Insect Biochemistry and Molecular Biology 46 (2014) 31–42

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

Comparative analysis of two phenologically divergent populations of the pine processionary moth (*Thaumetopoea pityocampa*) by *de novo* transcriptome sequencing

Bernhard Gschloessl [a,*], Heiko Vogel [b], Christian Burban [c], David Heckel [b], Réjane Streiff [a], Carole Kerdelhué [a]

[a] INRA, UMR CBGP (INRA/IRD/CIRAD/Montpellier Supagro), Campus International de Baillarguet, CS30016, F-34988 Montferrier-sur-Lez Cedex, France
[b] Max Planck Institute for Chemical Ecology, Department of Entomology, 07745 Jena, Germany
[c] INRA, UMR1202 BIOGECO, 69 Route d'Arcachon, F-33612 Cestas Cedex, France

ARTICLE INFO

Article history:
Received 10 September 2013
Received in revised form
11 January 2014
Accepted 13 January 2014

Keywords:
Pine processionary moth
Thaumetopoea pityocampa
de novo transcriptome assembly
Next Generation Sequencing
Population comparative analyses
SNP detection
Phenology

ABSTRACT

The pine processionary moth *Thaumetopoea pityocampa* is a Mediterranean lepidopteran defoliator that experiences a rapid range expansion towards higher latitudes and altitudes due to the current climate warming. Its phenology – the time of sexual reproduction – is certainly a key trait for the local adaptation of the processionary moth to climatic conditions. Moreover, an exceptional case of allochronic differentiation was discovered *ca.* 15 years ago in this species. A population with a shifted phenology (the summer population, SP) co-exists near Leiria, Portugal, with a population following the classical cycle (the winter population, WP). The existence of this population is an outstanding opportunity to decipher the genetic bases of phenology. No genomic resources were so far available for *T. pityocampa*. We developed a high-throughput sequencing approach to build a first reference transcriptome, and to proceed with comparative analyses of the sympatric SP and WP. We pooled RNA extracted from whole individuals of various developmental stages, and performed a transcriptome characterisation for both populations combining Roche 454-FLX and traditional Sanger data. The obtained sequences were clustered into *ca.* 12,000 transcripts corresponding to 9265 unigenes. The mean transcript coverage was 21.9 reads per bp. Almost 70% of the *de novo* assembled transcripts displayed significant similarity to previously published proteins and around 50% of the transcripts contained a full-length coding region. Comparative analyses of the population transcriptomes allowed to investigate genes specifically expressed in one of the studied populations only, and to identify the most divergent homologous SP/WP transcripts. The most divergent pairs of transcripts did not correspond to obvious phenology-related candidate genes, and 43% could not be functionally annotated. This study provides the first comprehensive genome-wide resource for the target species *T. pityocampa*. Many of the assembled genes are orthologs of published Lepidoptera genes, which allows carrying out gene-specific re-sequencing. Data mining has allowed the identification of SNP loci that will be useful for population genomic approaches and genome-wide scans of population differentiation to identify signatures of selection.

1. Introduction

The pine processionary moth (hereafter, PPM) *Thaumetopoea pityocampa* (Lepidoptera, Notodontidae) is an insect pest occurring over the Mediterranean basin and the Atlantic coasts of France, Spain and Portugal (Kerdelhué et al., 2009). It causes considerable damage to pinewoods over its distribution range, and its gregarious, urticating larvae are responsible for severe public and animal health concern (Rodríguez-Mahillo et al., 2012; Vega et al., 2004). Its distribution range is in part driven by winter temperatures, as larval development mainly occurs during the coldest months.

Abbreviations: PPM, pine processionary moth; SP, summer population; WP, winter population; SNP, single nucleotide polymorphism; NGS, Next Generation Sequencing; SFF, standard flowgram file; NR, non-redundant NCBI protein database; GO, Gene Ontology; CDS, coding sequence; OHR, Ortholog Hit Ratio; UniProtKB, Uniprot knowledgebase; RBH, reciprocal best hit; Indel, insertion or deletion; UTR, untranslated region; N50, contig length for which half of the assembly is represented by contigs of this size or longer.

* Corresponding author. Tel.: +33 4 30 63 04 19; fax: +33 4 99 62 33 45.
E-mail address: Bernhard.Gschloessl@supagro.inra.fr (B. Gschloessl).

0965-1748/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ibmb.2014.01.005

Consistently, the PPM has been shown to expand northward and towards higher altitudes due to the current climate warming (Battisti et al., 2005).

This species typically has one generation per year, although prolonged pupal diapause can delay adult emergence by one to several years. Adults emerge in summer, mate and lay eggs in the following hours or days. After *ca.* 30 days of embryonic development, larvae hatch and develop in autumn and winter. The caterpillars spin characteristic silk nests where they gather during daytime, while they feed on pine needles at night. At the end of larval development, usually between January and March, the colony leaves the nest in a typical head-to-tail procession in search for an underground pupation site. After an obligate nymphal diapause, adults emerge the following summer. Local phenological variations are supposed to be adaptive responses in the moth populations and allow the species to occur under various environmental conditions: reproduction tends to take place in early summer in regions where winters are harsh, and the first, susceptible larval stages thus develop before the first frost; in contrast, reproduction takes place in late summer in regions with hot summers and mild winters, thereby avoiding larval mortality due to high temperatures (Huchon and Démolin, 1970). Phenology is thus a major trait involved in local adaptation of the PPM.

Interestingly, a population showing a shifted phenology was discovered in Portugal *ca.* 15 years ago in a coastal pine forest named Mata Nacional de Leiria. In this very peculiar population, reproduction occurs in spring and larvae develop all over the summer. Individuals with the classical life cycle co-occur in the same forest. Due to its summer larval development, the shifted population has been called the "summer population" (SP) while the sympatric population exhibiting a classical cycle is referred to as the "winter population" (WP) (Pimentel et al., 2006). The SP larvae never face winter conditions, and consequently do not spin nests, although they still show a gregarious behaviour. On the contrary, the most susceptible larval stages experience very high temperatures that are expected to be lethal in that species (Huchon and Démolin, 1970). Experimental approaches have shown that the first larval instars of the SP survive significantly better under high temperatures than the sympatric WP larvae (Santos et al., 2011b), suggesting a physiological adaptation. SP is phylogenetically very close to the sympatric WP according to mitochondrial and ITS sequences while microsatellite data suggest that current gene flow is very restricted between both populations (Santos et al., 2011a, 2007). This unique situation corresponds to a plausible recent allochronic differentiation, where gene flow is hampered by a shift in time of the reproductive period. The PPM found in the Leiria pine forest provides an unprecedented opportunity to study the genetic bases of phenology and adaptation to high temperatures.

We present here a *de novo* transcriptome sequencing approach to study and compare genes expressed in the SP and WP occurring in Leiria. We focussed on the late developmental stages (last larval instar, pupae and adults of both sexes) in which the genes involved in phenology (here, mostly the regulation of pupal diapause) and in reproduction are likely to be expressed. Data combine low- and high-throughput sequencing technologies (Sanger and 454 sequencing, respectively). The goal of the present study was fourfold: (i) build a *de novo* reference transcriptome for *T. pityocampa*, and significantly increase at the same time the genomic resources for this insect pest, which is phylogenetically distant from most studied Lepidoptera species (Mutanen et al., 2010); (ii) identify gene-targeted single nucleotide polymorphisms (SNP) for future genome wide analyses of diversity and differentiation; (iii) identify the sets of genes specifically expressed (or absent) in the shifted SP; (iv) identify the most divergent homologous genes between the SP and WP at the nucleotide level. These two latter points are the first steps towards the comprehension of the genetic architecture of phenology, *i.e.* of the trait responsible for the allochronic differentiation occurring in the Leiria forest and a major trait in PPM local adaptation.

2. Material and methods

2.1. Laboratory protocols

2.1.1. Sampling, RNA purification and isolation

All samples were initially collected in the field in the Mata Nacional de Leiria, Portugal (39°47′N 8°58′W). Larvae were sampled about one month after the L4 to L5 molt, *i.e.* at the end of the last larval instar, while still aggregated in the nest. Pupae were sampled about two months after the procession and about 4–5 months before adult emergence. Sampling of the adults took place one to two days after emergence; virgin females were collected, while the males had possibly mated. Concerning the SP, pupae were sampled between November 15th and December 17th 2007. Ten individuals were deep frozen at −80 °C immediately after field collection, while the other pupae were kept under laboratory conditions at the Instituto Superior de Agronomia of the University of Lisbon until adult emergence. Ten individuals were then frozen as adults in April 2008. L5 larvae were sampled in the field in September 2008, and immediately frozen at −80 °C. Concerning the WP, L5 larvae were collected in Leiria forest in January 2008. Ten of those were immediately frozen at −80 °C, while the others were kept in the laboratory at the Instituto Superior de Agronomia and fed with *Pinus pinaster* branches until pupation. Ten pupae were frozen from this rearing in July and ten adults in August 2008. Samples were sent in dry ice to the Max Planck Institute for Chemical Ecology (Jena, Germany) to proceed with RNA isolation and sequencing procedures. Total RNA was isolated from five L5 larvae (which could not be sexed), five pupae (males and females) and five adults (males and females) per population. The other individuals were kept frozen in Lisbon to ensure that the whole procedure could be repeated.

TRIzol Reagent (Invitrogen) was used to isolate the RNA according to the manufacturer's protocol. The RNA was precipitated overnight at −20 °C and the dried pellet was dissolved in 90 μl RNA Storage Solution (Ambion). An additional DNAse (Turbo DNAse, Ambion) treatment was then applied prior to the second purification step to eliminate any contaminating DNA. The DNAse enzyme was removed and the RNA was further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20 μl of RNA Storage Solution (Ambion). RNA integrity and quantity was verified on an Agilent 2100 Bioanalyser using the RNA Nano chips (Agilent Technologies, Palo Alto, CA), using both the high resolution gel and electropherogram views provided, even though the standard RIN method is not applicable to insect RNA because the 28S band tends to break (Winnebeck et al., 2010). RNA quantity was determined on a NanoDrop ND-1000 spectrophotometer. For each developmental stage of each *T. pityocampa* population two different high-quality RNA extractions were generated. Equal amounts of total RNA of all stages were subsequently pooled for each population (leading to one final RNA pool per population, including L5 larvae, pupae and adults).

2.1.2. Normalisation and construction of the cDNA library

For each of the RNA pools, a full-length enriched, normalised cDNA library was generated using a combination of the SMART cDNA library construction kit (Clontech) and the Trimmer Direct cDNA normalisation kit (Evrogen). We generally followed the manufacturer's protocol except for some important modifications, as described in Vogel et al. (2010). In brief, 2 μg of total RNA was

used for each cDNA library. Reverse transcription was performed with a mixture of several reverse transcription enzymes for 60 min at 42 °C and 90 min at 50 °C. Each step of the normalisation procedure was carefully monitored to avoid the generation of artefacts and over-cycling, following the protocol detailed in Vogel and Wheat (2011). The optimal condition for double-stranded cDNA synthesis was empirically determined by subjecting the cDNA to a range of cycle numbers and their products checked by electrophoresis. The optimal cycle number was defined as the maximum number of PCR cycles without any signs of over-cycling. Optimisation of the complete cDNA normalisation procedure was essentially performed as described in Vogel and Wheat (2011). Each cDNA library was then purified and concentrated using the DNA Clean and Concentrator kit (Zymogen) and size fractionated with SizeSep 400 spin columns (GE Healthcare) that resulted in a cut-off at about 150 base pairs (bp). Normalised cDNAs were then split in two independent lots, one for each sequencing procedure (*i.e.* Sanger and 454, see below).

2.1.3. Sanger sequencing and data cleaning

The full-length enriched cDNAs were cut with *Sfi*I and ligated to pDNR-LIB plasmid (Clontech). Ligations were transformed into *Escherichia coli* ELECTROMAX DH5α-E electro-competent cells (Invitrogen). Plasmid mini-preparation from bacterial colonies grown in 96 deep-well plates was performed using the 96-well robot plasmid isolation kit (NextTec) on a Tecan EVO Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5′ termini of directionally cloned cDNA libraries was carried out on an ABI 3730xl automatic DNA sequencer (PE Applied Biosystems) in the genomics group of the Entomology department at the Max Planck Institute for Chemical Ecology, Jena (Germany). A total of 10,994 Sanger sequences was generated (see Table 1). By assembling the Sanger reads of each population into contigs (data not shown) we verified that the normalisation procedure was efficient and that low redundancy occurred. Vector clipping and quality trimming using stringent conditions (*e.g.* high quality sequence trimming parameters) was done with the Lasergene software package (DNAStar Inc.). Initial BLAST searches to exclude potential contaminations in the normalisation and cloning procedure were conducted on a local server against the National Center for Biotechnology Information (NCBI) bacteria and fungi databases using BLASTX with an E-value cut-off of 1e-20 and the UniVec database using custom Perl scripts.

Table 1
Sequencing and assembly features of the PPM, SP and WP transcriptomes.

	PPM	SP	WP
Raw number of 454 reads	873,281	467,082	406,199
Number of 454 reads after cleaning	870,176	465,703	404,473
Mean 454 trimmed read length [bp] (min–max)	319.07 (220–675)	307.68 (220–593)	332.18 (240–675)
Number of Sanger reads	10,994	5290	5704
Mean Sanger trimmed read length [bp] (min–max)	550 (334–1904)	515 (334–1832)	583 (382–1904)
Number of assembled reads	594,072	309,200	247,247
Number of singletons	99,602	67,550	69,035
Size of assembled transcriptome [Mbp]	14.9	9.4	10.1
Coverage (mean read count per bp)	21.9	17.1	14.0
Number of unigenes	9265	6696	7141
Number of transcripts[a]	12,011	8119	8524
Mean transcript length [bp]	1239	1156	1189
Median transcript length [bp]	1048	1001	1042
N50 transcript length [bp][b]	1517	1386	1421
Number of exons	13,627	9092	9623
Mean exon count per transcript	1.3	1.3	1.3
Mean exon length [bp][c]	915	889	895
Median exon length [bp][c]	754	758	775

[a] Including alternative transcripts.
[b] N50: contig length for which half of the assembly is represented by contigs of this size or longer.
[c] Each contig of an alternative transcript was only counted once as exon.

2.1.4. Next Generation Sequencing and data cleaning

The two cDNA libraries resulting from the normalisation procedure were also subjected to Next Generation Sequencing (NGS) with the Roche 454 FLX platform using Titanium chemistry at GATC Biotech (Konstanz, Germany). Normalised cDNA material was purified on DNA Clean & Concentrator columns (Zymo Research, Irvine, USA) for direct sequencing. For each of the *T. pityocampa* populations half a microtiter plate was run. The short read sequences were retrieved from the standard flowgram file (sff) provided by the sequencing centre using the application *sff_extract* (version 0.2.8, http://bioinf.comav.upv.es/sff_extract, accessed 09/04/2013, default options). Subsequently, we ran an in-house sequence filtering pipeline on the 454 short read sequences which consisted in the following consecutive steps: 5′ and 3′ 454 adapter sequences (provided by the sequencing centre) were searched for by *cross_match* (Phrap/Cross_match/Swat package, http://www.phrap.org, accessed on 09/04/2013). Then, eventual bacterial contaminations were detected in the 454 reads by searching with cross_match a set of ribosomal RNA sequences (obtained from the NCBI EST and Nucleotide databases). SeqClean (http://sourceforge.net/projects/seqclean/, accessed on 09/04/2013) clipped poly-A/poly-T tails and those sequence parts which were marked by cross_match with Xs (adapters and contaminations). Finally, TrimSeq (EMBOSS package version 6.3.1, Rice et al., 2000) was applied to remove Xs and undetermined nucleotides (Ns) which were not cleaned by the former programs at both sequence ends.

2.2. Bioinformatics data analyses

2.2.1. Assembly of PPM and population-specific transcriptomes

We assembled three *de novo* transcriptomes from different subsets of the data. The “PPM reference transcriptome” was built with the sequence reads obtained from both populations, whereas the “WP-” and “SP-transcriptomes” were constructed by assembling only the population-specific sequences. All three assemblies combined 454 short reads and Sanger sequences, and were done with Newbler (version 2.5.3, Roche), as recommended in Kumar and Blaxter (2010). This programme reconstitutes alternative transcripts; a so-called “contig” represents an exon, while an “isotig” describes a transcript eventually composed of several contigs connected by bridging sequence reads. In some cases though, a single exon can represent a complete transcript. Alternative transcripts (or splice variants) are isotigs with at least one common contig. Finally, an “isogroup” comprises all alternative transcripts that share a common contig subset. As a consequence, an isogroup should represent, in theory, one unigene.

Assembly parameter settings for all three assemblies were: (i) a maximum of 500 contigs per isogroup, (ii) a minimum contig length of 100 bp and (iii) a maximum of 100 isotigs per isogroup. We also applied the following options: “-cdna” for transcriptome assembly, “-notrim” to disable additional quality filtering and “-het” to take into account the sequence variability due to the pooling of heterozygous individuals. Finally, reads were clustered if they had a minimum overlap of 40 bp (“-ml” option) and a minimum sequence identity of 97% (“-mi” option).

In fine, the PPM-, SP- and WP-transcriptomes contained both Newbler isotigs and large contigs (>500 bp) not belonging to any

isotig. The reads that were not assembled (*i.e.* not included in any contig) formed the set of singletons. They were used for functional annotation analyses, but they were not included in comparative analyses. The raw sequence reads and the assembled transcripts were deposited as NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject/232643).

2.2.2. Functional annotation

The assembled transcripts were automatically annotated by BLASTX against the Non-Redundant NCBI protein database (hereafter NR database, version of 5th March 2013). Up to twenty best NR hits per transcript were retained, with an E-value cut-off of 1e-5 and a minimum sequence alignment similarity of 55%.

Gene Ontology (GO, Ashburner et al., 2000) Slim terms were assigned to the NR-annotated transcripts using a local Blast2GO pipeline *b2g4pipe* (Conesa et al., 2005, version 2.5.0) with access to a local GO MySQL database (version of April 2013). The application Annex (Myhre et al., 2006) was used to further optimize the GO term identification by crossing the three GO categories (biological process, molecular function and cellular component) in order to search for name similarities, GO term relationships (indicated in gene annotations submitted by users) and enzyme relationships within metabolic pathways (Kyoto Encyclopedia of Genes and Genomes, Kanehisa and Goto, 2000).

2.2.3. Assembly completeness

The occurrence and completeness of coding sequences (CDS) within transcripts were estimated with FrameDP (Gouzy et al., 2009, version 1.2.0) using the default parameters. FrameDP uses probabilistic interpolated Markov models taking into account protein similarities, eventual frame shifts and the occurrence of start and stop codons to determine whether a transcript is likely to have a CDS ("coding/non-coding potential"). We enabled the FrameDP self-training procedure, using the SwissProt database of April 2013 as reference database. FrameDP then created a training set of transcripts showing significant identity (through BLASTX searches) against the reference database (see Gouzy et al., 2009 for more details).

Besides, we used a slightly modified version of the Ortholog Hit Ratio (OHR) developed by O'Neil et al. (2010) to estimate the completeness of *de novo* transcriptome assemblies. The OHR was originally defined as the ratio of the number of bases in the BLAST hit region of a *de novo* transcript to the length of the best match silkworm (*Bombyx mori*) protein; it therefore gives a close estimate of the completeness of the putative coding region of the *de novo* transcripts compared to a Lepidoptera model. In the present study, we extended the OHR definition by comparing each annotated assembled transcript to its very best BLAST hit with several Lepidoptera protein sequence databases rather than to *B. mori* only: the Lepidoptera reference protein set ($n = 102{,}216$) used here included 14,623 *B. mori* predicted proteins (SilkDB version 2.0: consensus gene set merged by GLEAN, Duan et al., 2010), 70,688 butterfly protein sequences from ButterflyBase (namely ButterflyBase_pro.fsa.gz, version 08/20/2007, http://butterflybase.ice.mpg.de/datasets.php), 651 Lepidoptera sequences downloaded from SwissProt *via* the Sequence Retrieval System at the European Bioinformatics Institute (srs.ebi.ac.uk, downloaded the 01/10/2013) and 16,254 proteins of *Danaus plexippus* (version 1.0.20, 08/09/2013, ftp://ftp.ensemblgenomes.org/pub/metazoa/release-20/fasta/danaus_plexippus/pep/). We then used as ratio the length of the ungapped alignment of the transcript sequence divided by the total sequence length of the very best BLAST hit (BLASTX, E-value 1e-5). Hence, an OHR of 1 indicated that the assembled transcript covered the entire hit sequence, while an OHR >1 (respectively, <1) indicated that the *de novo* transcript was longer (respectively, shorter) than the most similar reference sequence.

2.2.4. Phenology-related set of candidate genes

We established a reference set of proteins encoded by genes potentially involved in insect phenology (*i.e.* a set of phenology-related candidate genes). In order to select genes for which function and role in insect life cycle was experimentally confirmed in the literature, we searched within the NCBI PubMed database (http://www.ncbi.nlm.nih.gov/pubmed, accessed on 09/04/2013) for scientific manuscripts using the keywords "phenology", "circadian clock" or "clock", and restricting the search procedure to Insecta. The corresponding protein sequences were then downloaded from the Uniprot knowledgebase (UniProtKB, December 2012) using either gene or function names ("Gene Name" or "Description"). The reference list was completed by a UniProtKB search within Insecta for proteins with function names "diapause", "clock" or "circadian". For both datasets, if Lepidoptera sequences were available for a given protein, we only kept these. In total, the complete phenology-related set of candidate genes contained 1307 reference protein sequences (1159 and 148 protein sequences found by the two searches, respectively; Additional file 1), which included redundant information, *i.e.* different sequences of the same gene from a wide range of taxa. However, no clustering analysis was performed to reduce this redundancy since we intended to avoid merging paralogous sequences (or splitting orthologous sequences) in this set. This list was used to identify genes potentially involved in the allochronic differentiation between the SP and WP within the transcripts identified below as differentially expressed or divergent between the studied populations (see "comparative analyses of the SP and WP transcriptomes" below). Any assembled transcript was identified as a phenology-related candidate gene if it corresponded to a reciprocal best hit (RBH), by bi-directional BLAST analyses (BLASTX and TBLASTN, E-value threshold of 1e-5) against the phenology-related set of proteins defined above.

2.3. Comparative analyses of the SP and WP transcriptomes

2.3.1. Population-specific PPM transcripts and unigenes

Each PPM transcript was subjected to a BLAST search (BLASTN, E-value 1e-5) against the SP- and WP-assembled transcriptomes. PPM transcripts which had significant similarity with only one of the SP or WP data sets were considered as "private transcripts" (*i.e.* SP- or WP-specific). As a second step, we assigned all PPM transcripts to their respective unigenes, and we thereby identified "private unigenes", *i.e.* PPM unigenes comprising transcripts that had a significant homology with either only the SP- or only the WP-data. To avoid artefacts due to low coverage and sequencing errors, we further restricted the results to the private unigenes having a minimal average coverage of five reads/bp, as recommended by Gilles et al. (2011).

Enrichment analyses were performed to test whether private transcripts uncovered GO terms categories enriched in either the SP or WP, with a two-tailed Fisher's Exact Test in the Blast2GO web application (http://www.blast2go.com/webstart/makeJnlp.php, accessed on 09/04/2013, with access to the local GO database). A false discovery rate (FDR) correction for multiple testing was applied (Benjamini and Hochberg, 1995).

2.3.2. Divergence patterns between homologous SP- and WP-transcripts

In order to identify the homologous transcripts present in both populations and to avoid retrieving paralogs, we used stringent criteria for the search. Bi-directional BLASTN analyses were thus

conducted between the SP- and the WP-transcriptomes, and homologs were defined as the RBH with a minimum E-value cut-off of 1e-10 and an alignment sequence overlap over 90%. The sequence identity between SP and WP homologs was retrieved from the BLAST outputs.

2.3.3. Detection of SNPs

SNPs were identified within and between populations by mapping the 454 and Sanger reads of each population against the set of assembled exon sequences found in the PPM reference transcriptome. Mapping was done with the 64-bit SSAHA program (Ning et al., 2001 version 2.5.5), which was specifically developed to map long read sequences against a reference sequence set. The SAMtools (Li et al., 2009) application *mpileup* and the PoPoolation2 toolkit (Kofler et al., 2011a, 2011b) were used to estimate genotype likelihoods and to proceed with SNP calling. Strict criteria were applied to avoid false positive SNPs: (i) only reads with a good sequencing quality (PHRED-scaled mapping quality value $\geq$20) were considered; (ii) SNPs were only defined in regions with a minimum read depth of ten reads/bp in each population, while positions with a coverage of 100 and more reads were ignored, because they were likely located in duplicated regions (SAMtools manual, http://samtools.sourceforge.net/mpileup.shtml, accessed on 09/04/2013); (iii) an allele variant was only called when there were at least two corresponding reads within a same population; (iv) regions containing insertions and deletions (indels) were also filtered, as alignment programs often fail to align short read sequences in such regions (Harismendy et al., 2009; Krawitz et al., 2010). Moreover, the five neighbouring base pairs on both sides of indels were also excluded.

We defined three SNP categories. The first category included the SNPs that were polymorphic within both the SP and WP ("within-population SNPs"). The second category corresponded to SNPs that were polymorphic in one population and monomorphic in the other one ("SP- or WP-specific SNPs"). The third category included the "diagnostic SNPs", *i.e.* positions where the base was fixed within each population and differed between the SP and WP.

Finally, with the aid of an in-house Perl script we characterized the position of each SNP within the genic region (*i.e.* CDS, 5′ or 3′ untranslated regions (UTR) or introns (in case of pre-mRNAs)). We further determined if the polymorphisms located within CDS regions were silent (synonymous) or non-silent (non-synonymous), by applying the EMBOSS *transeq* tool for sequence translation (EMBOSS package version 6.3.1, Rice et al., 2000).

3. Results and discussion

3.1. The PPM reference transcriptome: assembly quality and functional annotation

3.1.1. Sequencing and assembly characteristics

We obtained 467,082 short reads for the SP and 406,199 for the WP from the Roche 454 sequencing. After sequence cleaning, we retained 465,703 short reads for the SP and 404,473 for the WP. Mean read lengths were 308 bp and 332 bp for the SP and WP, respectively (Table 1). This set was completed by 5290 SP and 5704 WP Sanger long read sequences, which were on average 515 bp long for the SP and 583 bp long for the WP.

After assembly of the entire data set using Newbler, we obtained 13,627 exons, which were assigned to 12,011 transcripts and 9265 unigenes (Table 1). Each unigene had on average 1.3 alternative transcripts, which themselves were constituted on average of 1.3 exons. The mean transcript length was 1239 bp (N50 = 1517 bp), while the mean exon length was 915 bp. The average transcript coverage was 22 reads per bp, which is close to the results obtained for reference transcriptomes of other non-model insect species (*e.g.* Ewen-Campen et al., 2011). 99,602 short reads (11.4%) could not be assembled and remained as singletons. As expected, the statistics obtained for the SP- and WP-specific assemblies were slightly below the values obtained for the PPM reference transcriptome in terms of number of transcripts, numbers of unigenes and coverage, while the contigs and transcripts lengths were similar (Table 1).

3.1.2. Low impact of homopolymers on the assemblies

When using 454 technology, sequencing errors in homopolymer regions can hamper a proper DNA assembly (Gilles et al., 2011; Huse et al., 2007). In the present work, the three *de novo* transcriptomes fulfilled the minimum criterion of five reads/bp proposed by Gilles et al. (2011), as the average coverage per transcript ranged between 14 and 22. Moreover, for each data set (PPM, SP and WP), we further analysed the lengths of homopolymers both in the set of cleaned sequence reads and in the assembled transcripts for each nucleotide (A, C, G, T). The mean lengths of the longest homopolymers were below 2.7 in all cases. The results presented here were thus probably not affected by the occurrence of artificial homopolymers.

3.1.3. Number of unigenes

In Eukaryotes, it has been suggested that genome size is not linearly related to the actual number of genes (Gregory, 2005; Hou and Lin, 2009). In Lepidoptera, genome sizes vary from *ca.* 280 Mbp for *D. plexippus* (Zhan et al., 2011) to 1.9 Gbp for *Euchlaena irraria* (Gregory and Hebert, 2003; Gregory et al., 2007). Yet, the estimated numbers of genes range from *ca.* 14,000 in *Lymantria dispar* (Sparks and Gundersen-Rindal, 2011) or *B. mori* (Duan et al., 2010) to *ca.* 18,000 in *Plutella xylostella* (You et al., 2013). Lower numbers of unigenes were found in some species, *e.g.* 9033 unigenes in the Noctuid moth *Spodoptera littoralis* (Legeai et al., 2011) or 6593 unigenes in the Sphingid *Manducta sexta* (Pauchet et al., 2010). In both cases though, these low numbers of genes were likely due to single-tissue sampling (antennae and midgut, respectively). Based on these observations, the number of protein-coding genes in the PPM could range from 12,000 to 18,000, which suggests that the reference transcriptome assembled in the present study for *T. pityocampa* contains 50%–77% of the expected total number of genes. These somehow low proportions are likely due to a sampling effect, as we obtained RNA from some but not all the developmental stages, from a limited number of individuals per stage, and not from samples exposed to extreme biotic or abiotic conditions. The number of sequenced genes will be enhanced in future studies by including embryos and early instar larvae. Moreover, transcriptome assembly is not optimal when using heterozygous individuals; yet, rearing of the PPM is challenging and hampers the production of inbred lines.

3.1.4. Functional annotation

In the PPM reference transcriptome, 69.6% (n = 8361) of the transcripts and 26.5% (n = 26,344) of the singletons had counterparts in NR. A majority of the best similarity hits to the assembled transcripts (86.3%) found in NR corresponded to Lepidoptera sequences. More, the number of PPM orthologs found for a given reference species mostly reflected the number of entries that occurred in NR for that species. When taking into account only the species annotation of the very best NR similarity hit, as many as 4844 PPM transcripts blasted with *D. plexippus* sequences, *vs.* 823 with *Bombyx* spp., 611 with *Papilio* spp., 190 with *Helicoverpa* spp., 185 with *Spodoptera* spp. and 112 with *M. sexta* (Additional file 2).

Yet, 41.3% (n = 3457) of these NR annotations did not exhibit actual functional information, as they corresponded to

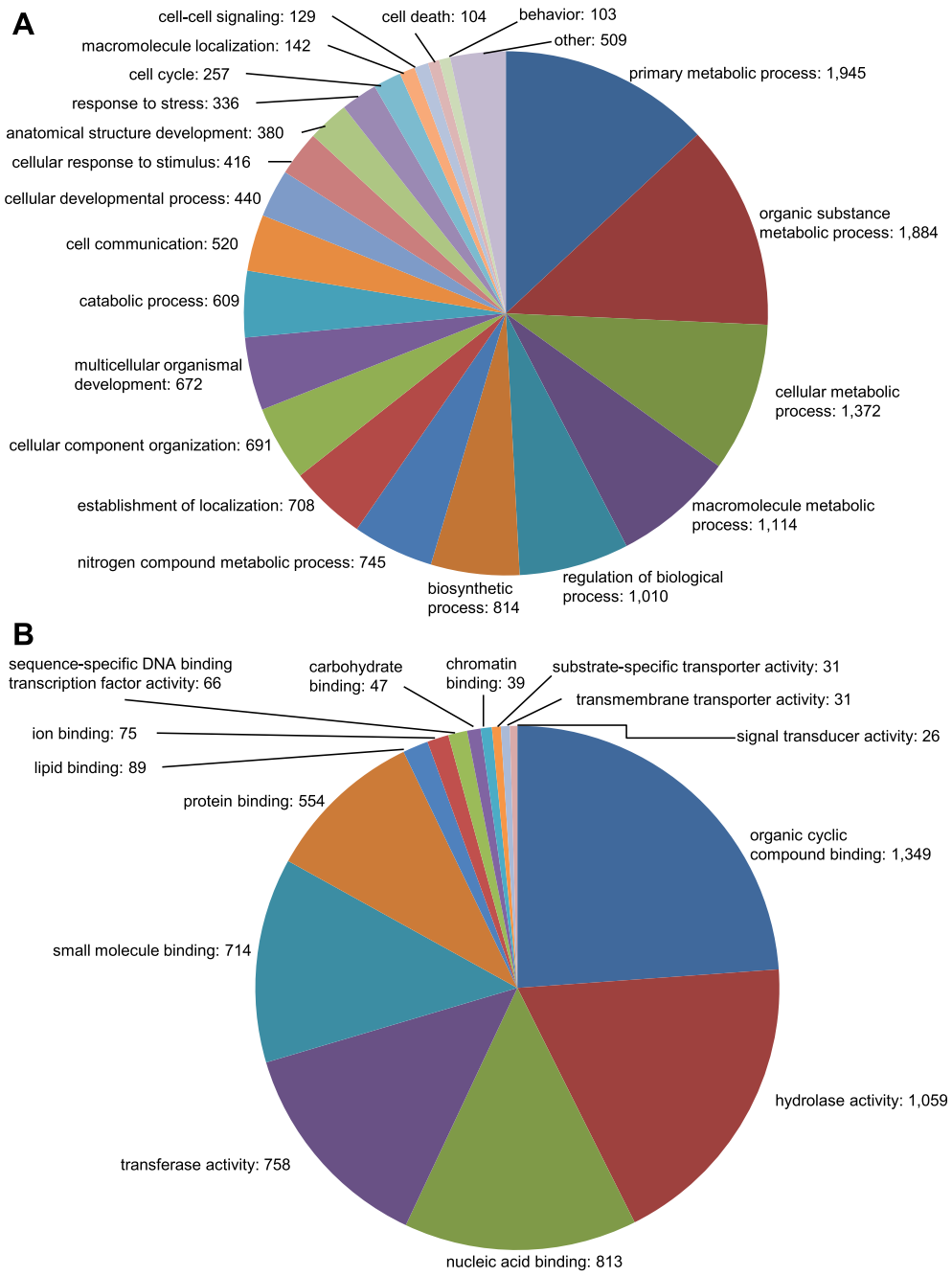

Fig. 1. Gene Ontology (GO) assignments for the PPM transcriptome. Level 3 GO assignments as predicted for their involvement in (A) biological process and (B) molecular function. The numbers of transcripts assigned to each GO term are given. In part A, the term "other" comprises: response to external stimulus (84), regulation of biological quality (73), cellular homeostasis (73), secondary metabolic process (71), response to abiotic stimulus (67), response to biotic stimulus (52), response to endogenous stimulus (36), cell growth (21), cell recognition (20), interspecies interaction between organisms (9), microtubule-based process (1), cellular localisation (1), cellular component movement (1).

"hypothetical protein" ($n = 3381$), "uncharacterized protein" ($n = 58$), "predicted protein" ($n = 12$) or "unknown protein" ($n = 6$). Moreover, 30.4% of the PPM reference transcriptome retrieved no hit at all in NR, and could correspond to specific genes. Other transcriptome analyses of Lepidoptera species reported similar, or even lower, proportions of annotated transcripts, such as *P. xylostella* (22% of NR annotations, He et al., 2012b), *M. sexta* (49% in NR and ButterflyBase, Pauchet et al., 2010) or *Maruca vitrata* (39% in NR, Margam et al., 2011). This feature suggests that Lepidoptera genes are highly divergent from the well-studied model organisms representing most of the entries in the public databases. Experimental identification of the function of these specific proteins would be essential for understanding specific traits. However, such experimental validations can rarely be undergone, in particular for non-model organisms.

To further functionally analyse the PPM reference transcriptome, we applied a Gene Ontology search. 26,202 GO terms could be assigned to 43% ($n = 5159$) of the annotated PPM transcripts. Similar or lower proportions of GO-annotated transcripts were reported for other Lepidoptera transcriptomes, such as *Zygaena filipendulae* (11%, Zagrobelny et al., 2009), *Heliothis virescens* (15%, Shelby and Popham, 2009), or *M. sexta* (31%, Grosse-Wilde et al., 2011). The spelling of gene and protein names is not standardised in public databases such as GenBank, which hinders the

association of GO-terms, even though the GO vocabulary was meant to be species-independent and applicable to all kind of organisms (Ashburner et al., 2000).

The category distributions of the GO-terms assigned to the PPM transcriptome were similar to the main GO-annotations described in previously published Lepidoptera transcriptomes (*e.g.* He et al., 2012b; Legeai et al., 2011; Li et al., 2012; Pascual et al., 2012). The GO terms found for *T. pityocampa* are shown in Fig. 1.

3.1.5. Assembly completeness

We estimated the proportion of transcripts likely to contain a coding region using FrameDP, and we compared the length of each *de novo* assembled gene to reference Lepidoptera data through the calculation of the Ortholog Hit Ratio (see Material and methods). These clues can be used as indications of the quality of coding gene reconstitutions.

In the PPM reference assembly, FrameDP predicted at least partial CDS in 67.6% ($n = 8115$) of the transcripts, and full-length CDS in 47.4% ($n = 5693$). Quite similar proportions were obtained in previous transcriptomic studies in insects (*e.g.* 62% in antennae of the moth *S. littoralis*, Jacquin-Joly et al., 2012) or plants (*e.g.* 52% for the oak species *Quercus petreae* and *Q. robur*, Ueno et al., 2010). The remaining 32.4% of the transcripts without any predicted CDS likely corresponded to 3′- or 5′ UTRs that can reach thousands of bp in invertebrates (Pesole et al., 2001), or to incomplete transcript assembly that impeded CDS prediction. Consistent with this latter hypothesis, the transcripts containing a predicted CDS region were on average almost twice as long as the transcripts without CDS (1419 *vs.* 866 bp). They had on average 1.1 CDS (mean length = 759 bp); when several CDS were predicted in a single transcript ($n = 811$), they were on average separated by 333 bp, suggesting the presence of introns, *i.e.* of pre-mRNA in the sequenced libraries.

We calculated the OHR for the transcripts that had a BLAST hit in at least one of the studied Lepidoptera databases, *i.e.* for around 71.2% ($n = 8547$) of the PPM transcripts. The mean OHR was 0.70, and 5.5% ($n = 663$) of the transcripts had an OHR above 1, *i.e.* were longer than their counterparts in the Lepidoptera resource databases (Fig. 2A). By contrast, a similar proportion of transcripts (6.8%, $n = 816$) had an OHR below 0.2. These poorly assembled transcripts corresponded to genes with low levels of expression and were actually under-represented in the data set in spite of the molecular normalisation procedure. Indeed, the average coverage of the transcripts having the lowest OHR (<0.2) values was only 7.5 reads/bp, while it was 21.7 reads/bp when considering all the transcripts for which an OHR could be calculated.

3.1.6. Identification of phenology-related candidate genes within the obtained reference transcriptome

We then specifically focused on candidate genes potentially involved in the phenological change observed in Leiria. Of the set of 1307 phenology-related candidate proteins we defined (see Material and methods, Section 2.2.4), 65% ($n = 849$) could be recovered in the PPM transcriptome, corresponding to 464 transcripts. The mean OHR calculated on these candidate genes was 0.76, and 43 (9.3%) had an OHR above 1 (Fig. 2B). Moreover, 318 of these transcripts were predicted by FrameDP to have a complete CDS and 129 to have a partial CDS. The average length of the 17 PPM transcripts without any CDS prediction was 1041 bp.

The reference transcriptome obtained in the present study is the first high throughput genomic resource developed for the target species *T. pityocampa*. Although still incomplete, this assembly contains several thousands of genes, *ca.* half of which are reconstructed at full length. Annotation and GO term analyses suggest that we recovered a wide diversity of the expressed genes, which is likely due to the pooling of various development stages sampled in natural populations, *i.e.* responding to various environmental conditions. Concerning our primary goal – identification of genes potentially involved in phenology – our results show that the transcriptomic resources built here for the PPM allowed the recovery of a significant amount of such phenology-related candidate genes. Targeted resequencing of these genes in natural populations will hence be feasible in the future, as specific PCR primers can now be developed to screen and compare gene-specific diversity and expression patterns. This work is the first necessary step towards the development of genome-wide assays.

3.2. Comparative analyses of the SP and WP-specific transcriptomes

3.2.1. Population-specific PPM transcripts and unigenes

In order to identify population-specific transcripts within the reference transcriptome, the PPM transcripts were aligned to both the SP- and WP-transcriptomes. 90.2% ($n = 10,835$) of the PPM

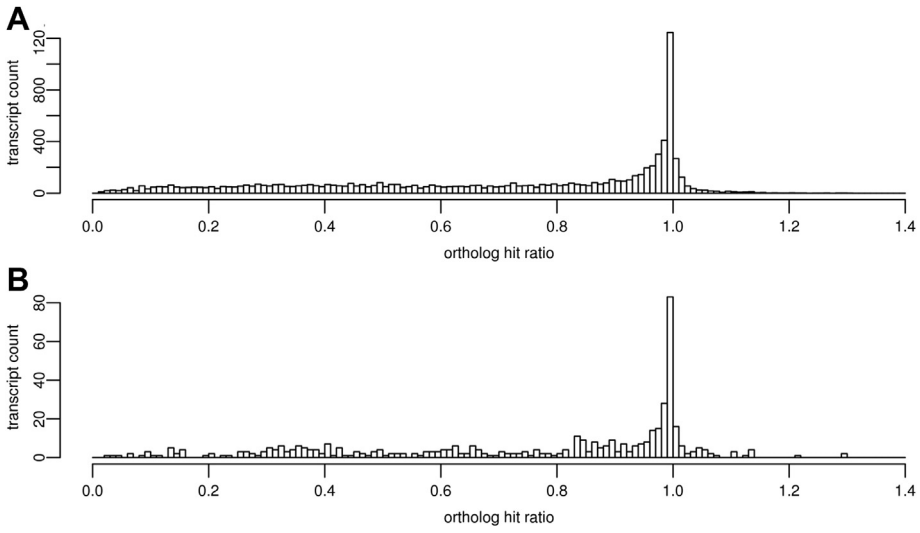

Fig. 2. Distribution of the number of transcripts per OHR class. (A) for all PPM transcripts with a similarity hit in the Lepidoptera protein set; (B) for the transcripts included in (A) and also corresponding to a phenology-related candidate gene. See text for details.

transcripts had at least one hit in one of the populations. The remaining 9.8% ($n = 1176$) were either essentially assembled from population singletons ($n = 1127$ transcripts), or corresponded to PPM-specific alternative transcripts containing additional assembled exons ($n = 49$ transcripts). In total, 782 PPM-transcripts had similarities only with SP-transcripts and 1002 only with WP-transcripts (Fig. 3A).

Besides, we estimated the number of SP and WP private unigenes that may differ from the number of private transcripts because of the occurrence of alternative transcripts (see Material and methods). We hence identified 680 PPM unigenes that corresponded exclusively to SP data, while 864 corresponded only to the WP (Fig. 3B). We then restricted the results to the "private unigenes" having a mean coverage above five reads per base pair (Fig. 3C). We thereby retained 242 SP-specific well-covered unigenes (corresponding to 252 transcripts) and 315 WP-specific unigenes (corresponding to 329 transcripts) (Additional files 3 and 4). No significant GO term enrichment was detected for any of these population-specific subsets when compared to the GO annotations obtained for the PPM unigenes. These genes could correspond to potential differentially expressed genes as defined by Logacheva et al. (2011). Yet, as the population SP- and WP-transcriptomes obtained in the present study are still incomplete (*ca.* 6700 and 7100 unigenes assembled in each population, respectively, see Table 1), and as some coding regions are only partially recovered, most of the population-specific unigenes identified here within the PPM reference transcriptome may well be due to a sampling bias rather than to an actual differential expression between the studied populations. The unigenes found in one population only could either be genes whose orthologs are actually not expressed in the other population, or genes that are expressed in both populations but are absent in the sequence data of one of them, due to some random fluctuations during the molecular procedures (Logacheva et al., 2011). Expression studies (*e.g.* using quantitative RT-PCR) would now be necessary to definitely conclude which of these genes displays differential expression between SP and WP.

Since one of our objectives was to identify a set of genes potentially involved in the allochronic differentiation between the SP and WP, we further examined the annotation of the well-covered, population-specific unigenes, and identified the ones that could correspond to phenology-related candidate genes. In total, 69 SP-specific and 115 WP-specific unigenes could be effectively annotated through NR search (Additional files 3 and 4), while 105 SP-specific and 105 WP-specific unigenes were associated to "hypothetical" or "unknown" proteins. Further, 68 SP-specific and 95 WP-specific unigenes did not have any NR counterpart. Finally, only two SP- and two WP-specific unigenes corresponded to phenology-related candidate genes, namely *timeless* and one juvenile hormone-inducible protein in the SP, and *takeout* and *ctrip* (goliath E3 ubiquitin ligase) in the WP. Phenology-related candidate genes were thus apparently not over-represented in the SP- nor in the WP-specific unigenes, as they corresponded to less than 1% of the population specific sets while they represented *ca.* 3% of the PPM transcriptome. Yet, as a high proportion of transcripts was not effectively annotated, the possibility remains that important genes involved in phenology in the PPM are included in the population-specific sets of transcripts, but could not be identified as such.

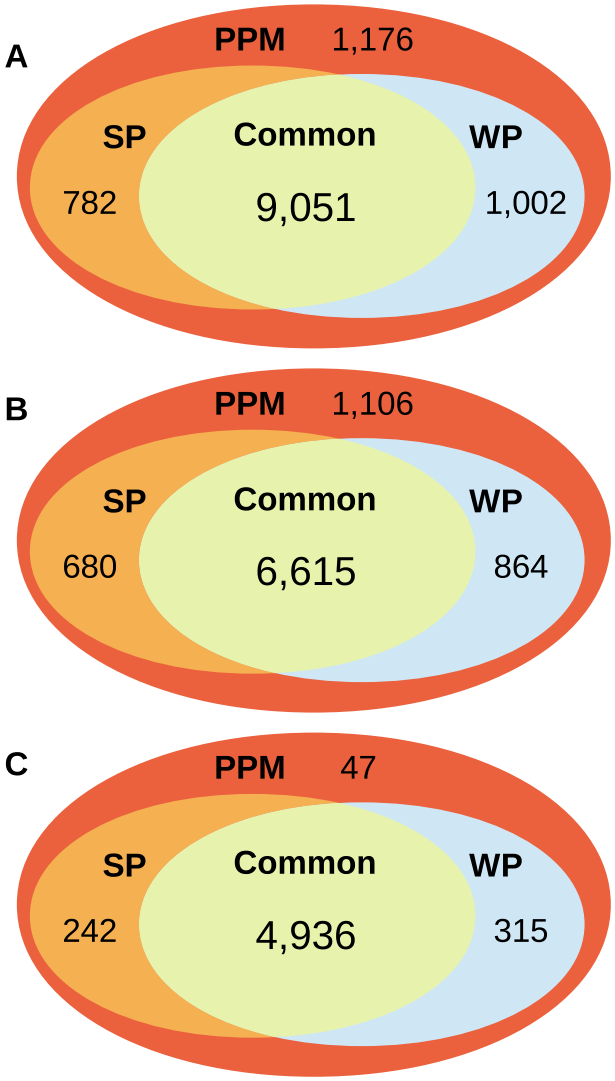

Fig. 3. Venn diagrams of specific and shared transcripts or unigenes between the PPM, SP and WP assemblies. (A) for the set of alternative transcripts; (B) for all unigenes; (C) for the unigenes with a minimal mean coverage of five reads/bp.

3.2.2. Divergence patterns between homologous SP- and WP-transcripts

We determined the divergence of homologous SP- and WP-transcripts pairs identified by reciprocal BLAST search. We identified 1295 pairs with nucleotide identities varying from 78.3% to 100% (mean identity = 99.5%) (Fig. 4). When ignoring the most divergent transcript pair, which most probably corresponded to

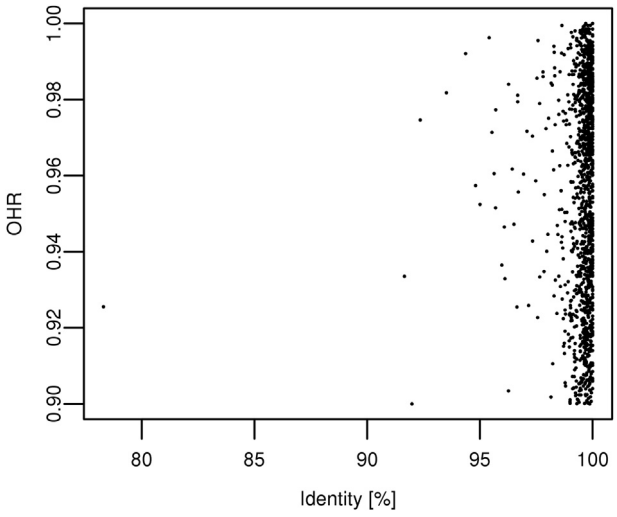

Fig. 4. Distribution of identity *versus* OHR of homologous transcripts identified in the SP and WP transcriptomes as reciprocal best hits.

Table 2
Annotations of the diagnostic SNPs.

Unigene ID	Transcript ID	Acc. number	NR function	Position	Base	Region	Codon substitution
isogroup01481	isotig03058	No hit	No hit	597	C -> T	CDS	TTC (F) -> TTT (F)
isogroup01919	isotig03551	EHJ74874	Nuclear orphan receptor	659	C -> T	CDS	GCC (A) -> GCT (A)
isogroup02186	isotig03818	EHJ79289	CRAL/TRIO domain-containing protein	1090	A -> G	CDS	GTA (V) -> GTG (V)
isogroup03063	isotig04695	EHJ72099	Hypothetical protein	960	T -> C	CDS	TGT (C) -> TGC (C)
isogroup02840	isotig04472	EHJ71781	Hypothetical protein	142	G -> A	5′ UTR	–
isogroup03366	isotig04998	No hit	No hit	534	G -> A	5′ UTR	–
isogroup00019	isotig00083, isotig00085	NP_001040287	Vacuolar ATP synthase subunit G	1225	T -> G	3′ UTR	–
isogroup00180	isotig00503	EHJ65449	Astryp1	4339	G -> A	3′ UTR	–
isogroup02027	isotig03659	BAM17986	ATP synthase-gamma chain	1779	G -> A	3′ UTR	–

paralogous sequences as they actually matched with two different PPM unigenes annotated as “serine protease” (Koonin, 2005), the lowest identity value reached 91.7%. Among those, 19 had counterparts in the set of phenology-related candidate genes (Additional file 5) and showed sequence identities from 95.7% to 100% (mean = 99.4%) between populations, which is close to the values obtained for other transcripts. Finally, 563 pairs (43%) were not functionally annotated as they had either no hit in NR or the corresponding hit was not informative (Additional file 6).

We then focused on the 112 most divergent homologous transcripts, with identities below 99%. We could assign NR functions to 65 such transcripts pairs, among which only two corresponded to phenology-related candidate genes. The first one was a diapause bioclock protein, while the other was identified as a juvenile hormone-inducible protein, which actually corresponds to a vast group of genes (Davey, 2000). We discovered as well divergent genes potentially involved in insect development such as arginase (Raghupathi Reddy and Campbell, 1969) and apoptosis inhibitor (He et al., 2012a). This set of divergent genes contained in addition chemosensory proteins, which typically can have a major role in adult communication and reproduction; whether individuals of the SP and WP have divergent olfactory phenotypes is still unknown. Several transcription and translation factors, as well as ribosomal proteins, were also identified as particularly divergent between species. Interestingly, even though SP and WP larvae were proved to have different thresholds of temperature survival (Santos et al., 2011b), the transcripts probably coding for various heat-shock proteins were very similar between populations.

Table 3
SNPs found within phenology-related candidate genes.

Unigene ID	Transcript ID	UniProtKB accession number	Function
isogroup00006	isotig00035	I4DPF8	Takeout/JHBP like protein
isogroup00034	isotig00135	C0KH33	Juvenile hormone epoxide hydrolase
isogroup00067	isotig00226	I4DIM4	Takeout/JHBP like protein
isogroup00710	isotig01572	O76484	Casein kinase II subunit alpha
isogroup00814	isotig01777	P22327	Acidic juvenile hormone-suppressible protein 1
isogroup01573	isotig03214	A5LFV6	Juvenile hormone acid methyl-transferase
isogroup01935	isotig03567	G6DFU6	Cullin 3
isogroup02473	isotig04105	A7LBQ0	L-lactate dehydrogenase
isogroup02516	isotig04148	Q1W696	Juvenile hormone epoxide hydrolase
isogroup02616	isotig04248	I4DJH4	Takeout/JHBP like protein
isogroup05081	isotig06713	C0MNP5	Serine/threonine phosphatase 2a
isogroup06089	isotig07721	Q402D8	Juvenile hormone binding protein

The comparative analyses conducted here provided (i) gene sets found specifically in one of the studied populations, and (ii) orthologs present in both populations showing differentiation at the sequence level between the SP and WP. The genes involved in the phenological shift could belong to both categories. Surprisingly, very few candidate genes potentially related to phenology could be identified in both sub-sets. Given the relatively limited sample size used per population and per development stage, the obtained gene coverage is probably not fully representative for each population, and we cannot rule out that the expression and divergence patterns found here are mostly due to a sampling bias (in terms of number of individuals and developmental stages included in the present study, or in terms of sequencing coverage and normalisation). Yet, another possibility is that the genes effectively responsible for the change in phenology are still un-annotated, either because the studied species is phylogenetically too far from the model Lepidoptera species represented in the reference databases, and orthologous genes are too divergent to enable a successful annotation *via* BLAST search; or because the genes of interest in the particular case of Leiria SP are actually still unknown. Moreover, the evolutionary and demographic histories of the SP are still poorly understood (Santos et al., 2011a, 2007). A relatively recent bottleneck has probably happened following the phenological shift of the founding individuals, causing stochasticity in the current patterns of genetic divergence over the genome. We suspect that this complex, recent demographic history may blur the effect of divergent selection acting on the causal gene(s), complicating its identification.

3.2.3. *In silico SNP detection*

We determined SNP positions by mapping the sequence reads of both the SP and WP datasets to the PPM exons. A total of 361,695 WP (88%) and 419,893 SP (89%) reads could be successfully aligned to the reference transcriptome. We detected 1451 SNP positions, located in 969 exons belonging to 748 PPM unigenes (1216 PPM alternative transcripts). Approximately half of the detected SNPs ($n = 734$) were located in transcripts containing a predicted CDS, including 77 within the 5′ UTR, 465 in the CDS and 153 in the 3′ UTR, while 39 SNPs were predicted to be located in putative introns. Hence, almost half of the SNPs were found in transcripts without a predicted coding region, although such transcripts corresponded to only 30% of the assembly. This suggests that polymorphic sites are preferentially found in non-coding regions of cDNAs that are probably less constrained.

A PPM exon had on average 0.11 SNP, corresponding to 0.11 SNP per kbp. Among the identified SNPs, 145 were polymorphic within each population (within-population SNPs), 1297 were polymorphic in one population and fixed in the other (565 SP-specific and 732 WP-specific SNPs), and the remaining nine

sites were diagnostic, *i.e.* fixed in each population and different between populations.

We then focused on SNPs located within CDS regions to determine whether they corresponded to synonymous or non-synonymous changes. We could identify 1008 codons containing a SNP, in which the most frequent substitutions were adenine substituted by guanine (A→G, 35%) and thymine substituted by cytosine (T→C, 23.6%), followed by C→T (13.5%) and G→A (10.5%). When examining each class of SNP separately, we found that the most frequent nucleotide substitutions were C→T (50%) in diagnostic SNPs, A→G (36.8%) in population-specific SNPs and A→G and G→A (each 19.2%) in within-population SNPs. In total, synonymous substitutions were present in 842 codons (diagnostic: 4, population-specific: 761, within-population: 77), whereas the other 166 codons encoded nonsynonymous amino acids (population-specific: 142, within-population: 24).

The nine diagnostic SNPs were located in nine unigenes (corresponding to ten transcripts) (Table 2); two of these SNPs were identified within the 5′UTR, four in the CDS and three in the 3′UTR. All four CDS-located diagnostic SNPs corresponded to synonymous changes. Seven of the nine unigenes containing diagnostic SNPs were assigned similarities in NR; among these, we identified the CRAL/TRIO domain-containing protein, which is essential for cell migration and growth (Johnson and Kornfeld, 2010), a trypsin-like serine protease (*D. plexippus* Astryp1 homolog), the vacuolar H + ATPase (the principal regulator of ion-transport processes in the lepidopteran midgut, Wieczorek et al., 2009) and an ATP synthase gamma chain (Table 2). In addition, we identified a nuclear orphan receptor, which belongs to a superfamily of transcription regulators involved in reproduction and development in eukaryotes (Bain et al., 2007). The two remaining unigenes were not effectively annotated, as they corresponded to "hypothetical" proteins. Among the 1451 identified SNPs, twelve were located in transcripts annotated as phenology-related candidate genes, including several proteins related to the juvenile hormone pathway, as well as casein kinase 2, cullin 3, serine/threonine-protein phosphatase 2A (*PP2A*) and lactate dehydrogenase (*ldh*) (Table 3).

Genotyping assays will be necessary to validate the SNPs identified here *in silico*, and to confirm their status (diagnostic SNP, or SNP found polymorphic in one population only) using a larger sample size. Previous studies using *de novo* transcriptomic assemblies have proved useful to identify SNPs and develop genotyping procedures such as multiplex PCR analyses and chip-based genotyping (Coates et al., 2011; Margam et al., 2011). These markers are promising in Lepidoptera, for which development and routine use of microsatellite loci are particularly challenging (Meglecz et al., 2004; Van't Hof et al., 2007). Pan-genomic SNP loci are increasingly being used in population genetics and genomics (Helyar et al., 2011), and the candidate SNPs identified for the PPM in the present study will be most useful to decipher population genetic structures and demographic histories, and to estimate the strength of evolutionary forces acting on the studied populations.

4. Conclusions

We have established a reference transcriptome for the pine processionary moth, *T. pityocampa* based on a combination of 454 and Sanger sequencing technologies. This study represents a fundamental progress towards the understanding of this plague of pine trees that causes severe allergic reactions in humans and animals. The present transcript set has a high sequencing coverage (22 reads/bp) and comprises 9265 unigenes. The majority of these *de novo* genes were homologs to already identified genes of Lepidoptera and were well reconstituted (mean OHR of 0.71). Yet, we stated that a large portion of PPM transcripts could not be functionally annotated: around 30% of the transcripts did not have any NR similarities and 41.3% of the transcripts with NR counterparts were of unknown function. This underlines the urgent need for research development in Lepidoptera to characterise their genes and to decipher their functions.

We also compared the expressed genes of two PPM populations showing contrasted reproduction time (allochrony). We applied several comparative analyses in order to identify candidate genes which might be related to the shift in life cycle. Divergence analyses of orthologous genes and SNP identifications pointed out a list of potentially interesting genes. Yet, many of those were not functionally annotated, and a few were identified as phenology-related candidate genes. We also discovered unigenes that were exclusively expressed in one of these PPM populations only. This latter study however has to be confirmed and detailed by quantitative RNA-Seq expression experiments in order to exclude eventual sampling biases, for instance due to lack of non-sequenced developmental stages. In the present study, we detected differences between the studied populations, but the sampling and sequencing design did not allow distinguishing between the evolutionary forces that caused these variations (local adaptation to environmental conditions, drift, demographic history, modifications in development cycle *etc.*). More, even though the main phenotypic divergence between the SP and WP is associated with pupal diapause termination, the possibility exists that regulation genes are expressed in earlier development stages, and were missed in our study. Since it is extremely difficult to rear PPM in the laboratory across several generations, we used samples from natural populations. These samples necessarily faced different environmental conditions and showed a high level of heterozygosity which renders sequence assembly and identification of differentially expressed genes more difficult.

A perspective of the present study will be to develop comparative transcriptomic analyses separately for the various developmental stages using new high-throughput sequencing technologies such as Illumina data. This work also gives deeper insights into the genomics of non-model Lepidoptera species. The established set of 1451 SNPs – once experimentally validated – will provide a basis for further population genomic approaches and genome-wide scans of population differentiation in order to identify signatures of selection and decipher the evolutionary forces involved in the SP/WP differentiation.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors greatly acknowledge Manuela Branco and Helena Santos (Instituto Superior de Agronomia, University of Lisbon, Portugal, http://www.isa.utl.pt/pt) for providing samples for RNA extractions. We are grateful to Henriette Ringys-Beckstein and Domenica Schnabelrauch (Max Planck Institute for Chemical Ecology, Department of Entomology, Jena/Germany, http://www.ice.mpg.de/ext/entomology.html) for general technical assistance and Sanger sequencing. We would like to thank Jérôme Gouzy and Sébastien Carrere of the institute Laboratoire Interactions Plantes Micro-organismes (INRA/CNRS Toulouse, France, http://www.toulouse.inra.fr/lipm) for their technical advice regarding data cleaning and assembly. We also thank Franck Dorkeld for keeping updated the local NR and GO databases. We are grateful to the

bioinformatics platform CNRS-UPMC ABIMS (http://abims.sb-roscoff.fr) for providing computational resources and support. This study was supported by the European NoE Evoltree (FP6-2004-GLOBAL-3) and by the French National Agency for Research (project GenoPheno, ANR-2010-JCJC-1705 01).

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ibmb.2014.01.005.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Bain, D.L., Heneghan, A.F., Connaghan-Jones, K.D., Miura, M.T., 2007. Nuclear receptor structure: implications for function. Annu. Rev. Physiol. 69, 201–220.

Battisti, A., Stastny, M., Netherer, S., Robinet, C., Schopf, A., Roques, A., Larsson, S., 2005. Expansion of geographic range in the pine processionary moth caused by increased winter temperatures. Ecol. Appl. 15, 2084–2096.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Methodol. 57, 289–300.

Coates, B.S., Bayles, D.O., Wanner, K.W., Robertson, H.M., Hellmich, R.L., Sappington, T.W., 2011. The application and performance of single nucleotide polymorphism (SNP) markers for population genetic analyses of Lepidoptera. Front. Genet. 2.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Davey, K.G., 2000. The modes of action of juvenile hormones: some questions we ought to ask. Insect. Biochem. Mol. Biol. 30, 663–669.

Duan, J., Li, R., Cheng, D., Fan, W., Zha, X., Cheng, T., Wu, Y., Wang, J., Mita, K., Xiang, Z., Xia, Q., 2010. SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. Nucleic Acids Res. 38, D453–D456.

Ewen-Campen, B., Shaner, N., Panfilio, K., Suzuki, Y., Roth, S., Extavour, C., 2011. The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. BMC Genomics 12, 61.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., Martin, J.-F., 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12, 245.

Gouzy, J., Carrere, S., Schiex, T., 2009. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25, 670–671.

Gregory, T.R., 2005. Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. 6, 699–708.

Gregory, T.R., Hebert, P.D.N., 2003. Genome size variation in lepidopteran insects. Can. J. Zool. 81, 1399–1405.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., Bennett, M.D., 2007. Eukaryotic genome size databases. Nucleic Acids Res. 35, D332–D338.

Grosse-Wilde, E., Kuebler, L.S., Bucks, S., Vogel, H., Wicher, D., Hansson, B.S., 2011. Antennal transcriptome of *Manduca sexta*. P. Natl. Acad. Sci. U. S. A. 108, 7449–7454.

Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., Frazer, K., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10, R32.

He, H.J., Hou, L., Wang, J.X., Zhao, X.F., 2012a. The apoptosis inhibitor survivin prevents insect midgut from cell death during postembryonic development. Mol. Biol. Rep. 39, 1691–1699.

He, W., You, M., Liette, V., Yang, G., Xie, M., Cui, K., Bai, J., Liu, C., Li, X., Xu, X., Huang, S., 2012b. Developmental and insecticide-resistant insights from the *de novo* assembled transcriptome of the diamondback moth, *Plutella xylostella*. Genomics 99, 169–177.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol. Ecol. Res. 11, 123–136.

Hou, Y., Lin, S., 2009. Distinct gene number-genome size relationships for Eukaryotes and non-Eukaryotes: gene content estimation for Dinoflagellate genomes. PLoS ONE 4, e6978.

Huchon, H., Démolin, G., 1970. La bioécologie de la processionnaire du pin. Dispersion potentielle – dispersion actuelle. Rev. For. Fr. 22, 220–234.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8, R143.

Jacquin-Joly, E., Legeai, F., Montagné, N., Monsempes, C., François, M.-C., Poulain, J., Gavory, F., Walker, W.B.I., Hansson, B.S., Larsson, M.C., 2012. Candidate chemosensory genes in female antennae of the noctuid moth *Spodoptera littoralis*. Int. J. Biol. Sci. 8, 1036–1050.

Johnson, K.G., Kornfeld, K., 2010. The CRAL/TRIO and GOLD domain protein TAP-1 regulates RAF-1 activation. Dev. Biol. 341, 464–471.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kerdelhué, C., Zane, L., Simonato, M., Salvato, P., Rousselet, J., Roques, A., Battisti, A., 2009. Quaternary history and contemporary patterns in a currently expanding species. BMC Evol. Biol. 9, 220.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C., 2011a. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS ONE 6, e15925.

Kofler, R., Pandey, R.V., Schlötterer, C., 2011b. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27, 3435–3436.

Koonin, E.V., 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39, 309–338.

Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S., Robinson, P.N., 2010. Microindel detection in short-read sequence data. Bioinformatics 26, 722–729.

Kumar, S., Blaxter, M., 2010. Comparing *de novo* assemblers for 454 transcriptome data. BMC Genomics 11, 571.

Legeai, F., Malpel, S., Montagne, N., Monsempes, C., Cousserans, F., Merlin, C., Francois, M.-C., Maibeche-Coisne, M., Gavory, F., Poulain, J., Jacquin-Joly, E., 2011. An expressed sequence tag collection from the male antennae of the Noctuid moth *Spodoptera littoralis*: a resource for olfactory and pheromone detection research. BMC Genomics 12, 86.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, G.P.D.P., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, Y., Wang, G., Tian, J., Liu, H., Yang, H., Yi, Y., Wang, J., Shi, X., Jiang, F., Yao, B., Zhang, Z., 2012. Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA sequencing. PLoS ONE 7, e43713.

Logacheva, M., Kasianov, A., Vinogradov, D., Samigullin, T., Gelfand, M., Makeev, V., Penin, A., 2011. *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12, 30.

Margam, V.M., Coates, B.S., Bayles, D.O., Hellmich, R.L., Agunbiade, T., Seufferheld, M.J., Sun, W., Kroemer, J.A., Ba, M.N., Binso-Dabire, C.L., Baoua, I., Ishiyaku, M.F., Covas, F.G., Srinivasan, R., Armstrong, J., Murdock, L.L., Pittendrigh, B.R., 2011. Transcriptome sequencing, and rapid development and application of SNP markers for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). PLoS ONE 6, e21388.

Meglecz, E., Petenian, F., Danchin, E., Coeur D'Acier, A., Rasplus, J.-Y., Faure, E., 2004. High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. Mol. Ecol. 13, 1693–1700.

Mutanen, M., Wahlberg, N., Kaila, L., 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc. R. Soc. Lond. B 277, 2839–2848.

Myhre, S., Tveit, H., Mollestad, T., Laegreid, A., 2006. Additional gene ontology structure for improved biological reasoning. Bioinformatics 22, 2020–2027.

Ning, Z., Cox, A.J., Mullikin, J.C., 2001. SSAHA: a fast search method for large DNA databases. Genome Res. 11, 1725–1729.

O'Neil, S.T., Dzurisin, J.D., Carmichael, R.D., Lobo, N.F., Emrich, S.J., Hellmann, J.J., 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics 11, 310.

Pascual, L., Jakubowska, A.K., Blanca, J.M., Canizares, J., Ferré, J., Gloeckner, G., Vogel, H., Herrero, S., 2012. The transcriptome of *Spodoptera exigua* larvae exposed to different types of microbes. Insect Biochem. Mol. Biol. 42, 557–570.

Pauchet, Y., Wilkinson, P., Vogel, H., Nelson, D.R., Reynolds, S.E., Heckel, D.G., Ffrench-Constant, R.H., 2010. Pyrosequencing the *Manduca sexta* larval midgut transcriptome: messages for digestion, detoxification and defence. Insect Mol. Biol. 19, 61–75.

Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., 2001. Structural and functional features of eukaryotic mRNA untranslated regions. Gene 276, 73–81.

Pimentel, C., Calvão, T., Santos, M., Ferreira, C., Neves, M., Nilsson, J.-Å., 2006. Establishment and expansion of a *Thaumetopoea pityocampa* (Den. & Schiff.) (Lep. Notodontidae) population with a shifted life cycle in a production pine forest, Central-Coastal Portugal. For. Ecol. Manag. 233, 108–115.

Raghupathi Reddy, S.R., Campbell, J.W., 1969. Arginine metabolism in insects. Role of arginase in proline formation during silkmoth development. Biochem. J. 115, 495–503.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Rodríguez-Mahillo, A.I., González-Muñoz, M., Vega, J.M., López, J.A., Yart, A., Kerdelhué, C., Camafeita, E., García Ortiz, J.C., Vogel, H., Toffolo, E.P., Zovi, D., Battisti, A., Roques, A., Moneo, I., 2012. Setae from the pine processionary moth (*Thaumetopoea pityocampa*) contain several relevant allergens. Contact Dermat. 67, 367–374.

Santos, H., Burban, C., Rousselet, J., Rossi, J.-P., Branco, M., Kerdelhué, C., 2011a. Incipient allochronic speciation in the pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera, Notodontidae). J. Evol. Biol. 24, 146–158.

Santos, H., Paiva, M.R., Tavares, C., Kerdelhué, C., Branco, M., 2011b. Temperature niche shift observed in a Lepidoptera population under allochronic divergence. J. Evol. Biol. 24, 1897–1905.

Santos, H., Rousselet, J., Magnoux, E., Paiva, M.R., Branco, M., Kerdelhué, C., 2007. Genetic isolation through time: allochronic differentiation of a phenologically atypical population of the pine processionary moth. Proc. R. Soc. Lond. B 274, 935–941.

Shelby, K.S., Popham, H.J.R., 2009. Analysis of ESTs generated from immune-stimulated hemocytes of larval *Heliothis virescens*. J. Invertebr. Pathol. 101, 86–95.

Sparks, M.E., Gundersen-Rindal, D.E., 2011. The *Lymantria dispar* IPLB-Ld652Y cell line transcriptome comprises diverse virus-associated transcripts. Viruses 3, 2339–2350.

Ueno, S., Le Provost, G., Leger, V., Klopp, C., Noirot, C., Frigerio, J.-M., Salin, F., Salse, J., Abrouk, M., Murat, F., Brendel, O., Derory, J., Abadie, P., Leger, P., Cabane, C., Barre, A., de Daruvar, A., Couloux, A., Wincker, P., Reviron, M.-P., Kremer, A., Plomion, C., 2010. Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. BMC Genomics 11, 650.

Van't Hof, A.E., Brakefield, P.M., Saccheri, I.J., Zwaan, B.J., 2007. Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera. Heredity 98, 320–328.

Vega, J., Vega, J.M., Moneo, I., Armentia, A., Caballero, M.L., Miranda, A., 2004. Occupational immunologic contact urticaria from pine processionary caterpillar (*Thaumetopoea pityocampa*): experience in 30 cases. Contact Dermat. 50, 60–64.

Vogel, H., Heidel, A.J., Heckel, D.G., Groot, A.T., 2010. Transcriptome analysis of the sex pheromone gland of the noctuid moth *Heliothis virescens*. BMC Genomics 11, 29.

Vogel, H., Wheat, C., 2011. Accessing the transcriptome: how to normalize mRNA pools. In: Orgogozo, V., Rockman, M.V. (Eds.), Molecular Methods for Evolutionary Genetics. Humana Press, New-York, pp. 105–128.

Winnebeck, E.C., Millar, C.D., Warman, G.R., 2010. Why does insect RNA look degraded? J. Insect Sci. 10, 159.

Wieczorek, H., Beyenbach, K.W., Huss, M., Vitavska, O., 2009. Vacuolar-type proton pumps in insect epithelia. J. Exp. Biol. 212, 1611–1619.

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., Zhan, D., Baxter, S.W., Vasseur, L., Gurr, G.M., Douglas, C.J., Bai, J., Wang, P., Cui, K., Huang, S., Li, X., Zhou, Q., Wu, Z., Chen, Q., Liu, C., Wang, B., Li, X., Xu, X., Lu, C., Hu, M., Davey, J.W., Smith, S.M., Chen, M., Xia, X., Tang, W., Ke, F., Zheng, D., Hu, Y., Song, F., You, Y., Ma, X., Peng, L., Zheng, Y., Liang, Y., Chen, Y., Yu, L., Zhang, Y., Liu, Y., Li, G., Fang, L., Li, J., Zhou, X., Luo, Y., Gou, C., Wang, J., Wang, J., Yang, H., Wang, J., 2013. A heterozygous moth genome provides insights into herbivory and detoxification. Nat. Genet. 45, 220–225.

Zagrobelny, M., Scheibye-Alsing, K., Jensen, N., Moller, B., Gorodkin, J., Bak, S., 2009. 454 pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides. BMC Genomics 10, 574.

Zhan, S., Merlin, C., Boore, J.L., Reppert, S.M., 2011. The Monarch butterfly genome yields insights into long-distance migration. Cell 147, 1171–1185.