sexual response of aphids to day-length must result from several cascades of events which are far from being understood. Unravelling the nature of the molecular events underpinning polyphenisms in aphids is thus becoming a necessary step towards (i) the elucidation on the coexistence of sexual and asexual reproductive modes in aphids and (ii) the development of innovative control strategies against these important crop pests.

Expressed sequence tags (ESTs) have become an effective means of gene discovery and therefore, in an effort to create a resource for gene discovery in aphids and to begin the characterization of their genetic complement, we have generated ESTs from the aphid *Rhopalosiphum padi*. This aphid causes serious damages to most cereals world-wide and is an efficient vector of cereal and barley yellow dwarf viruses (Gray and Gildow, 2003). Here, we describe the annotation of 726 new ESTs together with 4358 mRNA and EST sequences from different aphid species found in public DNA sequence databases.

2. Materials and methods

The cyclically parthenogenetic line, h3 (i.e. that alternates several parthenogenetic generations and a single sexual generation within the annual life cycle) of *R. padi* was isolated at Rennes (France) in 1992 from its winter host, the bird cherry tree, *Prunus padus* and has been reared since on wheat in the laboratory. It was maintained in conditions of continuous parthenogenetic reproduction under long photoperiod (16-h light/8-h dark) and warm temperature (18 °C). In order to enrich the cDNA library in transcripts up-regulated under sexual-reproductive mode, insects were placed in sex-induction conditions using a standard protocol (Simon et al., 1991) (Fig. 1). Briefly, fourth instar parthenogenetic nymphs (N4) were transferred on wheat seedlings (6 nymphs per plant) under short photoperiod and low temperature (10-h light/14-h dark at 12 °C). After 8 days, they became adults and gave birth to first instar parthenogenetic larvae (L1). About 100 of these larvae were individually transferred to new wheat seedlings. After two moults, third instar larvae (L3) were collected and immediately frozen in liquid nitrogen, and kept at −80 °C until use. These L3 larvae corresponded to future wingless parthenogenetic females induced to produce sexual forms (Fig. 1).

Total RNA from 74 whole-L3 larvae (see above) was extracted using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) in the RTL extraction buffer, following the manufacturer instructions. cDNA synthesis and cloning were performed with the Creator™ Smart™ cDNA Library Construction Kit (BD Biosciences Clontech, Palo Alto, CA, USA). Briefly, enriched full-length double-stranded cDNA were obtained by RT-PCR and ligated into the pDNR-LIB plasmid. Ligation products were electroporated in electrocompetent *Escherichia coli* TOP10 (Invitrogen, Paisley, United-Kingdom) cells. Bacterial colonies ($n = 1056$) were inoculated into 96-well plates containing selective LB medium and 10% (v/v) glycerol, and grown overnight in stand culture at 37 °C. Backup plates were also created. Plates were stored at −80 °C.

Polymerase chain reaction (PCR) was performed in 15 μl of final volume from 1 μl of defrosted bacterial glycerol stock as template and pDNR-lib forward primer (5′-GCCGCATAACTTCGTATAGCA-3′) and pDNR-lib reverse primer (5′-CCAGGATCTCC-TAGGGAAACA-3′) at 0.2 μM final concentration. The PCR consisted of 94 °C for 2 min, 94 °C for 30 s, 57 °C for 30 s, 72 °C for 2 min for 30 cycles and a final extension at 72 °C for 4 min (PCR Express, Hybaid-Promega, Madison, WI, USA). A 1 μl aliquot of each reaction was analysed on a 1% agarose gel and stained with ethidium bromide for the control of size and quality of the PCR products. Excess primers and nucleotides were removed by filtration on Sephadex (SigmaSpin Post-Reaction Clean-Up Plates Kit, Sigma, St Louis, Missouri, USA). A 1 μl aliquot of each reaction was analysed on a 1% agarose gel and stained with ethidium bromide for the determination of the concentration of the PCR products. The resulting purified PCR products were then rearranged in 96-well plates and used as templates (30–50 ng) for a sequencing reaction using 0.5 mM of the pDNR-lib forward primer with the ABI Prism BigDye v3.1 sequencing kit (ABI, Foster City, CA, USA). Sequencing reactions were performed at 94 °C for 1 min, 94 °C for 15 s, 57 °C for 7 s, 60 °C for 4 s for 50 cycles (PCR Express, Hybaid-Promega, Madison, WI, USA). The products of the sequencing reaction were purified by filtration on Superfine G-50 (Amersham Pharmacia Biotech), using the MultiScreen system (Millipore, Billerica, MA, USA). Samples were eluted in water and ready for capillary electrophoresis separation and detection by the automated multicapillary sequencer (AB 3100, Applied Biosystem, Foster City, CA, USA) at the sequencing facilities of OUEST-Genopole® (Roscoff, France). The name given to each EST corresponds to the name of the cDNA library (RpL3i for

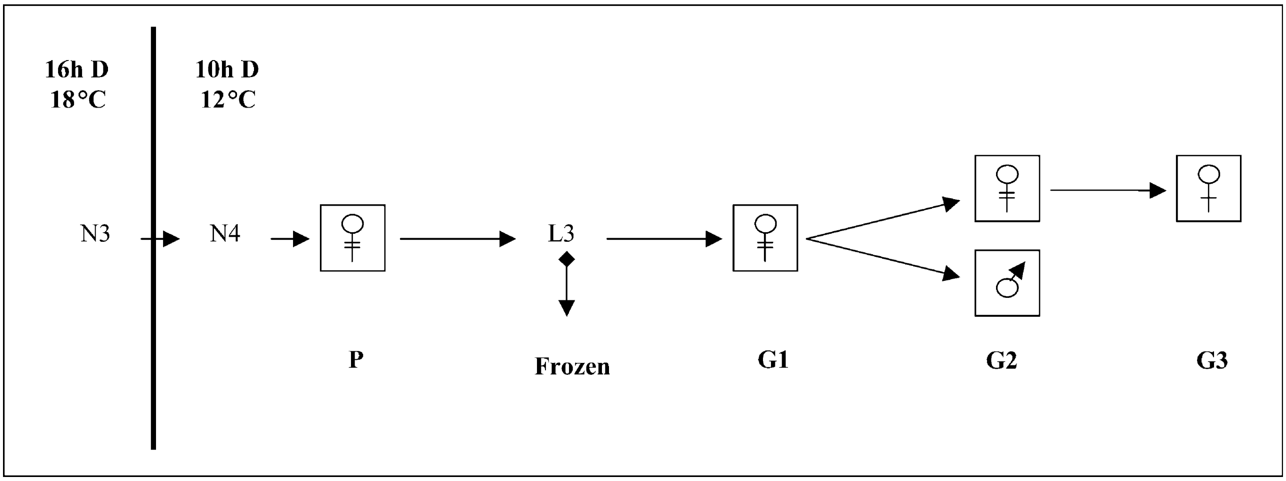

Fig. 1. Induction of sexual reproduction in *Rhopalosiphum padi*. This diagram follows the same denomination and representation as proposed by Ramos et al. (2003) for the pea aphid *Acyrthosiphon pisum* in order to facilitate comparison between the two species. Fourth instar parthenogenetic nymphs (N4) were transferred on wheat seedlings under short photoperiod and low temperature. Parthenogenetic adult (P) gave birth to first instar parthenogenetic larvae. After two moults, third instar larvae (L3) were collected and immediately frozen in liquid nitrogen. These L3 larvae corresponded to future wingless parthenogenetic females (G1) induced to produce (G2) males and winged parthenogenetic females (gynoparae) which will produce sexual females (G3). Adult stages are indicated in boxes.

Rhopalopsiphum padi, third instar larvae L3, after short-day induction), followed by the Roman number of the microplate, the letter of the row in the microplate and the Arabic number of the column in the microplate (eg: RpL3i-I-A1). Sequences have been deposited to GenBank database under the accession numbers CF799941–CF800414.

Full length sequences of 3 cDNA clones encoding proteins involved in differentiation or developmental processes were obtained sequencing both strands of the corresponding cDNA by designing internal specific primers. For each cDNA, assembly was performed by BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and alignments by Multalin (http://prodes.toulouse.inra.fr/multalin/multalin.html).

Using Phred (Ewing et al., 1998; Ewing and Green, 1998) as a base caller and cross-match (unpublished, see http://www.phrap.org), vectors, adaptors and low quality extremities of the sequences extracted from the ABI chromatograms were clipped out: 12 bases with a phred quality value under 10 on a window of 30 bp length, or stretch of >15 A, were used as a cut-off.

A total of 53 public EST and mRNA sequences from different aphid species (AF420231, AB005262, AB016720, AB039958, AB051572, AF165428, AF287291, AF411453, AF411454, AF411455, AF412814, AF412815, AF435075, AF448802, AF502081, AF502082, AF502083, AF502084, AF502397, AF527785, AF527786, AF527787, AJ489298, AJ457193, AY049740, AY162274, AY162275, AY217540, AY217542, AJ131759, AJ131760, AJ236786, AJ236787, AJ236788, AJ250348, AJ251838, AJ496197, X74554, X74555, X81887, X81888, AF233239, AF233241, AF233242, AF233243, AF233244, AF233245, AF233246, AF233247, AF233248, AF527783, AF527784, AF420231) as well as related information (library, tissue, development stage, vector...) were extracted from dbEST and EMBL. ESTs ($n = 4304$) from the brown citrus aphid *Toxoptera citricida* were extracted from Genbank's dbEST (CB814527–CB814982, CB832665–CB833296, CB854878–CB855147, CB909714–CB910020, CB936196–CB936346, CB449954–CB450759).

From all this collection, rRNA sequences, *E. coli* and yeast sequences were eliminated after identification (cross-match with a score > 100) against an invertebrate ribosomal sequence library (extracted from Genbank), or complete corresponding genomes. As aphids live in symbiosis with *Buchnera* bacteria (Baumann et al., 1995), *R. padi* ESTs were compared to the *Buchnera* sp. APS genome (http://buchnera.gsc.riken.go.jp/) (Shigenobu et al., 2000) and those presenting a cross-match with a score > 100 were eliminated.

The remaining sequences were clusterized using Biofacet (Glémet and Codani, 1997), based on a criteria of 96% similarity over 80 bps. Clusters served for contig alignment and a consensus sequence was attained using Cap3. The consensus were annotated through a NCBI-blast v2.2.6 by BlastX against SP-Trembl (version 24): only hits with an E-value $< 1.0 \times 10^{-5}$ were used for annotation. Two different contig sessions were created: one called "*R. padi*" made off *R. padi* ESTs and the 53 mRNA aphid sequences, and the second combining the "*R. padi*" contig version with the retrieved *T. citricida* ESTs.

A relational database was set up in order to store sequences and annotations, as well as links between sequences, clusters, contigs and libraries (Samson et al., 2003). To freely access these data, a web interface (http://urgi.infobiogen.fr///Projects/GPiDB/Interface/) was developed to allow users to get information on sequences, contigs, clusters (e.g. consensus, origin of its members, alignment, annotation, availability of the clone) and libraries (e.g. type, tissue, development stage). A tutorial is available on line (http://urgi.infobiogen.fr////Projects/GPiDB/Interface/gp_est_tutorial.html) which describes how to manipulate the different tools. Data can thus be viewed directly by sequence names or keywords to check for annotation, to execute personalized annotation tasks (ClustalW, Blast, Primer3), and to access to graphical visualization of the contigs (Samson et al., 2003). The whole set of sequences can be downloaded at http://urgi.infobiogen.fr///Projects/GPiDB/Interface/ for personal convenience. Annotation of the two different contig sessions was facilitated by a link to the GeneOntology (The Gene Ontology Consortium, 2001) through the AmiGO browser (http://www.godatabase.org/cgi-bin/go.cgi).

Codon frequencies were determined for sequences having a hit. First, sequences showing a frameshift (hits on different frames) have been removed from the set. Then, the putative open reading frames (ORFs) were determined as being the larger one from the first codon (ATG) to the first stop codon. Putative ORF were clustered by cellular functions (following the results of annotation, see above and Table 1) and codon usage was calculated for all the putative ORF belonging to one cellular function using Emboss cusp. Finally, codon usage was compared between cellular functions, by using Emboss codcmp.

3. Results and discussion

The *R. padi* cDNA library was constructed from total RNA extracted for whole L3 larvae produced under a short photoperiod regime (Fig. 1). About 200,000 bacterial colonies were obtained, and 1056 bacteria were individually kept as glycerol stocks. The average size of the cDNAs was 750 bp. About 30% of the amplified inserts were not sequenced either because of their small size (less than 500 bp including 478 bp of vector) or because of the presence of two amplified fragments detected after gel electrophoresis. Further analyses demonstrated that double bands were related either to the presence of two different bacterial colonies in the same glycerol stock, or to the annealing of primers within the cDNA (data not shown). A total of 726 sequences were obtained from the 5′ end of the selected cDNAs. The mean length of these sequences was 288 bp, and the median was 308 bp. The longest sequence was 604 bp. After the filtering of vector and adaptor sequences, 122 sequences shorter than 80 bp were removed (17% of the ESTs). ESTs identified as rRNA (four sequences) were also eliminated. Neither yeast nor *E. coli* contaminants were found. The endosymbiotic bacteria *Buchnera* are located in the abdominal part of aphids in a specific structure called a bacteriocyte (Baumann et al., 1995). *Buchnera* sequences (e.g. ketol-acid reductoisomerase, GTP-binding protein, porin, chaperonin) present in the *R. padi* EST collection were removed (139 sequences corresponding to 19% of the ESTs). Regions in the AT-rich bacterial DNA or RNA (probably co-extracted with the total aphid RNA) were probably able to act as a template for the olido(dT) primering during first strand cDNA synthesis. The final set of *R. padi* ESTs contained 461 sequences. The 4304 ESTs of the brown citrus aphid, *T. citricida* retrieved from dbEST (Hunter et al., 2003) were treated and edited by the same procedures: one EST corresponding to *E. coli* DNA, two ESTs from *Buchnera* and 13 ESTs of small size were eliminated. This *T. citricida* EST collection,—prepared from polyA RNA of whole aphids—was nearly not contaminated with *Buchnera* sequences, indicating that purification of mRNA might limit the risk of cloning and sequencing endosymbiont DNA.

Clusters and contigs were then produced with the 461 *R. padi* sequences plus the 53 mRNA sequences found in the public databases from various aphid species (*Acyrthosiphon pisum, Aphis fabae*, *A. gossypii*, *A. nerii* and *Myzus persicae*): this "*R. padi* contig" corresponds to version 1 of contig (see Materials and methods) at http://urgi.infobiogen.fr///Projects/GPiDB/Interface/. Among these 514 aphid sequences, 288 sequences were unique and the other 226 ESTs grouped in 73 contigs. The total of contigs was thus 361 (288 + 73). The number of ESTs in each contig ranged from 1 (288 contigs) to 13 (one contig). Fifty two sequences from the 53 extracted from public databases did not match to *R. padi* ESTs. Only one contig (CTG_RP_63.1-RpL3i-I-F1, encoding the ribosomal protein S4) grouped ESTs from *R. padi* with an aphid (*Aphis gossypii*) mRNA sequence from the public databases. Therefore, this set of ESTs represents up to 361 unique aphid contigs (or unigenes) from 514 sequences and, with one exception, is composed with sequences from one aphid species. This corresponds to a redundancy of 44% (number of ESTs in contigs/total number of ESTs) which is in the range of other insect EST collections (Mita et al., 2003). A high redundancy within the *R. padi* sequences was found for the ribosomal proteins (housekeeping proteins). The second contig (version 2 at http://urgi.infobiogen.fr///Projects/GPiDB/Interface/) put together these 514 aphid sequences and the 4288 ESTs from *T. citricida* (Hunter et al., 2003). This set formed 2457 contigs. More than half (253) of the 514 "*R. padi* contigs" did not group

Table 1
R. padi EST sequence similarities and abundance

Contig name per category	Protein homologue	Species	Accession number	Blast *E*-value	EST abundance
Chaperonin					7
CTG_RP_59.1-RpL3i-I-F10	HSC70	*Trichoplusia ni*	Q94805	8.0×10^{-38}	2
CTG_RP_100.1-RpL3i-I-F2	Stress-induced phosphoprotein 1	*Mus musculus*	Q8BPH3	4.0×10^{-29}	1
CTG_RP_177.1-RpL3i-IV-H7	90-kDa heat shock protein HSP83	*Spodoptera frugiperda*	Q9GQG6	3.0×10^{-15}	1
CTG_RP_195.1-RpL3i-IX-E6	70 kDa heat shock protein	*Bactrocera tau*	Q867Z1	1.0×10^{-30}	1
CTG_RP_337.1-RpL3i-XI-C10	Heat shock 70 kDa protein cognate	*Bombyx mori*	O76180	2.0×10^{-25}	1
CTG_RP_341.1-RpL3i-XI-D2	T-complex protein 1 gamma subunit	*Lepeophtheirus salmonis*	Q9U6Z3	2.0×10^{-29}	1
Differentiation					16
CTG_RP_43.1-RpL3i-III-D8	Chemosensory protein	*Leucophaea maderae*	Q8MTC3	6.0×10^{-23}	2
CTG_RP_75.1-RpL3i-I-A2	bgcn	*Drosophila melanogaster*	NM_166627.1	3.0×10^{-13}	1
CTG_RP_90.1-RpL3i-I-D10	CG17661 protein (programmed cell death)	*Drosophila melanogaster*	Q9VMB9	9.0×10^{-43}	1
CTG_RP_112.1-RpL3i-II-A6	CG12908 protein (epidermal growth factor)	*Drosophila melanogaster*	Q9V5J7	3.0×10^{-24}	1
CTG_RP_122.1-RpL3i-II-F11	Similar to pelota homolog	*Danio rerio*	Q7ZWC4	5.0×10^{-9}	1
CTG_RP_128.1-RpL3i-II-H8	CG17870-PF (14-3-3 protein)	*Drosophila melanogaster*	Q8MKV5	7.0×10^{-20}	1
CTG_RP_147.1-RpL3i-III-F4	Exuperantia 1	*Drosophila miranda*	Q9GNF9	6.0×10^{-20}	1
CTG_RP_157.1-RpL3i-IV-B2	Septin A	*Xenopus laevis*	Q9DE33	2.0×10^{-7}	1
CTG_RP_168.1-RpL3i-IV-F2	CG12139 protein (lipophorin receptor)	*Drosophila melanogaster*	Q9W343	2.0×10^{-51}	1
CTG_RP_184.1-RpL3i-IX-B4	SKI interacting protein	*Mus musculus*	Q9CV75	1.0×10^{-18}	1
CTG_RP_198.1-RpL3i-IX-F10	Activin receptor	*Xenopus laevis*	Q91962	1.0×10^{-13}	1
CTG_RP_258.1-RpL3i-VI-D8	COP9 signalosome complex subunit 2	*Drosophila melanogaster*	Q94899	1.0×10^{-50}	1
CTG_RP_273.1-RpL3i-VII-F1	CG2252 protein (RING)	*Drosophila melanogaster*	Q9W3L3	2.0×10^{-30}	1
CTG_RP_286.1-RpL3i-VIII-B9	CG7650 protein (phosducin)	*Drosophila melanogaster*	Q9VUR7	2.0×10^{-9}	1
CTG_RP_289.1-RpL3i-VIII-D1	Calmodulin	*Homo sapiens*	Q13942	7.0×10^{-22}	1
Metabolism					61
CTG_RP_27.1-RpL3i-IX-A1	CG9032 protein (ATP synthase)	*Drosophila melanogaster*	Q9VXN2	4.0×10^{-13}	3
CTG_RP_46.1-RpL3i-VII-E9	Ubiquitin	*Cyanidium caldarium*	Q9M3W6	1.0×10^{-34}	2
CTG_RP_47.1-RpL3i-V-B3	2-Cys thioredoxin peroxidase	*Aedes aegypti*	Q8WSF6	9.0×10^{-50}	2
CTG_RP_49.1-RpL3i-II-D2	Ubiquitin fusion protein	*Kluyveromyces lactis*	Q9Y854	5.0×10^{-27}	2
CTG_RP_52.1-RpL3i-V-F1	Putative FK506-binding protein	*Suberites domuncula*	Q966Y4	9.0×10^{-46}	2
CTG_RP_53.1-RpL3i-IX-A8	ADP/ATP translocase	*Anopheles gambiae*	Q86PG1	1.0×10^{-56}	2

(continued on next page)

Table 1 (*continued*)

Contig name per category	Protein homologue	Species	Accession number	Blast *E*-value	EST abundance
CTG_RP_60.1-RpL3i-VI-F6	ATP synthase A chain subunit 6 (EC 3.6.3.14)	*Schizaphis graminum*	Q9B6H5	3.0×10^{-8}	2
CTG_RP_76.1-RpL3i-I-A7	hnRNP	*Caenorhabditis elegans*	Q8WSM6	9.0×10^{-7}	1
CTG_RP_83.1-RpL3i-I-C1	CG1490 protein (ubiquitin protease)	*Drosophila melanogaster*	Q9VYQ8	2.0×10^{-20}	1
CTG_RP_92.1-RpL3i-I-D2	CG11015 protein (cytochrome *c* oxidase)	*Drosophila melanogaster*	Q9W5N8	1.0×10^{-24}	1
CTG_RP_97.1-RpL3i-I-E7	Cytochrome oxidase subunit I	*Uroleucon ambrosiae*	Q9B0T1	4.0×10^{-50}	1
CTG_RP_103.1-RpL3i-I-F7	Beta-glucosidase precursor	*Neotermes koshunensis*	Q8T0W7	2.0×10^{-12}	1
CTG_RP_106.1-RpL3i-I-G11	NACALPHA protein	*Drosophila melanogaster*	O16813	2.0×10^{-33}	1
CTG_RP_120.1-RpL3i-II-D4	Cytochrome *c* oxidase subunit II	*Aphis cornifoliae*	Q85JS9	1.0×10^{-33}	1
CTG_RP_137.1-RpL3i-III-C12	Dhm2 protein (5′-3′ exonuclease)	*Mus musculus*	O35651	3.0×10^{-25}	1
CTG_RP_143.1-RpL3i-III-E2	Peptidyl-prolyl *cis–trans* isomerase G precursor	*Tachypleus tridentatus*	O44073	1.0×10^{-29}	1
CTG_RP_144.1-RpL3i-III-F10	Heterogeneous nuclear ribonucleoprotein F	*Homo sapiens*	Q96AU2	8.0×10^{-13}	1
CTG_RP_152.1-RpL3i-III-G7	hnRNP protein	*Chironomus tentans*	Q23795	5.0×10^{-21}	1
CTG_RP_155.1-RpL3i-IV-A6	CG17397 protein (suppressor of RNA polymerase B)	*Drosophila melanogaster*	Q9W5P1	5.0×10^{-24}	1
CTG_RP_163.1-RpL3i-IV-E4	Inorganic pyrophosphatase	*Zygosaccharomyces bailii*	Q9C0T9	4.0×10^{-20}	1
CTG_RP_164.1-RpL3i-IV-E5	CNJB protein (DNA binding)	*Tetrahymena thermophila*	Q94821	5.0×10^{-12}	1
CTG_RP_169.1-RpL3i-IV-F6	ADP/ATP translocase	*Homo sapiens*	Q9H0C2	6.0×10^{-15}	1
CTG_RP_173.1-RpL3i-IV-G7	Ubiquitin fusion protein	*Pyrus pyrifolia*	Q9LLK2	5.0×10^{-26}	1
CTG_RP_175.1-RpL3i-IV-G9	Ribonucleotide reductase 2	*Aedes aegypti*	Q95VP8	3.0×10^{-28}	1
CTG_RP_182.1-RpL3i-IX-B11	Ahcy13 protein	*Drosophila melanogaster*	Q9VXV5	1.0×10^{-27}	1
CTG_RP_185.1-RpL3i-IX-C1	Elongation factor-2	*Scolopendra polymorpha*	Q9BNW7	1.0×10^{-33}	1
CTG_RP_189.1-RpL3i-IX-D12	Similar to dendritic cell protein	*Xenopus laevis*	Q7ZYU8	5.0×10^{-6}	1
CTG_RP_191.1-RpL3i-IX-D4	Nucleoporin 153	*Fugu rubripes*	Q9DD34	6.0×10^{-7}	1
CTG_RP_196.1-RpL3i-IX-E9	Abnormal wing disc-like protein	*Choristoneura parallela*	Q8MUR5	2.0×10^{-14}	1
CTG_RP_201.1-RpL3i-IX-F7	Proteasome	*Mus musculus*	Q8BWT0	1.0×10^{-11}	1
CTG_RP_213.1-RpL3i-V-A1	Aldehyde dehydrogenase	*Drosophila melanogaster*	O46056	3.0×10^{-22}	1
CTG_RP_217.1-RpL3i-V-B4	CG3174 protein (flavon-containing monooxygenase)	*Drosophila melanogaster*	Q9V9C9	3.0×10^{-26}	1
CTG_RP_218.1-RpL3i-V-B7	Phosphoribosylaminoimidazole carboxylase	*Danio rerio*	Q7ZUN6	2.0×10^{-38}	1
CTG_RP_226.1-RpL3i-V-D8	Translation initiation factor 3	*Drosophila melanogaster*	Q8MR84	4.0×10^{-07}	1
CTG_RP_229.1-RpL3i-V-F2	CG14214 protein (SEC 61)	*Drosophila melanogaster*	Q9VWE9	9.0×10^{-21}	1
CTG_RP_234.1-RpL3i-V-H5	ATP synthase c-subunit	*Dermacentor variabilis*	Q86G68	5.0×10^{-22}	1
CTG_RP_236.1-RpL3i-V-H9	UDP-glucuronosyltransferase 1A7	*Rattus norvegicus*	Q8VD43	6.0×10^{-24}	1

Table 1 (*continued*)

Contig name per category	Protein homologue	Species	Accession number	Blast *E*-value	EST abundance
CTG_RP_247.1-RpL3i-VI-B9	Glutamine synthetase	*Biomphalaria glabrata*	Q8IS07	1.0×10^{-27}	1
CTG_RP_254.1-RpL3i-VI-D10	CG6105 protein (ATP synthase γ subunit	*Drosophila melanogaster*	Q9VKM3	5.0×10^{-22}	1
CTG_RP_255.1-RpL3i-VI-D11	GH21728p (translation initiation factor 3)	*Drosophila melanogaster*	Q8MR49	2.0×10^{-41}	1
CTG_RP_274.1-RpL3i-VII-G4	Proliferating cell nuclear antigen	*Hyphantria cunea*	Q8MYA4	3.0×10^{-29}	1
CTG_RP_285.1-RpL3i-VIII-B7	Acyl-CoA delta-11 desaturase	*Heliothis zea*	Q9NB26	1.0×10^{-10}	1
CTG_RP_287.1-RpL3i-VIII-C11	CG5384 protein (tRNA guanine transglycosylase)	*Drosophila melanogaster*	Q9VKZ8	1.0×10^{-29}	1
CTG_RP_294.1-RpL3i-VIII-E2	dUTPase	*Mus musculus*	Q9CQ43	2.0×10^{-24}	1
CTG_RP_307.1-RpL3i-X-A4	SD19419p (uroporphyrinogen decarboxylase)	*Drosophila melanogaster*	Q8MRV9	2.0×10^{-14}	1
CTG_RP_312.1-RpL3i-X-B5	Glyceraldehyde 3-phosphate dehydrogenase	*Drosophila melanogaster*	Q8SXG8	1.0×10^{-10}	1
CTG_RP_323.1-RpL3i-X-F7	LD32039p (DNA binding protein)	*Drosophila melanogaster*	Q8MZ43	6.0×10^{-37}	1
CTG_RP_329.1-RpL3i-XI-A11	Putative Eip71CD protein	*Drosophila melanogaster*	Q9VUP4	3.0×10^{-14}	1
CTG_RP_336.1-RpL3i-XI-B9	Pyruvate carboxylase	*Aedes aegypti*	Q16921	1.0×10^{-23}	1
CTG_RP_343.1-RpL3i-XI-D8	CG7006 protein (ribosome biogenesis)	*Drosophila melanogaster*	Q9VC28	1.0×10^{-18}	1
CTG_RP_346.1-RpL3i-XI-E5	H+-ATPase subunit	*Sus scrofa*	Q9T2U6	2.0×10^{-9}	1
CTG_RP_355.1-RpL3i-XI-G7	Cytochrome *b*	*Siphateles bicolor*	Q85U34	2.0×10^{-12}	1
CTG_RP_356.1-RpL3i-XI-G8	SMT3 protein	*Drosophila melanogaster*	O97102	3.0×10^{-15}	1
Ribosomal proteins					122
CTG_RP_4.1-RpL3i-I-E4	Ribosomal protein L17/23	*Spodoptera frugiperda*	Q962Y9	2.0×10^{-63}	5
CTG_RP_5.1-RpL3i-I-D1	Ribosomal protein P2	*Ceratitis capitata*	O96934	8.0×10^{-16}	5
CTG_RP_6.1-RpL3i-I-H1	RE05022p (Ribosomal protein L71)	*Drosophila melanogaster*	Q9W3Z2	2.0×10^{-34}	5
CTG_RP_7.1-RpL3i-III-D5	Ribosomal protein L26	*Spodoptera frugiperda*	Q962T4	3.0×10^{-54}	5
CTG_RP_8.1-RpL3i-III-C1	QM protein (Ribosomal protein L10)	*Heliothis virescens*	Q95PD4	2.0×10^{-72}	5
CTG_RP_9.1-RpL3i-I-E6	CG7424 protein (Ribosomal protein L44)	*Drosophila melanogaster*	Q9VLT7	8.0×10^{-40}	5
CTG_RP_12.1-RpL3i-III-B1	Ribosomal protein S9	*Drosophila melanogaster*	Q9VT06	1.0×10^{-78}	4
CTG_RP_14.1-RpL3i-I-C4	Ribosomal protein L14	*Spodoptera frugiperda*	Q962T9	1.0×10^{-34}	4
CTG_RP_16.1-RpL3i-II-B8	Ribosomal protein L34	*Spodoptera frugiperda*	Q8WQI6	3.0×10^{-41}	4
CTG_RP_17.1-RpL3i-III-G10	CG7283 protein (ribosomal protein L10A)	*Drosophila melanogaster*	Q9VTP4	5.0×10^{-63}	4
CTG_RP_20.1-RpL3i-IV-D9	CG5827 protein (ribosomal protein)	*Drosophila melanogaster*	Q9VMU4	1.0×10^{-31}	3
CTG_RP_21.1-RpL3i-I-A11	Similar to ribosomal protein S8	*Bos taurus*	Q862P8	1.0×10^{-47}	3
CTG_RP_22.1-RpL3i-IX-F2	Ribosomal protein S21	*Spodoptera frugiperda*	Q962Q8	1.0×10^{-28}	3

(*continued on next page*)

Table 1 (*continued*)

Contig name per category	Protein homologue	Species	Accession number	Blast *E*-value	EST abundance
CTG_RP_24.1-RpL3i-II-E12	Ribosomal protein L8	*Spodoptera frugiperda*	Q95V39	2.0×10^{-75}	3
CTG_RP_25.1-RpL3i-IV-F7	Ribosomal protein L18	*Branchiostoma lanceolatum*	Q86LX2	3.0×10^{-42}	3
CTG_RP_29.1-RpL3i-III-D2	Ribosomal protein S23	*Spodoptera frugiperda*	Q962Q7	6.0×10^{-50}	3
CTG_RP_30.1-RpL3i-I-D5	CG2998 protein (ribosomal protein S28)	*Drosophila melanogaster*	Q9W334	1.0×10^{-20}	3
CTG_RP_31.1-RpL3i-III-D12	Ribosomal protein L29	*Spodoptera frugiperda*	Q95V37	6.0×10^{-18}	3
CTG_RP_34.1-RpL3i-IV-B10	Ribosomal protein S3	*Spodoptera frugiperda*	Q95V36	1.0×10^{-77}	3
CTG_RP_36.1-RpL3i-III-C11	Ribosomal protein L11	*Petromyzon marinus*	Q801H8	7.0×10^{-24}	3
CTG_RP_45.1-RpL3i-IV-A12	60S ribosomal protein L15	*Spodoptera frugiperda*	Q8I9V9	5.0×10^{-43}	2
CTG_RP_54.1-RpL3i-II-E5	Ribosomal protein L30	*Aequipecten irradians*	Q8ITC5	2.0×10^{-12}	2
CTG_RP_55.1-RpL3i-II-A9	Ribosomal protein L35A	*Spodoptera frugiperda*	Q962S9	1.0×10^{-36}	2
CTG_RP_57.1-RpL3i-III-B3	Ribosomal protein L32	*Spodoptera frugiperda*	Q962T1	5.0×10^{-58}	2
CTG_RP_58.1-RpL3i-I-B12	Ribosomal protein S15	*Spodoptera frugiperda*	Q962R4	7.0×10^{-50}	2
CTG_RP_63.1-RpL3i-I-F1	RpS4 protein	*Drosophila melanogaster*	Q9VU44	4.0×10^{-30}	2
CTG_RP_67.1-RpL3i-I-E12	CG3751 protein (ribosomal protein S24)	*Drosophila melanogaster*	Q9W229	3.0×10^{-33}	2
CTG_RP_68.1-RpL3i-I-E1	Ribosomal protein L36A	*Spodoptera frugiperda*	Q962S8	5.0×10^{-28}	2
CTG_RP_70.1-RpL3i-III-C7	Ribosomal protein S26	*Spodoptera frugiperda*	Q962Q4	1.0×10^{-50}	2
CTG_RP_77.1-RpL3i-I-A8	Ribosomal protein S16	*Spodoptera frugiperda*	Q95V31	3.0×10^{-26}	1
CTG_RP_78.1-RpL3i-I-A9	Ribosomal protein P0	*Aedes albopictus*	Q8MQT0	2.0×10^{-45}	1
CTG_RP_79.1-RpL3i-I-B3	Ribosomal protein S19	*Spodoptera frugiperda*	Q962R0	3.0×10^{-42}	1
CTG_RP_80.1-RpL3i-I-B5	Similar to ribosomal protein	*Xenopus laevis*	Q7ZYR6	5.0×10^{-52}	1
CTG_RP_85.1-RpL3i-I-C3	Ribosomal protein L13	*Xenopus laevis*	Q8AVQ1	4.0×10^{-15}	1
CTG_RP_96.1-RpL3i-I-E5	Ribosomal protein S10	*Branchiostoma belcheri*	Q86QR8	2.0×10^{-41}	1
CTG_RP_111.1-RpL3i-II-A2	RE28824p (ribosomal protein L12)	*Drosophila melanogaster*	Q9W1B9	5.0×10^{-38}	1
CTG_RP_121.1-RpL3i-II-E2	Ribosomal protein S18	*Branchiostoma belcheri*	Q8ISP0	3.0×10^{-9}	1
CTG_RP_150.1-RpL3i-III-G12	Ribosomal protein S26	*Spodoptera frugiperda*	Q962Q4	3.0×10^{-6}	1
CTG_RP_153.1-RpL3i-IV-A3	Ribosomal protein S5	*Spodoptera frugiperda*	Q95V33	7.0×10^{-34}	1
CTG_RP_156.1-RpL3i-IV-A9	Ribosomal protein L18a	*Xenopus laevis*	Q7ZYQ8	2.0×10^{-42}	1
CTG_RP_158.1-RpL3i-IV-B9	CG12740 protein (ribosomal protein L28)	*Drosophila melanogaster*	Q9VZS4	3.0×10^{-16}	1
CTG_RP_165.1-RpL3i-IV-E7	Ribosomal protein S19	*Aequipecten irradians*	Q8ITC3	3.0×10^{-14}	1
CTG_RP_176.1-RpL3i-IV-H2	Ribosomal protein S6	*Mus musculus*	Q8BT09	2.0×10^{-28}	1
CTG_RP_178.1-RpL3i-IV-H8	Ribosomal protein L38	*Spodoptera frugiperda*	Q962S5	2.0×10^{-22}	1

Table 1 (*continued*)

Contig name per category	Protein homologue	Species	Accession number	Blast *E*-value	EST abundance
CTG_RP_197.1-RpL3i-IX-F1	Similar to ribosomal protein L17	*Xenopus laevis*	Q7ZY53	2.0×10^{-22}	1
CTG_RP_237.1-RpL3i-VI-A10	60S ribosomal protein L27	*Drosophila melanogaster*	Q9VBN5	2.0×10^{-54}	1
CTG_RP_240.1-RpL3i-VI-A6	CG4046 protein (ribosomal protein S16)	*Drosophila melanogaster*	Q9W237	4.0×10^{-37}	1
CTG_RP_252.1-RpL3i-VI-C7	Ribosomal protein L44	*Chlamys farreri*	Q8MUE4	8.0×10^{-14}	1
CTG_RP_260.1-RpL3i-VI-F10	Ribosomal protein L7Ae-like	*Mus musculus*	Q9D0T1	4.0×10^{-39}	1
CTG_RP_278.1-RpL3i-VII-H10	Ribosomal protein S15A	*Spodoptera frugiperda*	Q962R3	2.0×10^{-26}	1
CTG_RP_281.1-RpL3i-VII-H5	Ribosomal protein L19	*Ictalurus punctatus*	Q90YU8	9.0×10^{-27}	1
CTG_RP_304.1-RpL3i-VIII-H7	Ribosomal protein L31	*Heliothis virescens*	Q9GP16	3.0×10^{-19}	1
CTG_RP_320.1-RpL3i-X-E8	Ribosomal protein L9	*Branchiostoma belcheri*	Q8ISP7	2.0×10^{-24}	1
CTG_RP_313.1-RpL3i-X-B9	60S acidic ribosomal protein P0	*Spodoptera frugiperda*	Q8WQJ2	4.0×10^{-40}	1
CTG_RP_327.1-RpL3i-X-H12	Ribosomal protein S11	*Heliothis virescens*	Q95P67	3.0×10^{-30}	1
CTG_RP_328.1-RpL3i-X-H5	Ribosomal protein S18	*Spodoptera frugiperda*	Q962R1	3.0×10^{-21}	1
CTG_RP_330.1-RpL3i-XI-A2	Ribosomal protein S18	*Spodoptera frugiperda*	Q962R1	2.0×10^{-28}	1
Structure					27
CTG_RP_3.1-RpL3i-II-B12	Histone H1	*Rhynchosciara americana*	Q963G2	3.0×10^{-20}	6
CTG_RP_19.1-RpL3i-I-F9	Alpha-4-tubulin	*Gecarcinus lateralis*	O01944	4.0×10^{-95}	3
CTG_RP_33.1-RpL3i-II-H3	Troponin T	*Periplaneta americana*	Q9XZ71	7.0×10^{-6}	3
CTG_RP_51.1-RpL3i-V-B9	Troponin I-like protein	*Haemaphysalis longicornis*	Q969A1	9.0×10^{-21}	2
CTG_RP_86.1-RpL3i-I-C5	Actin 1	*Culicoides sonorensis*	Q8WRE6	8.0×10^{-83}	1
CTG_RP_88.1-RpL3i-I-C7	RH04334p (Articulin)	*Drosophila melanogaster*	Q8SZM2	8.0×10^{-17}	1
CTG_RP_94.1-RpL3i-I-D7	DNM1 protein (dynamin)	*Homo sapiens*	Q86VD2	3.0×10^{-36}	1
CTG_RP_105.1-RpL3i-I-G10	Putative cytoskeletal actin	*Ciona intestinalis*	Q8I7E1	2.0×10^{-51}	1
CTG_RP_115.1-RpL3i-II-B6	Putative HMG-like protein	*Dermacentor variabilis*	Q86G70	8.0×10^{-15}	1
CTG_RP_136.1-RpL3i-III-B9	RE12057p (actin)	*Drosophila melanogaster*	Q8MZ23	5.0×10^{-27}	1
CTG_RP_145.1-RpL3i-III-F2	CG11274 (serine/arginine repetitive matrix 1 protein)	*Drosophila melanogaster*	Q9VU43	5.0×10^{-43}	1
CTG_RP_199.1-RpL3i-IX-F11	Histone H1	*Tigriopus californicus*	P92139	2.0×10^{-8}	1
CTG_RP_221.1-RpL3i-V-C8	Alpha-tubulin	*Spirometra erinaceieuropaei*	Q9NL73	2.0×10^{-53}	1
CTG_RP_228.1-RpL3i-V-F12	Histone H2	*Arabidopsis thaliana*	Q8GUH3	1.0×10^{-33}	1
CTG_RP_279.1-RpL3i-VII-H3	Actin	*Aplysia californica*	Q16942	2.0×10^{-18}	1
CTG_RP_309.1-RpL3i-X-A9	Actin 1	*Culicoides sonorensis*	Q8WRE6	6.0×10^{-59}	1
CTG_RP_319.1-RpL3i-X-D5	Tropomyosin	*Lepisma saccharina*	Q8T379	8.0×10^{-24}	1

with *T. citricida* contigs. This contig version is no further analysed in order to eliminate the risk of working on in silico-defined chimeric sequences from different aphid species.

After annotation, contigs from the "*R. padi* contig" were grouped into functional categories (Table 1). Gene Ontology annotation was also performed (Table 2) showing a high representation for "physiological process", "cell", "binding" and "catalytic activity" terms.

A cellular role could be assigned for 44% of the *R. padi* contigs (Table 1) on the basis of sequence similarity to proteins with known function in public databases using BlastX with an E-value $\leq 10^{-5}$. The remaining 56% fall into the "hypothetical" (6%) or "no hit" (50%) categories (Fig. 2) and the most abundant sequence (CTG_RP_1.1-RpL3i-I-D6 harbouring 13 ESTs) belongs to the "no hit" category. This proportion is similar to that observed for other insect EST projects as for *S. frugiperda* (Landais et al., 2003), *B. mori* (Mita et al., 2003) or *T. citricida* (Hunter et al., 2003). The most likely reason for this lack of similarity is that some of these sequences are probably too short or may be constituted by 5′- or 3′-unstranslated regions. A second possibility is that these partial sequences correspond to non-conserved domains of polypeptides: a longer sequence should allow a better identification of these ESTs. Finally, some of these proteins with unknown functions might correspond to aphid specific cellular functions not yet elucidated. Domazet-Loso and Tautz (2003) proposed that genes involved in environmental adaptation evolve quickly and might correspond to orphan sequences. As aphids are highly sensitive to environmental changes, it could be that many of these orphan genes correspond to sequences rapidly evolving. The largest proportion of functionally assigned sequences fall into three functional categories: metabolism (16%), ribosomal proteins (17%), and structure (5%). Of note is the presence of chaperonins (2%) as well as a high representation (38%) of house keeping genes involved in general cellular functions (categories 1, 2 and 3). The large proportion of housekeeping genes is lower than the one found for *S. frugiperda* EST collection which was constructed from cultured cells (Landais et al., 2003): the function of these cultured cells—even if they originated from ovarian cells—are likely to have most of their highly expressed genes directed mainly towards general metabolism, gene expression and cell division. The high proportion of housekeeping genes in our collection probably results from the use of whole-insects for preparing the cDNA library. Housekeeping genes being ubiquitously expressed within cells, their predominance may reduce the chance to identify other genes expressed within a narrow range of cells and/or following a specific challenge. A large number of the housekeeping *R. padi* ESTs belongs to the ribosomal protein gene family (34% of the housekeeping proteins and 17% of total ESTs) directly involved in mRNA translation and protein synthesis. The number (57 clusters, Table 1) of the different ribosomal proteins found in our collection might represent a large proportion of the 80 predictive ribosomal protein complement (Landais et al., 2003). These genes are known to be highly expressed and present in most cell types.

Some genes (4%) involved in specific developmental events were identified. Full length sequences were obtained for three of these inserts: *pelota*, *bgcn* and *exuperantia* because there are all involved in early steps of oocyte differentiation or embryo development in *Drosophila* (see below). None of these inserts corresponded to full length cDNA sequences. *R. padi bgcn*, *pelota* and *exuperantia* deduced amino acids sequences showed a high similarity to *Drosophila* orthologous (76%, 89% and 47%, respectively) as well as to vertebrate sequences (except for *exuperantia*) (Fig. 3). During aphid parthenogenesis, in contrast with sexual differentiation, oocytes are arrested in the first meiotic division and undergo embryogenesis at a 2*n* stage without fertilization by spermatozoids (Blackman, 1987). Very little is known about the molecular events involved in this process, but the regulation of the meiotic cycle is probably a key step before the bifurcation

Table 2
Gene Ontology annotation of the *R. padi* contig

Gene Ontology	Number (%) of contigs
Biological process	72 (100)
Behavior	0 (0)
Unknown	7 (9.7)
Cellular process	25 (34.7)
Development	13 (18.0)
Physiological process	27 (37.5)
Cellular component	37 (100)
Cell	27 (73)
Unknown	10 (27)
Extracellular	0 (0)
Unlocalized	0 (0)
Molecular function	50 (100)
Antioxydant activity	0 (0)
Apoptosis regulator activity	0 (0)
Binding	14 (28)
Catalytic activity	13 (26)
Cell adhesion molecule activity	0 (0)
Chaperone activity	3 (6)
Defense/immunity protein activity	0 (0)
Enzyme regulator activity	0 (0)
Unknown	3 (6)
Motor activity	0 (0)
Protein tagging activity	1 (2)
Signal transducer activity	5 (10)
Structural molecule activity	1 (2)
Transcription regulator activity	2 (4)
Translation regulator activity	0 (0)
Transporter activity	8 (16)

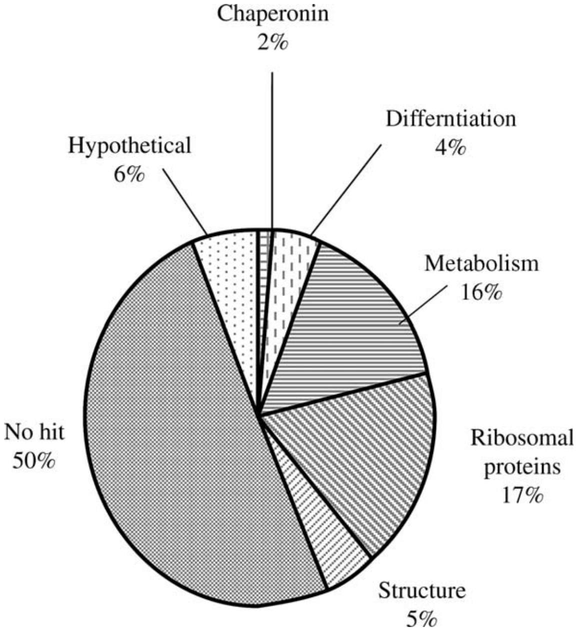

Fig. 2. Distribution by functional categories of 361 transcripts forming the "*R. padi* contig".

towards the two developmental orientations. The *pelota*-related gene (CTG_RP_122.1-RpL3i-II-F11) was identified that may regulate early divisions of aphid oocytes since in *Drosophila* mutations in *pelota* genes not only caused defects in spermatogenesis but also affected mitotic divisions during the development of ovaries (Adham et al., 2003). Contig CTG_RP_75.1-RpL3i-I-A2 was similar to a cystoblast differentiation factor (*begnin gonial cell neoplasm*) of *Drosophila melanogaster*. This gene is involved in the regulation of the asymmetric division of germline stem cells giving the cystoblast from which oocytes will later differentiate (Ohlstein et al., 2000). The *exuperantia* gene (CTG_RP_147.1-RpL3i-III-F4) is involved in the prelocalization of a small set of maternally expressed genes in *D. melanogaster* embryos (Macdonald et al., 1991). The presence of these two later genes in *R. padi* suggests regulation mechanisms for oocyte and embryos development involving syncytial structure and mRNA localization. It is noteworthy that we did not find in our EST catalog the *ApSDI-1* gene described by Ramos et al. (2003): it encodes an amino-acid transporter in neurons and is expressed in short-days reared pea aphids (*Acyrthosiphon pisum*). This gene is probably lowly expressed and a substractive approach was necessary to identify it in *A. pisum*.

In order to more precisely describe the abundant "no hit" sequences of the "*R. padi* contig", we performed a BlastX comparison on the orphan contigs with the filter off, to eventually detect sequences of low complexity which might have been lost in the original filtered BlastX. Additional ($n = 23$) contigs were found to have similarity to known polypeptides (Table 3) which are rich however in repeat domains.

A Student test comparing the mean length of the 189 sequences showing a hit with the not hit 172 sequences indicated that no hit sequences were smaller (risk α of 1×10^{-32}). GC contents were also compared between sequences with and without a hit: no hit sequences were richer in GC than sequences showing a hit ($p < 0.001$, χ^2 test), indicating the probable presence of non-translated regions. The high GC content observed for sequences having no hit can also be an indicator of either a 5′-untranslated region or a non-protein coding RNA. Post-genomic approaches should help in understanding the role of these proteins in aphid biology.

Codon bias was often detected in *Drosophila* genes. We analysed by gene categories (as defined in Table 1) the codon usage of the *R. padi* sequences having a hit. A χ^2 test showed that all protein categories except the "metabolism" class have a codon usage significantly different from the average codon usage (i.e. pooling all the sequences from all categories). A codon usage bias has been described in *D. melanogaster* for ribosomal proteins. Our preliminary analysis also indicated a codon usage bias for ribosomal proteins of *R. padi*. However, our collection of total *R. padi* sequences is still limited for that kind of studies as it contains no full-lengths sequences. Furthermore, the cDNA clones of *R. padi* were randomly picked from the library clones and probably corresponded (because of their small quantity: 1056) mainly to highly expressed genes; our sample is thus probably biased. A larger collection of ESTs and a precise analysis gene by gene (including a comparison with orthologs from other insects) will be necessary to show whether some aphid genes present significant codon usage bias.

EST sequences can be a source of molecular markers particularly for microsatellites sequences. At least two microsatellite motifs have been detected in CTG_RP_28.1-RpL3i-III-G9 and CTG_RP_194.1-RpL3i-IX-E4, corresponding to (TGG)9 and (TTA)7 repeats, respectively (Fig. 4). One small and one imperfect repeats were also detected in CTG_RP_146.1-RpL3i-III-F3 [(TTA)4] and CTG_RP_261.1-RpL3i-XI-F3 [(TGG)3 found twice separated by 77 bases], respectively (Fig. 4). Three microsatellite repeats (CTG_RP_28.1-RpL3i-III-G9, CTG_RP_194.1-RpL3i-IX-E4, CTG_RP_350.1-RpL3i-XI-F3) were found in putative ORFs while CTG_RP_146.1-RpL3i-III-F3 was probably within the 3′-unstranslated region. None of these contigs correspond to known ORFs.

In conclusion, ESTs provide a valuable resource for gene discovery related to the regulation of reproduction mode in aphids. The next step will be the use of these ESTs for cDNA array-based technologies and comparative hybridizations in order to identify aphid

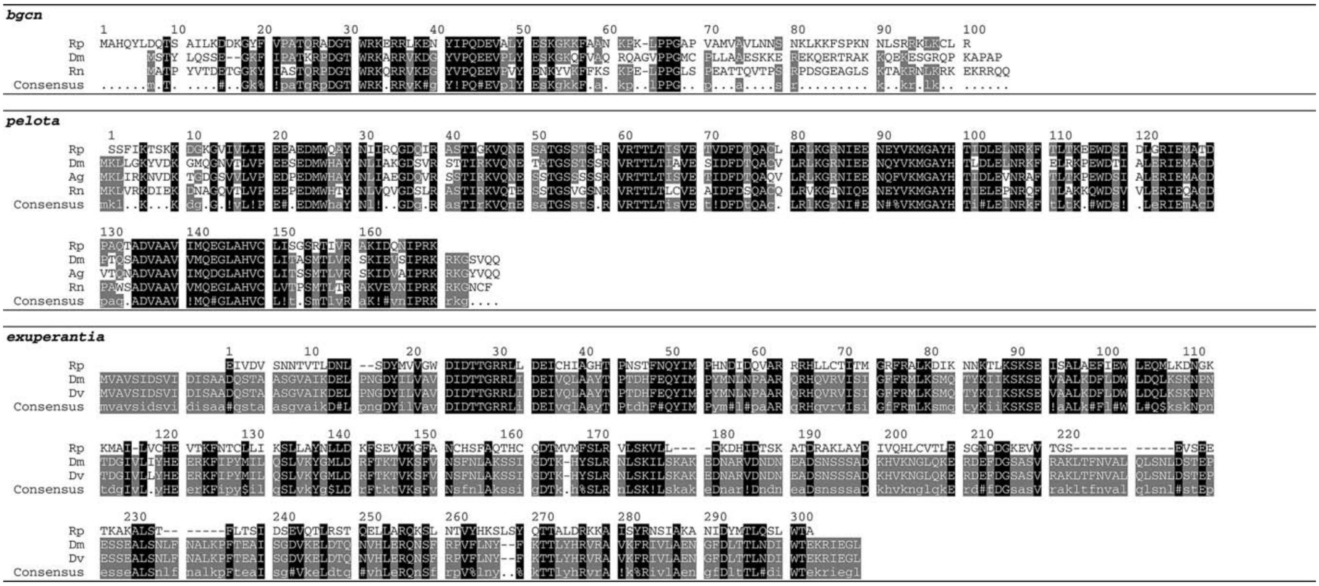

Fig. 3. Comparison of the predicted amino acid sequences of three *R. padi* cDNA inserts with other animal sequences. The *R. padi* sequences are shown on the upper lines in single letter code. The predicted amino acid sequences of *Anopheles gambiae* (Ag), *Drosophila melanogaster* (Dm), *Drosophila viridis* (Dm) and *Rattus norvegicus* (Rn) are shown below in optimal alignment (Corpet, 1998). Common amino acids are shown in black (present in all compared sequences) or grey (present in the majority of the sequences). Gaps are indicated (-).

genes involved in the switch between the development of asexual and sexual morphs. The present data represent a first step towards the identification and annotation of transcript complements in aphids.

Table 3
R. padi contigs sequence similarities from the "no hit" category after a BLASTX, filter off

Contig name	Protein homologue	Species	Accession number	*E*-value	Feature
CTG_RP_1.1-RpL3i-I-D6	Ribosomal protein L41	*Drosophila melanogaster*	Q962S2	7×10^{-7}	R/K rich
CTG_RP_15.1-RpL3i-III-C8	Putative VrrB	*Bacilus subtilis*	Q9KI89	3×10^{-19}	H rich
CTG_RP_18.1-RpL3i-IX-G10	Hypothetical protein	*Anopheles gambiae*	Q7QEH2	8×10^{-7}	none
CTG_RP_26.1-RpL3i-I-A12	Histone H1	*Triticum aestivum*	P27806	8×10^{-12}	P rich
CTG_RP_28.1-RpL3i-III-G9	Cold and drought-regulated protein CORA	*Medicago sativa*	Q07202	5×10^{-15}	G rich
CTG_RP_35.1-RpL3i-I-D11	Hypothetical protein	*Caenorhabditis elegans*	Q20013	2×10^{-9}	none
CTG_RP_56.1-RpL3i-VI-H8	Peptidyl-prolyl *cis–trans* isomerase	*Spodoptera frugiperda*	Q26486	3×10^{-10}	none
CTG_RP_132.1-RpL3i-III-A9	Mucin 5	*Homo sapiens*	Q8WWQ4	7×10^{-6}	none
CTG_RP_134.1-RpL3i-III-B12	Hypothetical protein	*Anopheles gambiae*	Q7PNM6	8×10^{-8}	none
CTG_RP_160.1-RpL3i-IV-D12	Orotidine 5′-phosphate decarboxylase	*Buchnera aphidicola*	Q8K9Q1	2×10^{-12}	none
CTG_RP_172.1-RpL3i-IV-G5	Hypothetical protein	*Anopheles gambiae*	Q7QII3	2×10^{-13}	Q rich
CTG_RP_174.1-RpL3i-IV-G8	DNA mismatch repair protein mutL	*Buchnera aphidicola*	Q8K913	7×10^{-19}	K rich
CTG_RP_179.1-RpL3i-IX-A10	PmbA protein homolog	*Buchnera aphidicola*	Q8KA30	5×10^{-17}	none
CTG_RP_206.1-RpL3i-IX-G5	Hypothetical protein	*Danio rerio*	Q7ZU84	4×10^{-16}	E/D rich
CTG_RP_225.1-RpL3i-V-D5	90-kDa heat shock protein HSP83	*Spodoptera frugiperda*	Q9GQG6	8×10^{-11}	none
CTG_RP_248.1-RpL3i-VI-C1	Trigger factor (TF)	*Buchnera aphidicola*	Q8K991	2×10^{-16}	K rich
CTG_RP_266.1-RpL3i-VI-H1	Hypothetical protein	*Drosophila melanogaster*	Q9W1R2	4×10^{-16}	G rich
CTG_RP_293.1-RpL3i-VIII-E11	DNA gyrase subunit A	*Buchnera aphidicola*	Q8K9W2	2×10^{-14}	none
CTG_RP_314.1-RpL3i-X-C12	Nucleoplasmin-like protein	*Drosophila melanogaster*	Q27415	6×10^{-6}	E rich
CTG_RP_326.1-RpL3i-X-G5	Ribosomal protein S3A	*Spodoptera frugiperda*	Q95V35	2×10^{-6}	K rich
CTG_RP_334.1-RpL3i-XI-B11	Hypothetical protein	*Drosophila melanogaster*	Q9VNX6	1×10^{-7}	G rich
CTG_RP_339.1-RpL3i-XI-D1	Chorion protein s18 precursor	*Ceratitis capitata*	Q9NFX7	5×10^{-8}	A/S rich
CTG_RP_354.1-RpL3i-XI-G5	Probable nucleoporin Nup54	*Drosophila melanogaster*	Q9V6B9	2×10^{-7}	Q rich

CTG_RP_28.1-RpL3i-III-G9

GGGGACAGTATACATCACACTTTCAAAGACGAATATTCAGTAGTCTCAAGAATCATGAAATCTTACACAGCGATA
TTAGCTTTGTGTTTTGTCGTTCTCGTAATGACTCAACAAGCTACTTCAGAACCAGCTCCCGAACCACGCAAGAAT
GA**TGGTGGTGGTGGTGGTGGTGGTGG**AAATCACCATGGAACTGGTGGAAATTATGGAAATGGAAATT
ACGGACATGGACATGGTGGACACGGTGGACACGGTGGCAACGGACATAATACACACGGAAGTAGCGGACATGGAC
AACACGGACACGGAGGTGGCGGACACGGACACGGAGGACATCATTAAATTTAAAACAATTTCATTTTCATCCAGT
ACAATAATATAAAAAAAAACTATTATTTTTTTTCTTATCAATCTATGTATTATCTTTAATGCAAAATCCCGAATA
AACAATAATTTCTCATGGAAATTTTCACCTA

CTG_RP_194.1-RpL3i-IX-E4

GGGGATATCGT**TTATTATTATTATTATTATTA**TAAATTTTTTTTCGTTTCGTGTTTTCCGTTTTCGGCGTC
CACCGCCGTTCGCCGCCAACGGCTGTCGCGTCGCCTTACGACAACGATAACGACGACGACAACAACAACAACGGC
GATAAACCACAACAAACCGACAACGACGGTGCGCGTGTTGTTCCGAGTACGACCGCACGGCATTCGCTGCTCCGT
CCGTCCGTCGCGGTGTCCGCGTCCGATGGGGATATCAACGTGCAC

CTG_RP_146.1-RpL3i-III-F3

GGGGACTGCATGGACCCAGAAGATGAAGACGGAATGATCCCTTATGTCCCGTTCGTAAAAAAAGTGATGGTTCGC
CTGAGCGACATGAAGTAAATTCCATTTTAAGCGTTTAAAGTTAAAAAAAAAACATTAAAAAATTTAAACTACCCG
GACCATTATCCCGTTTCTGGCCAAAGTTTGCGGAAAGCCCGTACCGGCCGACGACAACCCATTCAACAAGTAAAA
CTACACGCGCATCGCTCCGACAAAAGCAAATACACGTNATTATTTATTTA**TTATTATTATTA**ATATAATTATT
ATAATTGCAAAAAAAAAAAA

CTG_RP_350.1-RpL3i-XI-F3

GGGGTTACAATGTTTACCCGAACCGCCGGACAGTTGGCTGGCCAGTTGTTTACGTAAATCGATCTTGCGCGACAC
CTGCGGCGG**TGGTGGTGG**CGGCTCCGCGATTGACGGAGTGACGACGATGGTGGCGGCGGGCATGGAGGTGGTG
GTTACGTTGGTTGTAGCAGTAG**TGGTGGTGG**TGTTGTTGTACGTTGACATATGGGCCTCCTCGACTTTGACAG
AACGGTTTAACATTTTCATAGAATGTTCTTCCATCCATCTCTATAAAAAAATATCAAACGAAATATTGTGTTTAC
TTTAAAA

Fig. 4. Microsatellite sequences found in four contigs of *R. padi*. Microsatellites are underlined in bold.

4. Acknowledgments

We thank Gérard Devauchelle (INRA, St Christol-les-Alès, France) for giving us access to unpublished data on *S. frugiperda* ESTs, Wayne Hunter (USDA, Fort Pierce, Florida) for sharing EST data and Claude Rispe (INRA Rennes, UMR BiO3P) for reading the manuscript. This work has been supported by INRA ("Santé des Plantes et Environnement" Department) and an INRA postdoctoral grant to B. Sabater-Muñoz. Morgan Perennou, Erwan Corre and Antoine Lostanlen (CNRS Roscoff, OUEST-Genopole®, France) are acknowledged for their advice in EST sequencing. We thank Julien Elie for his technical help in PCR amplification.

References

Adham, I.M., Sallam, M.A., Steding, G., Korabiowska, M., Brinck, U., Hoyer-Fender, S., Oh, C., Engel, W., 2003. Disruption of the *pelota* gene causes early embryonic lethality and defects in cell cycle progression. Molecular Cell Biology 23, 1470–1476.

Baumann, P., Baumann, L., Lai, C.Y., Rouhbakhsh, D., Moran, N.A., Clark, M.A., 1995. Genetics, physiology and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. Annual Review Microbiology 49, 55–94.

Blackman, R.L., 1987. Reproduction, cytogenetics and development. In: Minks, A.K., Harrewijn, A.P. (Eds.), Aphids, their biology, natural enemies and control, 2A. Elsevier, Amsterdam, Oxford, New York, Tokyo, pp. 163–195.

Corpet, F., 1998. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research 16, 10881–10890.

Dixon, A.F.G., 1998. Aphid ecology: an optimization approach, second ed. Chapman & Hall, London.

Domazet-Loso, T., Tautz, D., 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Research 13, 2213–2219.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8, 175–185.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8, 186–194.

Glémet, E., Codani, J.J., 1997. LASSAP, a Large Scale Sequence Comparison Package (note: LASSAP is now Biofacet™). Computational Applied Bioscience 13, 137–143.

Gray, S., Gildow, F.E., 2003. Luteovirus-aphid interactions. Annual Review Phytopathology 41, 539–566.

Hardie, J., Nunes, M.V., 2001. Aphid photoperiodic clocks. Journal Insect Physiology 47, 821–832.

Hunter, W.B., Dang, P.M., Bausher, M.G., Chaparro, J.X., McKendree, W., Shatters, R.G., McKenzie, C.L., Sinisterra, X.H., 2003. Aphid biology: expressed genes from the alate *Toxoptera citricida*, the brown citrus aphid. Journal Insect Science 3, 23.

Landais, I., Ogliastro, M., Mita, K., Nohata, J., Lopez-Ferber, M., Duonor-Cérutti, M., Shimada, T., Fournier, P., Devauchelle, G., 2003. Annotation pattern of ESTs from *Spodoptera* SF9 cells and analysis of the ribosomal protein genes reveal insect-specific features and unexpectedly low codon usage bias. Bioinformatics 19, 2343–2350.

Macdonald, P.M., Luk, S.K., Kilpatrick, M., 1991. Protein enriched by the *exuperantia* gene is concentrated at sites of *bicoid* mRNA accumulation in *Drosophila* nurse cells but not in oocytes or embryos. Genes Development 5, 2455–2466.

Mita, K., Morimyo, M., Okano, K., Koike, Y., Nohata, J., Kawasaki, H., Kadono-Okuda, K., Yamamoto, K., Suzuki, M.G., Shimada, T., Goldsmith, M.R., Maeda, S., 2003. The construction of an EST database for *Bombyx mori* and its application. Proceedings National Academy Science USA 100, 14121–14126.

Ohlstein, B., Lavoie, C.A., Vef, O., Gateff, E., McKearin, D.M., 2000. The *Drosophila* cystoblast differentiation factor, benign genial cell neoplasm, is related to DExH-box proteins and interacts genetically with bag-of-marbles. Genetics 155, 1809–1819.

Ramos, S., Moya, A., Martinez-Torres, D., 2003. Identification of a gene overexpressed in aphids reared under short photoperiod. Insect Biochemistry Molecular Biology 33, 289–298.

Samson, D., Legeai, F., Karsenty, E., Reboux, S., Veyrieras, J.B., Just, J., Barillot, E., 2003. GenoPlante-Info (GPI): a collection of databases and bioinformatics resources for plant genomics. Nucleic Acids Research 31, 179–182.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H., 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 407, 81–86.

Simon, J.-C., Dedryver, C.A., Pierre, J.S., 1991. Identifying bird cherry-oat aphid *Rhopalosiphum padi* emigrants, alate exules and gynoparae: application of multivariate methods to morphometric and anatomical features. Entomologia Experimenta Applicata 59, 267–277.

The Gene Ontology Consortium, 2001. Creating the gene ontology resource: design and implementation. Genome Research 11, 1425–1433.