



# Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*

Raphaël Leblois, Mathieu Gautier, Audrey Rohfritsch, Julien Foucaud, Christian Burban, Maxime Galan, Anne Loiseau, Laure Sauné, Manuela Branco, K. Gharbi, et al.

## ► To cite this version:

Raphaël Leblois, Mathieu Gautier, Audrey Rohfritsch, Julien Foucaud, Christian Burban, et al.. Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*. *Molecular Ecology*, 2018, 27 (1), pp.264-278. 10.1111/mec.14411 . hal-02626319

**HAL Id: hal-02626319**

**<https://hal.inrae.fr/hal-02626319>**

Submitted on 11 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*

### Citation for published version:

Leblois, R, Gautier, M, Rohfritsch, A, Foucaud, J, Burban, C, Galan, M, Loiseau, A, Sauné, L, Branco, MDO, Gharbi, K, Vitalis, R & Kerdelhué, C 2017, 'Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*', *Molecular Ecology*.  
<https://doi.org/10.1111/mec.14411>

### Digital Object Identifier (DOI):

[10.1111/mec.14411](https://doi.org/10.1111/mec.14411)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Peer reviewed version

### Published In:

Molecular Ecology

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



1       **Deciphering the demographic history of allochronic differentiation in the**  
2                               **pine processionary moth *Thaumetopoea pityocampa***

3  
4   R. Leblois<sup>1,5,\*</sup>, M. Gautier<sup>1,5,\*</sup>, A. Rohfritsch<sup>1</sup>, J. Foucaud<sup>1</sup>, C. Burban<sup>2</sup>, M. Galan<sup>1</sup>, A.  
5   Loiseau<sup>1</sup>, L. Sauné<sup>1</sup>, M. Branco<sup>3</sup>, K. Gharbi<sup>4</sup>, R. Vitalis<sup>1,5</sup> and C. Kerdelhué<sup>1</sup>

6   **Full postal addresses**

7   1. CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, 755 avenue du  
8   Campus Agropolis, CS 300 16, F-34988 Montferrier sur Lez cedex, France

9   2. INRA, UMR1202 BIOGECO (INRA – Université de Bordeaux), 69 Route d'Arcachon, F-  
10   33612 Cestas cedex, France

11   3. Centro de Estudos Florestais (CEF), Instituto Superior de Agronomia (ISA), University of  
12   Lisbon, Lisbon, Portugal

13   4. Edinburgh Genomics, School of Biological Sciences, University of Edinburgh, Edinburgh,  
14   EH9 12 3JT, UK

15   5. Institut de Biologie Computationnelle (IBC), Université de Montpellier, 95 rue de la  
16   Galera, 34095 Montpellier, France

17   \* equal author contributions

18  
19   **Keywords (4-6):** population genomics, RAD-sequencing, demographic inference, detection  
20   of selection, pine processionary moth, phenology

21  
22   **Name, address, fax number and email of corresponding author**

23   Carole Kerdelhué, INRA, UMR CBGP, 755 avenue du Campus Agropolis, CS 300 16. F-  
24   34988 Montferrier-sur-Lez cedex, France. Fax: + 33 4 99 62 33 45; e-mail:  
25   Carole.Kerdelhue@inra.fr

26  
27   **Running title:** Demographic history of an allochronic population

28 **ABSTRACT**

29 Understanding the processes of adaptive divergence, which may ultimately lead to speciation,  
30 is a major question in evolutionary biology. Allochronic differentiation refers to a particular  
31 situation where gene flow is primarily impeded by temporal isolation between early and late  
32 reproducers. This process has been suggested to occur in a large array of organisms, even  
33 though it is still overlooked in the literature. We here focused on a well-documented case of  
34 incipient allochronic speciation in the winter pine processionary moth *Thaumetopoea*  
35 *pityocampa*. This species typically reproduces in summer and larval development occurs  
36 throughout autumn and winter. A unique, phenologically shifted population (SP) was  
37 discovered in 1997 in Portugal. It was proved to be strongly differentiated from the sympatric  
38 "winter population" (WP), but its evolutionary history could only now be explored. We took  
39 advantage of the recent assembly of a draft genome and of the development of pan-genomic  
40 RAD-seq markers to decipher the demographic history of the differentiating populations and  
41 develop genome scans of adaptive differentiation. We showed that the SP diverged relatively  
42 recently, i.e. few hundred years ago, and went through two successive bottlenecks followed  
43 by population size expansions, while the sympatric WP is currently experiencing a population  
44 decline. We identified outlier SNPs that were mapped onto the genome, but none were  
45 associated with the phenological shift or with subsequent adaptations. The strong genetic drift  
46 that occurred along the SP lineage certainly challenged our capacity to reveal functionally  
47 important loci.

48    **INTRODUCTION**

49    Ecological speciation in sympatry, the process by which adaptation to contrasting ecological  
50    conditions drives the divergence of co-occurring populations, has received growing attention  
51    in the last 12 years (Rundle & Nosil, 2005). The fate of diverging populations maintaining a  
52    certain level of gene flow, and the conditions in which speciation can still occur are central  
53    questions in evolutionary biology (Smadja & Butlin, 2011). A mechanism possibly causing  
54    sympatric speciation is allochronic differentiation, which occurs when differences in breeding  
55    time within a species lead to temporal assortative mating and limit gene flow between early  
56    and late reproducers (Alexander & Bigelow, 1960). Isolation-by-time can further lead to  
57    adaptation-by-time (Hendry & Day, 2005) when divergent selection operates between  
58    contrasting environmental conditions encountered at the different breeding times. This  
59    process remains largely unexplored in the literature, but has been suggested to occur in a large  
60    array of organisms such as plants (Devaux & Lande, 2008; Savolainen et al., 2006; Weis et  
61    al., 2005), birds (Friesen et al., 2007), fishes (Limborg, Waples, Seeb, & Seeb, 2014), corals  
62    (Rosser, 2015, 2016) or insects (Santos, Burban, et al., 2011; Sota et al., 2013; Yamamoto &  
63    Sota, 2009, 2012; Yamamoto, Beljaev, & Sota, 2016). Many more examples probably remain  
64    to be discovered, and only 9 study cases were identified in a recent review of 200 papers as  
65    examples of "true allochronic speciation" (Taylor & Friesen, 2017). To go beyond the  
66    description of such case studies and disentangle the evolutionary scenarios underlying  
67    allochronic differentiation, much remains to be done; in particular, the initial reduction of  
68    migration rate between the diverging populations and the underlying genomic mechanisms  
69    remain to be explored in most cases.

70    The recent advent of high-throughput genomic techniques as well as statistical advances for

analysing large-scale datasets have opened unprecedented opportunities to address ecological, evolutionary and genetic questions in non-model organisms (Hasselmann, Ferretti, & Zayed, 2015). Genome-wide data have been proven to be powerful for estimating the age of demographic events (McCoy, Garud, Kelley, Boggs, & Petrov, 2014), retrieving fine-scale population genetic structures (Szulkin, Gagnaire, Bierne, & Charmantier, 2016), and identifying phylogeographic patterns (Derkarabetian, Burns, Starrett, & Hedin, 2016). More, even if studying wild populations of non-model organisms is still a major challenge, population genomic approaches have allowed identification of genomic regions underlying phenotypic characteristics or traits involved in local adaptation (e.g., Berdan, Mazzoni, Waurick, Roehr, & Mayer, 2015; Guo, DeFaveri, Sotelo, Nair, & Merilä, 2015; Hohenlohe, 2014). Here, we used population genomics in a non-model insect species to disentangle the evolutionary scenario of allochronic differentiation, followed by adaptation to new environmental conditions.

We focused on one of the few cases identified by Taylor and Friesen (2017) as a well-documented example of "true incipient allochronic speciation", namely the pine processionary moth *Thaumetopoea pityocampa* (Dennis & Schiffermüller). This species is a well-studied pest of pine trees over the Mediterranean basin. Its caterpillars bear urticating hair, causing public and animal health concern (Battisti, Holm, Fagrell, & Larsson, 2011; Battisti, Larsson, & Roques, 2017; Rodríguez-Mahillo et al., 2012). Briefly, *T. pityocampa* reproduces in summer and larval development occurs through autumn and winter all over its range. Reproduction immediately follows adult emergence, as adults have a very limited lifespan of 1-2 days. In 1997, a population of *T. pityocampa* showing a shift in phenology (reproduction in spring and larval development in summer) was discovered in the Mata Nacional de Leiria (MNL) in Portugal, where it co-occurred with individuals following the typical biological

95 cycle (Pimentel et al., 2006; Santos et al., 2007). This unique shifted population is known as  
96 the "Summer Population" (SP) as opposed to all other known populations that are referred to  
97 as "Winter Populations" (WPs), in relation to the development time of the conspicuous larvae.  
98 The SP was initially restricted to a small area of the Mata Nacional, and has been slowly  
99 expanding along the coast since then (Godefroid et al., 2016). Strikingly, all larval stages of  
100 the SP develop under radically different environmental conditions compared to the typical  
101 WPs, experiencing much higher temperatures that were so far supposed to be lethal to early  
102 larval stages (Huchon & Démolin, 1970; Santos, Paiva, Tavares, Kerdelhué, & Branco, 2011).  
103 Understanding the scenario of this divergence is thus of interest in the context of current  
104 climate warming.

105 Previous studies have brought significant preliminary knowledge about the genetic and  
106 ecological characteristics of the peculiar SP. Analysis of a fragment of the mitochondrial COI  
107 gene and of the ITS1 region showed a high sequence similarity between the SP and the  
108 sympatric WP, which suggested a local origin of the SP, while microsatellites revealed a high  
109 differentiation between the SP and all studied Iberian populations (Santos, Burban, et al.,  
110 2011; Santos et al., 2007). Moreover, a recent study showed that some individuals belonging  
111 to the SP genetic cluster emerge during the WP reproductive season, and are referred to as  
112 "LateSP individuals" (Burban et al., 2016). This study also documented signs of rare  
113 hybridization between the two allochronic populations. Consistently, hybrids between SP and  
114 WP individuals could be obtained in laboratory conditions, and the time of adult emergence (a  
115 proxy for breeding time) was shown to be highly heritable (Branco, Paiva, Santos, Burban, &  
116 Kerdelhué, 2017). These patterns suggested that the SP originated from the WP, following a  
117 phenological shift of a few individuals, and that gene flow between the SP and the WP is now  
118 highly reduced but not absent. Yet, the population genetic data relied on a limited number of

microsatellite markers, and did not allow us to characterize the successive stages of the divergence between the SP and the WP.

The objectives of the present work were to uncover major characteristics of this prime example of allochronic differentiation and significantly move towards the fulfilment of the criteria proposed by Taylor and Friesen (2017) by deciphering the evolutionary history of the SP and characterizing its different stages. In particular, we aimed at (i) inferring the timing of the divergence, (ii) measuring the migration rate between diverging populations at different stages to determine if the differentiation occurred in the presence or absence of gene flow; (iii) determining the extent of population size changes, in particular to decipher if the SP experienced a strong bottleneck during the primary divergence step; and (iv) characterizing genomic regions possibly involved in the phenological shift and subsequent adaptations. To achieve these aims, we took advantage of RAD-seq technology (Baird et al., 2008; Davey & Blaxter, 2011) and the recent release of a first draft genome for *T. pityocampa* (Gschloessl et al., Submitted) to obtain a large number of informative loci genotyped in the SP and in two WPs occurring in the same region. We used these loci to explore complex demographic scenarios including drift, migration, and variation in population size, and to perform genome-wide scans for signatures of selection. We could thereby successfully disentangle the main characteristics of the on-going allochronic differentiation process.

## MATERIALS AND METHODS

### Biological material

A total of 180 individuals (adult or larvae) of *T. pityocampa* were collected in Portugal



between May 2008 and September 2010 following Santos, Burban, et al. (2011). These individuals originated from three distinct populations or sampling sites: two were collected in the MNL (39°47'N 8°58'W) and corresponded to the Winter and Summer populations from Leiria (referred to as LWP and LSP), and one winter population was collected in the Setubal peninsula, near Apostiça (38°34'N 9°07'W), ca. 150 km south from Leiria, at the same elevation and longitude, and was hereafter denoted as AWP. All individuals were sampled from the host plant *Pinus pinaster* Aiton.

Forty L5 larvae (i.e., 5<sup>th</sup> larval stage) belonging to the AWP were collected in December 2010; 40 males, 10 females and 20 L5 larvae belonging to the LWP were sampled in 2008-2010 and 60 males and 10 females belonging the LSP were sampled in 2008 – 2010. For the LSP and LWP, we used two sub-samples in each case. The first one included individuals assigned to the Winter or the Summer population based on the phenology observed in the field following Santos, Burban, et al. (2011) ( $n = 40$  for each population, sub-samples referred to as LSP1 and LWP1). The second sub-sample gathered males caught with pheromone traps and previously genotyped using 17 microsatellite loci, from which we excluded the individuals assigned as LateSP, F1 and F2 following Burban et al. (2016); these sub-samples ( $n = 30$  in each population) will be referred to as LSP2 and LWP2. The exact sampling design is described in Table 1.

## **RAD-sequencing and SNP calling**

### *RAD-libraries*

We carried out RAD tag sequencing (Baird et al., 2008) using both individual DNA and population pools of DNA. Individual data were used in combination with the pools to explore

the possible causes of the somewhat unexpected results yielded from the LWP1 pool (see Results). In total, we constructed six *Pst*I-digested paired-end (PE) RAD libraries as described in Gautier et al. (2013).

Twenty LSP1 (10 males and 10 females) and 20 LWP1 (10 males and 10 females) samples were barcoded and processed individually in libraries #1 to #3 (RAD Ind-Seq using 40 barcodes in total, see Table 1). The DNAs of each of the three sampled populations (all the 40 LSP1, all the 40 LWP1 and all the 40 AWP individuals respectively) were pooled and each population pool was barcoded with three different barcodes (9 barcodes in total, library #4). The RAD libraries #1 to #4 were then combined and PE sequenced ( $2 \times 101$  bp) on two Illumina HiSeq2000 lanes in the Edinburgh Genomics facility.

To replicate the experiment using only individuals genetically assigned to the LSP and LWP clusters following Burban et al. (2016), we further constructed 1 LSP2 and 1 LWP2 RAD libraries. Library #5 was constructed using 20 LWP2 males individually barcoded and a pool of all the 30 LWP2 DNAs that was identified with ten different barcodes. Finally, library #6 was the counterpart of library #5 for the LSP2 batch. Libraries #5 and #6 were each PE sequenced ( $2 \times 101$  pb) on a single Illumina HiSeq2000 lane on The Edinburgh Genomics facility.

### *Bioinformatic analyses*

Reads were first demultiplexed according to their barcode into individual and pool sequences using the default options of the *process\_radtags* program of the STACKS package (version 0.99994) (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe,

Bassham, Amores, & Cresko, 2013), including *-q* to remove poor quality reads. PCR duplicates were further discarded using the *clone\_filter* program (STACKS v0.99994). The remaining reads were trimmed by removing the first 5 bases and keeping the next 90 bases for reads 1 and keeping the first 95 bases for reads 2. The reads originating from the same population identified with different barcodes or from the same individuals run on different lanes were merged to increase coverage. We decided to discard three LSP1 and two LWP1 individuals from further analyses because their final coverage was too low, hence resulting in 75 genotyped individuals. The number of remaining PE reads for the Ind-Seq datasets varied from 996,796 to 20,480,652 with a total of ca. 304 millions (111,015,950; 81,905,824; 66,817,320 and 44,207,850 PE reads for the LSP1; LWP1; LWP2 and LSP2 individuals, respectively). Similarly, ca. 300 millions of PE reads were kept for the Pool-Seq datasets (59,692,064; 86,867,660; 91,755,446; 37,102,870 and 23,485,726, PE reads for the AWP; LSP1; LWP1; LSP2 and LWP2 pools, respectively). RADseq PE reads were then mapped against the indexed *Tpit*-SP V1 assembly (Gschloessl et al., Submitted) using the *bwa aln* and *bwa sampe* commands of the BWA 0.6.2 program (Li & Durbin, 2009) with default options to generate bam files for each of the 75 remaining individuals (38 LWP, i.e., 18 LWP1 + 20 LWP2; and 37 LSP, i.e., 17 LSP1 + 20 LSP2) and the 5 pool samples (Table 1). Between 56.11% and 67.39% RAD sequences were mapped and properly paired onto the genome for the different datasets (mean insert size: 286 bp).

#### *Generation of the Ind-Seq SNP dataset (gIS)*

The RAD Ind-Seq bam files were processed using the *mpileup* command of SAMTOOLS 0.1.19 (Li et al., 2009) and the same default options as above to obtain LWP and LSP *mpileup*

files. SNP and genotype calling were then performed separately for each of these two files using the *bcftools view* command and the resulting vcf files were merged using the *vcf-merge* program from the VCFTOOLS 0.1.12 package (Danecek et al., 2011) after filtering variants using the *vcfutils.pl varFilter* command from the SAMTOOLS suite with default options and *-w 5 -d 200*. Because of the high heterogeneity in the observed within-SNP and within-individual read coverages, we performed additional filtering steps to obtain a genotyping dataset as comprehensive as possible. First, all the genotypes with a read coverage  $DP < 5$  or  $DP > 1,000$  or a Phred quality  $GQ < 20$  were treated as missing data. The resulting number of genotype calls varied between 7,272 and 180,600 (with a median of 38,130). We thus decided to focus on the 40 individuals (28 LSP, i.e., 16 LSP1 + 12 LSP2; and 12 LWP, i.e., 10 LWP1 + 2 LWP2) with more than 35,000 genotype calls. We then discarded all the SNPs that were called on less than 90% of these 40 individuals leading to a total of 6,488 remaining SNPs. The resulting Ind-Seq genotyping dataset – hereafter referred to as gIS – had the following characteristics: (i) the individual genotyping call rate varied between 83.6% and 99.9% with a median equal to 96.3%; (ii) the individual mean read coverage varied between 10.8 and 72.1 with a median equal to 16.1; (iii) the SNP genotyping call rate varied between 92.5% and 100% with a median equal to 95.0%; and (iv) the SNP minor allele frequency varied between 0.012% and 0.5% with a median equal to 0.14%.

#### *Generation of the Pool-Seq SNP datasets (rPS and pPS)*

The five RAD Pool-Seq bam files were processed using the *mpileup* command of SAMTOOLS with default options and *-d 5000 -q 20*. The resulting file was further processed using a custom *awk* script to compute read counts for each alternative base after discarding bases with

a BAQ quality score  $< 25$ . A position was then considered as variable if (i) it had a coverage of more than 20 and less than  $c_i^{(max)}$  reads in each population  $i$ , where  $c_i^{(max)}$  represented the 95<sup>th</sup> percentile of the coverage of all positions for population  $i$ ; (ii) only two different bases were observed across all the five pools; and (iii) the minor allele was represented by at least one read in two different pool samples.

The final data read count for the Pool-Seq dataset - hereafter referred to as the rPS dataset - consisted of 58,210 SNPs with mean (median; max) coverage equal to 34.96 (32; 67) for the LWP2 pool; 53.71 (50; 105) for the LSP2 pool; 76.21 (72; 164) for the AWP pool; 122.9 (116; 244) for the LWP1 pool; and 124.6 (121; 253) for the LSP1 pool, respectively.

For applications requiring allele count data (joint PCA of individual and Pool-Seq data, computation of the SFS, see below), we used the following approach. Let  $a_{ij}$  represent the number of reads of the reference allele and  $c_{ij}$  the coverage for SNP  $i$  in population (pool)  $j$  with haploid sample size  $n_j$ . We further denote similarly  $y_{ij}$  the allele count for the reference allele in the sample and  $n_{ij}$  the haploid sample size ( $n_{ij} \leq n_j$ ) for SNP  $i$  in population (pool)  $j$ . The pPS dataset consists of the  $y_{ij}$ 's and  $n_{ij}$ 's, which were computed as follows:

$$\text{i) if } c_{ij} \leq n_j \text{ then } \hat{n}_{ij} = c_{ij} \text{ and } \hat{y}_{ij} = a_{ij}$$

$$\text{ii) if } c_{ij} > n_j \text{ then } \hat{n}_{ij} = n_j \text{ and}$$

$$\text{ii.1) if } a_{ij} = 0 \text{ or } a_{ij} = c_{ij} \text{ then } \hat{y}_{ij} = a_{ij};$$

$$\text{ii.2) if } 0 < a_{ij} < c_{ij} \text{ then } \hat{y}_{ij} = (n_j - 1) \wedge (1 \vee [n_j \times (a_{ij} / c_{ij})])$$

where  $\wedge$  and  $\vee$  stand for the maximum and the minimum, respectively. Note that formally ii) provides the maximum likelihood estimate of the  $y_{ij}$ 's under the assumption that the  $a_{ij}$ 's follow a binomial distribution  $a_{ij} \sim \text{Bin}(y_{ij} / n_j, n_{ij})$ . This approximation thus amounts in

assuming equal contribution of each individual of the pool to the Pool-Seq read data.

## **Population genetic diversity and structure**

### *Estimation of $F_{ST}$ from Pool-Seq data*

Pairwise and across populations  $F_{ST}$  were estimated using the estimator by Weir and Cockerham (1984) from the pPS data set. Even though this standard estimator was developed to measure differentiation from allele count data and therefore should be used cautiously when considering Pool-Seq data, the inherent biases are expected to be limited here given the haploid pool size, sequencing coverage and level of differentiation of the populations under study (Hivert, Gautier, & Vitalis, pers. comm.).

### *Estimation and visualisation of the scaled covariance matrix of the population allele frequencies*

To further assess the overall structuring of genetic diversity, we estimated the scaled covariance matrix of allele frequencies ( $\Omega$ ) across the five samples using the software BAYPASS (Gautier, 2015) with default options. When applied to read count data (rPS data), the Bayesian model underlying BAYPASS provides an accurate estimate of  $\Omega$  by integrating over the unobserved allele count estimation. An eigenvalue decomposition of the resulting  $\Omega$  matrix was further performed using the R function *svd()* to represent its major axis of variation. This latter approach amounts to performing a PCA that accounts rigorously for the specificities of the Pool-Seq data in the estimation of the covariance matrix.

## Joint Principal Component Analyses of individual (gIS) and pool-Seq (pPS) data

A total of 742 SNPs were in common between the individual gIS and the pPS pool datasets. We combined both datasets to obtain a matrix  $\mathbf{X} = \{x_{ij}\}$  (742 SNPs x 45 columns) of allele counts in 40 diploid individual samples ( $n_j = 2$  for  $j = 1$  to 40) and 5 pool samples ( $n_{41} = n_{AWP} = 80$  and  $n_j = n_{LSP1} = n_{LWP1} = n_{LSP2} = n_{LWP2} = 60$  for  $j = 42$  to 45) resulting in a total of 360 haploid individuals. To account for the differences in sample size, we defined a SNP weight vector  $\mathbf{w} = \{w_j\}$  where  $w_j = 1/180$  for  $j = 1$  to 40,  $w_{41} = w_{AWP} = 40/180$  and  $w_j = w_{LSP1} = w_{LWP1} = w_{LSP2} = w_{LWP2} = 30/180$  for  $j = 42$  to 45. We then computed the (observed) allele frequencies  $f_{ij} = x_{ij}/n_j$  for each SNP  $i$  and sample  $j$ , and the overall mean weighted allele frequency  $p_i = \sum_j w_j x_{ij}$  for each SNP  $i$ . Note that  $f_{ij}$  was set to  $p_i$  when  $x_{ij}$  was missing. Finally, we computed the standardized allele frequency matrix  $\mathbf{M} = \{m_{ij}\}$  where  $m_{ij} = (x_{ij} - p_i)/(p_i(1 - p_i))$ .

A weighted Principal Component Analysis (PCA) was then carried out based on the matrix  $\mathbf{M}$  and using  $\mathbf{w}$  as a (row) weight vector with the *dudi.pca()* function of the R package *ade4* (Chessel, Dufour, & Thioulouse, 2004).

## Demographic inferences

### Three-population tests of admixture

$F_3$  statistics provide a formal test for population admixture in three population trees (Patterson et al., 2012). A significantly negative  $F_3$  statistics for a ( $P1$ ;  $P2$ ,  $P3$ ) configuration supports an admixed origin of population  $P1$  with two ancestral source populations related to  $P2$  and  $P3$  respectively. Note however that the reverse is not necessarily true, e.g., the  $F_3$  statistics might not be significantly negative in this same configuration if  $P1$  experienced strong drift after the

admixture event.  $F_3$ -based tests were carried out for the possible topologies using the rPS Pool-Seq dataset. To account for the additional sampling level introduced in Pool-Seq experiments (i.e., the sampling of read sequences in the DNA pool), the following unbiased estimator relying on read count data and haploid pool sizes was used:

$$\widehat{F}_3(A, (B, C)) = \frac{1}{I} \sum_{i=0}^I [\widehat{\alpha}_i(A) + \widehat{\beta}_i(B, C) - \widehat{\beta}_i(A, B) - \widehat{\beta}_i(A, C)]$$

with:

$$\text{i) } \widehat{\alpha}_i(P) = \frac{1}{n_P - 1} \left( \frac{n_P a_{iP} (a_{iP} - 1)}{c_{iP} (c_{iP} - 1)} - \frac{a_{iP}}{c_{iP}} \right)$$

$$\text{ii) } \widehat{\beta}_i(P, Q) = \frac{a_{iP}}{c_{iP}} \frac{a_{iQ}}{c_{iQ}}$$

where, for SNP  $i$ ,  $a_{iP}$  (resp.  $a_{iQ}$ ) represents the number of reads of the reference allele and  $c_{iP}$  (resp.  $c_{iQ}$ ) the coverage in population  $P$  (resp.  $Q$ ) with haploid sample size  $n_P$ . To assess the significance of the departure of each statistic to the null hypothesis ( $F = 0$ ),  $Z$ -scores were computed as the ratio of the  $\widehat{F}_3$  mean to its standard deviation both estimated over 5,000 bootstrap samples.

### *Estimation of tree topology and divergence times using KIMTREE*

For a given tree topology, we estimated divergence times using KIMTREE 1.3 (Gautier & Vitalis, 2013), with the standard MCMC parameters recommended in the user manual. KIMTREE is a hierarchical Bayesian model where the allele frequencies are modelled along each branch of a population tree using Kimura's time-dependent diffusion approximation for genetic drift (Kimura, 1964). The support of the different topologies was assessed using a



317 Deviance Information Criterion (DIC) computed as described in Gautier and Vitalis (2013),  
 318 and up to a constant term, with slight modifications for Pool-Seq data:

$$DIC = \frac{2}{T} \sum_{t=1}^T \sum_{i=1}^I \sum_{j=1}^J -2 \log \left[ \binom{c_{ij}}{a_{ij}} \left( \frac{y_{ij}(t)}{n_j} \right)^{a_{ij}} \left( 1 - \frac{y_{ij}(t)}{n_j} \right)^{c_{ij}-a_{ij}} \right] \\ - \sum_{i=1}^I \sum_{j=1}^J -2 \log \left[ \binom{c_{ij}}{a_{ij}} \left( \frac{\widehat{y}_{ij}}{n_j} \right)^{a_{ij}} \left( 1 - \frac{\widehat{y}_{ij}}{n_j} \right)^{c_{ij}-a_{ij}} \right]$$

319 where  $y_{ij}(t)$  represents the sampled allele count value for SNP  $i$  in pool  $j$  at the  $t^{\text{th}}$  MCMC  
 320 iteration (out of  $T$ ) and  $\widehat{y}_{ij}$  the posterior mean of the corresponding allele count computed  
 321 over the  $T$  MCMC samples.

322

### 323 *Inference of complex demographic histories using FASTSIMCOAL2*

324 To explore more complex demographic scenarios, we analyzed the joint site frequency  
 325 spectrum (SFS) of the three sampled populations using the approach of Nielsen (2000)  
 326 implemented in FASTSIMCOAL2 2.5.2.21 (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, &  
 327 Foll, 2013). This approach uses coalescent simulations to infer the likelihood of the observed  
 328 SFS under any demographic model, and performs well even in situation where the events are  
 329 recent (Excoffier et al., 2013). The analyses were run on the folded SFS, i.e., using the  
 330 observed counts of the minor allele, obtained from the pPS datasets of the LSP2, LWP2 and  
 331 AWP samples (LSP2 and LWP2 were chosen based on the  $F_3$  statistics, see Result section)  
 332 with a pool size of 30, i.e., the haploid size of the smallest pool. We directly estimated the  
 333 scaled parameters of the models in a coalescent or a diffusion time scale (i.e.,  $2N\mu$ ,  $2Nm$ ,  $T/N$   
 334 and  $\mu T$ ) and then inferred the canonical parameters (divergence time  $T$  expressed in

generations, migration rates, and population size expressed in number of genes  $N$ , i.e., twice the number of diploid individuals) using the mutation rate  $\mu = 2.9 \cdot 10^{-9}$  mutations per generation per SNP, as recently estimated for the Lepidoptera *Heliconius melpomene* (Keightley et al., 2015). Note that this estimate of mutation rate is lower than that of *Drosophila melanogaster* (Haag-Liautard et al., 2007), and is expected to be more appropriate for our Lepidoptera model species. Estimated divergence times thus tend to be older than if we had used the *Drosophila* mutation rate.

We first considered a simple model of pure divergence and drift (DivDrift), which corresponds to the KIMTREE model, for the three possible topologies. This analysis allowed us to compare inferences of tree topologies and scaled divergence times obtained with two different methods, and therefore to check that the FASTSIMCOAL algorithm performed well when used with the pPS data set. We could identify the most likely topology in this simple model and then further increase step-by-step the complexity of the scenario. First we incorporated past fluctuations in population size (DivDriftVar), second we allowed migration between all populations (DivDriftMig), and third we considered both variations in population size and migration (DivDriftVarMig). All models were compared using the Akaike Information Criteria (AIC). Inferences of the canonical and/or scaled parameters were only considered for the simple DivDrift model for comparison with KIMTREE, and for the best supported model. Finally, 95% confidence intervals (CI) were built using parametric bootstrap as explained in Excoffier et al. (2013). Detailed inference settings and parameter ranges explored for each analysis are described in Appendix S1 and Table S1 (Supporting information).

# **Whole genome scans for adaptive differentiation using SELESTIM and BAYPASS**

Whole genome scans for adaptive differentiation were carried out by looking for overly differentiated SNPs using both SELESTIM version 1.1.3 (Vitalis, Gautier, Dawson, & Beaumont, 2014) and BAYPASS version 2.1 (Gautier, 2015) that are both handling Pool-Seq data. SELESTIM is based on a diffusion approximation for the distribution of allele frequencies in a subdivided population, which explicitly accounts for selection. In particular, SELESTIM assumes that each and every locus is targeted by selection to some extent, and estimates the strength of selection at each locus in each population. For each analysis, twenty-five short pilot runs (1,000 iterations each) were set to adjust the proposal distributions for each model parameter and, after a 100,000 burn-in period, 100,000 updating steps were performed with a thinning interval of 40 steps. Candidate markers under selection were selected on the basis of the distance between the posterior distribution for the locus-specific coefficient of selection and a "centering distribution" derived from the distribution of a genome-wide parameter that accounts for the among-locus variation in selection strength. SELESTIM uses the Kullback–Leibler divergence (KLD) as a distance between the two distributions, which is calibrated using simulations from a posterior predictive distribution based on the observed data (Vitalis et al., 2014). Hereafter, we report candidate markers with KLD values above the 99.9% quantile of the so-obtained empirical distribution of KLD.

In BAYPASS, we identified candidate markers using the  $XtX$  differentiation measure (Günther & Coop, 2013). This metrics might be viewed as a SNP-specific  $F_{ST}$  that explicitly corrects for the scaled covariance of population allele frequencies (matrix  $\Omega$ ), making it robust to the unknown demographic history relating the populations. The  $XtX$  was estimated using default options of BAYPASS. Pairwise correlations of the  $XtX$  estimates across ten independent runs were all found to be above 0.995 demonstrating the stability of the estimates. As described in

Gautier (2015), the  $XtX$  was calibrated based on a posterior predictive distribution obtained by analyzing a pseudo-observed dataset of 250,000 SNPs generated under the inference model with hyper-parameters fixed to their respective posterior means as estimated from the analysis of the original data. Hereafter, we report candidate markers with  $XtX$  values above the 99.9% quantile of the so-obtained empirical distribution of  $XtX$ . To identify the population of origin of the signal for overly differentiated SNPs, we examined the posterior means of the standardized population allele frequencies defined for each SNP  $i$  and population  $j$  as:

$$\mathbf{X}_i = \{x_{ij}\}_{j \in (1, \dots, J)} = \frac{1}{\sqrt{\pi_i(1 - \pi_i)}} \Gamma^{-1} \boldsymbol{\alpha}_i$$

where  $\boldsymbol{\alpha}_i$  represents the (unobserved) vector of population allele frequencies,  $\pi_i$  represents the across population allele frequency, and  $\Gamma$  the Cholesky decomposition matrix of  $\Omega$  (i.e.,  $\Omega = \Gamma\Gamma^T$ ). Although the standardized allele frequencies ( $x_{ij}$ ) are expected to be independent and identically normally distributed under the null model (Günther & Coop, 2013), the Bayesian (hierarchical) model-based estimation procedure leads to shrink their estimated posterior mean. As a result, they were each calibrated as the  $XtX$ , i.e., using their respective empirical distribution obtained from the analysis of the pseudo-observed dataset described above.

## RESULTS

### Population genetic diversity and structure

The multi-locus  $F_{ST}$  across the five samples was equal to 0.259 while pairwise population  $F_{ST}$  varied from 0.038 for the (LSP1;LSP2) pair to 0.374 for the (AWP;LSP2) pair (Table 2). The

LWP1 and LWP2 samples appeared differentiated, with a pairwise  $F_{ST}$  equal to 0.068. Nevertheless, both the LWP1 and LWP2 samples were found closer to the AWP ( $F_{ST}$  equal to 0.095 and 0.125 for the (AWP;LWP1) and (AWP;LWP2) pairs, respectively) than to the LSP ( $F_{ST}$  ranging from 0.302 to 0.368 depending on the sub-samples representing LSP and LWP).

We estimated the scaled covariance matrix of allele frequencies  $\Omega$  across the five samples using BAYPASS, and performed an eigenvalue decomposition of that matrix, which results in a principal component analysis accounting for the specificities of the Pool-Seq data. As shown in Fig. 1A, the first axis of variation (PC1) accounted for 93.5% of the total genetic variation and separated the samples according to the phenology of their underlying population (i.e., LSP1 and LSP2 vs. LWP1, LWP2 and AWP). The second axis of variation (PC2) that only accounted for 4.17% of the total variation was associated with a geographic gradient since it separated the Leiria samples (LSP1, LSP2, LWP1 and LWP2) from the Apostiça sample (AWP). Importantly, the coordinates of both the LSP1 and LWP1 samples on PC1 were found closer to the origin than their corresponding LSP2 and LWP2 counterparts. This result suggests the presence of LateSP individuals and/or hybrids in either LSP1, LWP1 or both.

This latter result was supported by the joint analysis of a subset of the Pool-Seq data together with the 28 LSP and 12 LWP genotyped individuals that had more than 35,000 genotype calls.

Fig. 1B shows the first factorial plan of a joint PCA performed on 742 SNPs that were in common between the Pool-Seq dataset (pPS) and the individual dataset (gIS). Although the number of SNPs was lower and the data for pool samples were projected onto their corresponding haploid sample size, the overall picture displayed in Fig. 1B was qualitatively similar to that of Fig. 1A. Interestingly, based on their coordinate on PC1, at least 3 out of the 12 genotyped LWP1 individuals appeared to be either LateSP or introgressed individuals.

When ignoring these 3 individuals, the coordinates of LWP individuals on PC1 were very

close to that of the LWP2 pool sample. Conversely, the PC1 coordinates for all LSP individuals remained close to those of both the LSP1 and LSP2 pool samples. All results (PCA and BAYPASS) thus suggested a higher variability across the LWP samples than across the LSP ones. They also revealed that some LateSP and introgressed individuals were included in the LWP1 pool that contained individuals that were only phenotypically assigned to their "phenological" population.

### Demographic inference

#### *F<sub>3</sub>-based tests of admixture*

Three-population tests were carried out for all the 30 possible configurations among the five pool samples (Table S2, Supporting information). Six configurations resulted in significant negative  $F_3$ -statistics. They corresponded to the four configurations that tested the LWP1 sample against another WP sample (AWP or LWP2) and a LSP sample (either LSP1 or LSP2) as source populations: (i) (LWP1; AWP, LSP1) with  $Z = -9.02$ ; (ii) (LWP1; AWP, LSP2) with  $Z = -9.45$ ; (iii) (LWP1; LSP1, LWP2) with  $Z = -17.8$ ; and (iv) (LWP1; LSP2, LWP2) with  $Z = -11.7$ . This result confirmed the inclusion of LateSP individuals in the LWP1 pool, as suggested by the PCA. The two other configurations displaying significantly negative  $F_3$ -statistics tested the LSP1 sample against the LSP2 sample and either the LWP1 or AWP as samples representative of the WP: (i) (LSP1; AWP, LSP2) with  $Z = -4.00$ ; (ii) (LSP1; LSP2, LWP1) with  $Z = -3.57$ . On the contrary, considering the LWP2 sample as representative of the LWP did not result in a significantly negative  $F_3$  (Table S2, Supporting information). In that case, the signal of admixture might be hidden by the stronger drift in LSP1, the  $F_{ST}$  of the (LWP2, LSP1) pair being higher than that of the (LWP1, LSP1) pair. We here recall that the

results of the  $F_3$  test should be interpreted only when significant. As both the PCA and  $F_3$ -statistics suggested that the LWP1 pool probably contained LateSP individuals, we further used only the LSP2 and LWP2 samples as representing the LSP and LWP to infer the demographic history of the LSP.

#### *Inferring the tree topology and divergence times under various scenarios*

We first ran KIMTREE on the Pool-Seq rPS data to compare the four possible topologies relating the AWP, LSP (using LSP2 sample) and LWP (LWP2 sample) under a pure-drift model of divergence (Fig. 2). The DIC gave the strongest support to the (LSP,(AWP,LWP)) tree (Fig. 2). Interestingly, the branch length relating the LSP to the root population (ancestral to the winter and summer populations) revealed a strong signature of drift ( $\tau_{LSP} = 0.383$ ).

We then analyzed the joint SFS of the three populations using the estimated allele count data pPS for different demographic models. Considering a simple model of divergence and drift (DivDrift), the best-supported tree according to the AIC corresponded to the best supported tree obtained with KIMTREE (Table S3, Supporting information). In the following steps, we thus only considered the topology (LSP,(LWP,AWP)) (Table 3). SFS analyses under this model lead to precise estimates of scaled parameters such as population size ratios ( $N_P / N_{WP} = 7.9$  [5.4; 9.9] and  $N_{WP} / N_{LSP} = 26$  [17; 35]), and of the four scaled divergence times for the different branches of tree, that can be directly compared to those inferred from KIMTREE and appear to be highly consistent (Fig. 2). Indeed, we estimated  $T_P / N_{LSP} = 0.35$  [0.34; 0.37] (to compare with  $\tau_{LSP} = 0.383$  in Fig. 2);  $T_{WP} / N_{LWP} = 0.085$  [0.081; 0.092] (to compare with  $\tau_{LWP} = 0.099$  in Fig. 2);  $T_{WP} / N_{AWP} = 0.089$  [0.081; 0.092] (to compare with  $\tau_{AWP} = 0.117$  in Fig. 2); and  $(T_P - T_{WP}) / N_{WP} = 0.11$  [0.094; 0.12] (to compare with  $\tau_{P4} = 0.107$  in Fig. 2). This overall

good agreement between the KIMTREE and SFS analyses suggest that estimating the SFS from the inferred allele counts (pPS dataset) provides robust results (KIMTREE analyses being based on the read count rPS dataset).

We further investigated more complex models by including migration between the populations (DivDriftMig model), variation in population sizes (DivDriftVar model) or both (DivDriftVarMig model). Comparison of AICs for these four demographic models showed that the data strongly supported the DivDriftVarMig model detailed in Fig. 3 (Table S4, Supporting information). Most of the 24 canonical parameters of this latter model, i.e., all population sizes and divergence times as well as some migration rates, were inferred with good precision (Table 3). The few exceptions concerned some migration rates for which CIs were relatively broad.

Overall, the SFS analyses suggested that the ancestral SP and WP diverged relatively recently, ca. 560 generations ago (with a confidence interval CI ranging from 448 to 2280), both experiencing a concomitant bottleneck. Then the ancestral SP experienced a first expansion, followed by a second bottleneck ca. 69 (CI = 35 - 216) generations ago, and a second strong expansion until present. From the ancestral WP, LWP and AWP diverged ca. 207 generations ago (CI = 95 - 526). Note that age estimates depend on the mutation rate used, which is lower in *Heliconius* butterflies (Keightley et al., 2015) than in *Drosophila* (Haag-Liautard et al., 2007). LWP recently experienced a relatively severe contraction while AWP showed an expansion event. Accordingly, negative growth rates (corresponding to expansions in the coalescence analyses) were inferred for all but the LWP. Finally, inferred migration rates were relatively large for the pairs (AWP,LWP) and (LSP,LWP), with especially large values for the migration from LWP to AWP and to a lesser extent from LSP to LWP. On the contrary,



inferred migration rates were lowest for the pair (LSP,AWP) as well as for the migration from SP to WP, i.e., the ancestral populations.

# **Genome-scan for adaptive differentiation**

We performed genome scans for adaptive differentiation across the three population samples (AWP, LWP2 and LSP2) using both the SELESTIM and BAYPASS software packages. Out of the 54,040 analyzed SNPs (4,170 SNPs from the original rPS dataset were discarded since monomorphic in the three analyzed population pools), 12 were found outlier with SELESTIM and 73 with BAYPASS; 11 were in common between the two analyses (Fig. 4; Fig. S1 and Table S5, Supporting information). However, we found no locus presumably involved in the phenological shift or subsequent ecological adaptation in the LSP. Indeed, among the outlier SNPs, none displayed extreme value for either the population-specific coefficient of selection estimated with SELESTIM, or the standardized allele frequencies estimated with BAYPASS in the LSP2 sample only. Instead, most outliers displayed outstanding differentiation in both the LWP and LSP (data not shown).

We then mapped the outlier SNPs onto the recently obtained draft genome (Gschloessl et al., Submitted) and used the associated gene prediction and transcriptomic resources to annotate the SNPs which fell within or near (< 2,000 pb) a potential gene. The 74 SNPs mapped to 63 different scaffolds; 7 of these SNPs were located within a gene (5 in introns of 4 different predicted or reconstructed genes, 2 in exons of 2 genes), 7 were located in the vicinity of 5 different predicted or reconstructed genes. Only two of the corresponding genes could be annotated, and corresponded to a transcription domain-associated protein of *Operophtera brumata* and an E3 ubiquitin-protein ligase RFWD2-like of the Pyralidae *Amyelois*

*transitella*. These results are detailed in Tables S5 and S6, Supporting information.

## DISCUSSION

In this study, we analyzed an illustrative example of "true allochronic differentiation" (sensu Taylor & Friesen, 2017) between sympatric populations of the pine processionary moth. Our results allowed us to decipher when and how the primary divergence occurred (bottleneck intensity, levels of gene flow), which allows us to propose hypotheses about the circumstances of the differentiation and the subsequent history of the populations.

### **The primary divergence: a fairly recent allochronic event associated to a strong bottleneck and an abrupt disruption of gene flow**

Tree-based analyses suggested that the phenologically-shifted SP first diverged from the common ancestor of the two studied WPs, which differentiated more recently from one another. The long branch leading to the SP suggested that this population experienced very strong drift. The model was significantly improved by including changes in population sizes and migration between populations, suggesting that the demographic history associated with the allochronic event is relatively complex. We could infer in detail this evolutionary scenario. The common ancestor of the SP and the WP is supposed to have been present for a long time (estimated to 900,000 years), with large population sizes ( $10^5$  to  $10^6$  reproducing individuals), which is consistent with the continuous occurrence of the pine processionary moth in the

refugial areas of the Iberian Peninsula during the Ice Ages (Rousset et al., 2010). The divergence of the SP was estimated to have occurred ca. 560 years ago, and it was associated with a very strong founder event (ancestral population size estimated to a few tens of individuals), while a bottleneck event occurred in the ancestral WP. One of the main questions about sympatric differentiation is to know whether it occurred in the presence or absence of gene flow in the first steps of the divergence, and how migration evolved over time (Powell, Forbes, Hood, & Feder, 2014; Smadja & Butlin, 2011). The question of the levels of gene flow can shed light on the differentiation process and impact the possible fate of the diverging populations. In the particular case of allochronic differentiation, the shift in breeding time can occur progressively, an overlap in reproductive times of the incipient populations then maintaining gene flow (with some similarities between isolation-by-distance and isolation-by-time in this case, see Hendry & Day, 2005). Conversely, it can also appear as an abrupt phenological change that would immediately disrupt gene flow and lead to an "automatic" complete assortative mating, acting as an "automatic magic trait" *sensu* Servedio, Doorn, Kopp, Frame & Nosil (2011). Our results showed that the first step of the differentiation occurred in a context of highly limited gene flow between the ancestral SP and WP (migration rate  $10^{-5}$  to  $10^{-8}$ ). This corroborates the hypothesis of a sudden event of divergence, resulting in an immediate barrier to gene flow between the two incipient populations.

Allochrony can in some situations evolve as a by-product of another primary driver of speciation, such as host plant shift followed by alteration of breeding time to match with the new host's phenology (Powell et al., 2014). It is not possible from our results to rule out the hypothesis that the two populations primarily diverged due to other factors, and that allochrony evolved more recently and would now be the main differentiated trait. On the other hand, no host or habitat change is associated with the differentiation of the SP. The land

was once covered by mixed forests and shrubs (AFN - Autoridade Florestal Nacional, 2012), and then sowed with *P. pinaster* during the XIII<sup>th</sup> and early XIV<sup>th</sup> century. The divergence of the SP probably occurred after this large afforestation program, which took place ca. 700 years ago, when *P. pinaster* was already predominant in this region. We thus conclude that in the particular case of the pine processionary moth, allochrony can still be hypothesized to be the initial driver of divergence. It is very likely that the periods of adult activity of the two diverging populations did not overlap in the early phases of their differentiation, immediately disrupting gene flow. A scenario of an initial mutation in key genes involved in seasonal rhythms or affecting diapause termination which first occurred by chance and drove the differentiation event in a very limited number of founder individuals can thus be favoured (Schluter, 2009), and would be consistent with the high heritability found in experimental rearing (Branco et al., 2017). This information is crucial for our understanding of the allochronic differentiation process.

We obtained a relatively recent estimate of the divergence time, but our results suggest that the SP was already present few hundreds years before its discovery in 1997 (Santos, Burban, et al., 2011; Santos et al., 2007). No mention was found in the historical archives of the Mata Nacional de Leiria (MB, pers. obs.), even though these archives contain much information because the national park has been a major wood production area for more than seven centuries. Yet, it is also possible that the ancestral SP evolved in the same region, but outside the limits of the park, and remained undocumented in historical times. In a recent study, Godefroid and collaborators (2016) showed that the current distribution of the LSP is limited by the high summer temperature occurring elsewhere in Portugal, even though larvae of this population were proved to cope better with higher temperatures than larvae of Portuguese and French WPs (Santos, Paiva, et al., 2011). In the first steps of the differentiation, milder

environmental conditions could have favoured the success of the diverging population. Interestingly, a period of colder climate known as the Little Ice Age occurred between years 1300 and 1900, including in Portugal, bringing favourable climatic conditions (Abrantes et al., 2005; Bartels-Jónsdóttir, Knudsen, Abrantes, Lebreiro, & Eiríksson, 2006). Other phenotypic trait divergences between the SP and the WP were documented, with obvious adaptations to the environmental changes experienced by the SP eggs and larvae due to the shift in breeding time (Rocha et al., 2017; Santos, Paiva, et al., 2011; Santos, Paiva, Rocha, Kerdelhué, & Branco, 2013), consistent with the concept of "adaptation-by-time" proposed by Hendry and Day (2005). Whether such phenotypic changes occurred over ca. 500 years or whether they occurred over some tens of generations as previously suggested (Santos, Burban, et al., 2011; Santos et al., 2007), these adaptations can still be considered as rapid.

**Recent demographic changes in the diverging population and increased recent gene flow**

The best demographic model we obtained further suggested that the SP experienced a recent bottleneck ca. 70 years ago, which reduced the population to a few hundred reproducers at most. The SP then expanded again until its high current population size (between 25,000 and 100,000 individuals). The cause of this recent and drastic reduction in size is difficult to characterize, and could be due to a local climatic or epidemiological event or to human activities (e.g., local habitat destruction, forest fire, management options). This bottleneck actually coincides with the recent establishment of intensive planning and forest management in the MNL. The first forest plan dates back from 1892 and was intensified during the 1960s, including management by clear-cuts and development of 120 km of forest roads (AFN - Autoridade Florestal Nacional, 2012). This major demographic event is consistent with the

fact that the SP remained undetected in the recent history and was discovered only recently during a very severe and thus conspicuous outbreak in 1997 (Pimentel et al., 2006; Santos et al., 2007). Parallel to the SP history, our model also suggested a complex scenario for the studied WPs. AWP and LWP diverged ca. 200 years ago, with a very strong founder event in Apostiça as the estimated population size reached 43 individuals only. This event could be linked to human activities and to the deforestation process that occurred to provision wood and agricultural goods, which dramatically decreased forest land in the region of Lisbon (Devy-Vareta, 1985). This probably tended to fragment the PPM habitat and strongly reduced its populations in the vicinity of Lisbon. It is worth noting that in the recent years, the population size has strongly increased in Apostiça, which is consistent with the recent afforestation activity, whereas the Leiria WP tended to decrease. Whether the decline of the WP observed in the MNL could be linked to possible competition between the sympatric summer and winter populations should now be tested. Monitoring tools could moreover allow us to determine if this is a long-term trend or if the local LWP would increase again. On the other hand, our results consistently show that AWP was closely related to LWP, and could not be used as an outgroup as we initially planned. A thorough phylogeographic study of Portuguese and/or Iberian populations would now be helpful to understand the genetic structure of populations in this PPM clade (Rousselet et al., 2010) and to develop further demographic analyses.

To complete the picture, our results suggested that some gene flow currently occurs between existing populations. Not surprisingly, in the best demographic model, migration rates were maximal between the two WPs but they were also relatively high in both directions between the two sympatric LSP and LWP (ca.  $10^{-3}$ ). This is consistent with the recent identification of few hybrid individuals by Burban and collaborators (2016). Interestingly, our results suggest

that the level of gene flow between the sympatric populations is higher today than in earlier stages of differentiation. This could be explained by the recent geographic and demographic expansion observed in the SP, which could have increased the probability of contact and thus introgression between the two populations. We could also hypothesize that plasticity in reproductive time plays a role by allowing some degree of overlap in reproductive time between the two populations, which can possibly vary over time as occurs in some plants (Devaux & Lande, 2008). Some individuals belonging to the SP but emerging during the LWP reproductive season were recently identified with molecular markers (Burban et al., 2016). Such "LateSP" individuals can only be identified through genotyping, and could also allow some introgression between the two populations. The results further showed that assigning individuals from their phenology alone can lead to erroneous mixing of some LateSP individuals in the LWP1 pool, and that robust results could only be obtained when pooling genetically well-characterized individuals. Preliminary observations suggest that some of these LateSP correspond to the last-emerging SP individuals, i.e., to events at the tail-end of the distribution of SP emergence time in July, during the early WP season. Other LateSP actually emerge very late, after the WP season, and could correspond to a dysfunction in diapause termination (Burban et al., 2016). The origins and fate of these categories of LateSP remain to be studied.

**Identifying and interpreting signatures of selection**

All of the SNPs identified by BAYPASS and SELESTIM as presumably under selection displayed population-specific signatures associated with both the LSP and the LWP, which did not allow us to clearly identify a pattern linked to the phenotypic evolution of the SP. It is

likely that the strong drift experienced by the SP and the high level of differentiation between the SP and both LWP and AWP ( $F_{ST} > 0.3$ ) impedes optimal use of genomic scans of adaptation. A similar challenge in revealing functionally important loci due to a stronger than expected background differentiation was encountered by Lozier, Jackson, Dillon, & Strange (2016) in their study of *Bombus* colour patterns. Moreover, even if RAD-seq was proved to be a powerful approach to easily develop population genomic studies for non-model organisms, the technique only allows us to analyze a reduced proportion of the genome, which increases the likelihood of missing the genomic region truly targeted by selection (Lowry et al., 2017; but see McKinney, Larson, Seeb, & Seeb, 2017). Our study also pointed a major challenge in arthropod genomics, which is the low proportion of functionally annotated genes. We could annotate only two of the genes in the vicinity of the detected candidate SNP, which strongly limits the functional interpretation of the results. Moreover, the draft genome currently available for *T. pityocampa* has low scaffold sizes (Gschloessl et al., Submitted), which explains why most of the identified SNP were found in different genomic fragments. Improving the genome assembly will greatly increase our analyzing capacities.

### **Perspectives and future directions**

Several studies have recently identified candidate genes involved in circadian and seasonal rhythms and in diapause termination, and their roles and interactions are increasingly understood (Denlinger, 2002; Derks et al., 2015; Wadsworth & Dopman, 2015). In particular, there is increasing evidence that genes involved in circadian rhythms are also involved in reproductive cycles (Fuchikawa et al., 2010; Levy, Kozak, Wadsworth, Coates, & Dopman, 2015; Ragland, Egan, Feder, Berlocher, & Hahn, 2011; Ragland & Keep, 2017). One possible



We thank H. Santos for sampling the PPM individuals analyzed in this paper. We are grateful to B. Gschloessl who facilitated access to the genome assembly and provided the SNP annotations, and to A. Estoup for fruitful discussions in the early steps of the project. The authors also wish to thank Pierre-Alexandre Gagnaire for the RAD adaptors and H  l  ne Holota from TAGC facility for the Covaris sonication of the RAD libraries.

31

assistance of A. Dehne- Garcia. Edinburgh Genomics is funded in part by grants from the UK Natural Environment Research Council (R8/H10/56) and Medical Research Council (G0900740). The work in Portugal was partially funded by the FCT national project UID/AGR/00239/2013.

## REFERENCES

- Abrantes, F., Lebreiro, S., Rodrigues, T., Gil, I., Bartels-Jonsdottir, H., Oliveira, P., . . . Grimalt, J. O. (2005). Shallow-marine sediment cores record climate variability and earthquake activity off Lisbon (Portugal) for the last 2000 years. *Quaternary Science Reviews*, 24, 2477–2494. doi:10.1016/j.quascirev.2004.04.009
- AFN - Autoridade Florestal Nacional (2012). *Estratégia para a gestão das matas nacionais*. Lisboa, Portugal: Relatório Direção Nacional de Gestão Florestal.
- Alem, S., Streiff, R., Courtois, B., Zenboudji, S., Limousin, D., & Greenfield, M. D. (2013). Genetic architecture of sensory exploitation: QTL mapping of female and male receiver traits in an acoustic moth. *Journal of Evolutionary Biology*, 26(12), 2581–2596. doi:10.1111/jeb.12252
- Alexander, R. D., & Bigelow, R. S. (1960). Allochronic speciation in field crickets, and a new species, *Acheta veletis*. *Evolution*, 14(3), 334–346.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), e3376. doi:10.1371/journal.pone.0003376
- Bartels-Jónsdóttir, H. B., Knudsen, K. L., Abrantes, F., Lebreiro, S., & Eiríksson, J. (2006). Climate variability during the last 2000 years in the Tagus Prodelta, western Iberian Margin: benthic foraminifera and stable isotopes. *Marine Micropaleontology*, 59, 83–103. doi:10.1016/j.marmicro.2006.01.002
- Battisti, A., Holm, G., Fagrell, B., & Larsson, S. (2011). Urticating hairs in arthropods: their nature and medical significance. *Annual Review of Entomology*, 56(1), 203–220. doi:10.1146/annurev-ento-120709-144844
- Battisti, A., Larsson, S., & Roques, A. (2017). Processionary moths and associated urtication risk: global change-driven effects. *Annual Review of Entomology*, 62(1), 323–342. doi:10.1146/annurev-ento-031616-034918
- Berardi, L., Branco, M., Paiva, M.-R., Santos, H., & Battisti, A. (2015). Development time plasticity of the pine processionary moth (*Thaumetopoea pityocampa*) populations under laboratory conditions. *Entomologia*, 3(1), 273. doi:10.4081/entomologia.2015.273
- Berdan, E. L., Mazzoni, C. J., Waurick, I., Roehr, J. T., & Mayer, F. (2015). A population genomic scan in *Chorthippus* grasshoppers unveils previously unknown phenotypic divergence. *Molecular Ecology*, 24(15), 3918–3930. doi:10.1111/mec.13276
- Branco, M., Paiva, M.-R., Santos, H., Burban, C., & Kerdelhué, C. (2017). Experimental evidence for heritable reproductive time in 2 allochronic populations of pine processionary moth. *Insect Science*, 24(2), 325–335. doi:10.1111/1744-7917.12287
- Burban, C., Gautier, M., Leblois, R., Landes, J., Santos, H., Paiva, M.-R., . . . Kerdelhué, C.

- (2016). Evidence for low-level hybridization between two allochronic populations of the pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Biological Journal of the Linnean Society*, 119(2), 311-328. doi:10.1111/bij.12829
- Catchen, J., Amores, A., Hohenlohe, P. A., Cresko, W., & Postlethwait, J. (2011). *Stacks*: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, 1, 171-182. doi:10.1534/g3.111.000240
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). *Stacks*: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124-3140. doi:10.1111/mec.12354
- Chessel, D., Dufour, A.-B., & Thioulouse, J. (2004). The ade4 package - I : One-table methods. *R News*, 4(1), 5-10.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. doi:10.1093/bioinformatics/btr330
- Davey, J. W., & Blaxter, M. L. (2011). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5), 416- 423. doi:10.1093/bfgp/elq031
- Denlinger, D. L. (2002). Regulation of diapause. *Annual Review of Entomology*, 47, 93-122.
- Derkarabetian, S., Burns, M., Starrett, J., & Hedin, M. (2016). Population genomic evidence for multiple Pliocene refugia in a montane-restricted harvestman (Arachnida, Opiliones, *Sclerobunus robustus*) from the southwestern United States. *Molecular Ecology*, 25(18), 4611-4631. doi:10.1111/mec.13789
- Derks, M. F. L., Smit, S., Salis, L., Schijlen, E., Bossers, A., Mateman, C., . . . Megens, H.-J. (2015). The genome of winter moth (*Operophtera brumata*) provides a genomic perspective on sexual dimorphism and phenology. *Genome Biology and Evolution*, 7(8), 2321-2332. doi:10.1093/gbe/evv145
- Devau, C., & Lande, R. (2008). Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, 275(1652), 2723-2732. doi:10.1098/rspb.2008.0882
- Devy-Vareta, N. (1985). Para uma geografia histórica da floresta portuguesa. *Revista da Faculdade de Letras-Geografia*, 1, 47-67.
- Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *Plos Genetics*, 9(10), e1003905. doi:10.1371/journal.pgen.1003905
- Franchini, P., Fruciano, C., Spreitzer, M. L., Jones, J. C., Elmer, K. R., Henning, F., & Meyer, A. (2014). Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. *Molecular Ecology*, 23(7), 1828-1845. doi:10.1111/mec.12590
- Friesen, V. L., Smith, A. L., Gomez-Diaz, E., Bolton, M., Furness, R. W., Gonzalez-Solis, J., & Monteiro, L. R. (2007). Sympatric speciation by allochrony in a seabird. *Proceedings of the National Academy of Sciences of the USA*, 104(47), 18589-18594. doi:10.1073/pnas.0700446104
- Fuchikawa, T., Sanada, S., Nishio, R., Matsumoto, A., Matsuyama, T., Yamagishi, M., . . . Miyatake, T. (2010). The clock gene cryptochrome of *Bactrocera cucurbitae* (Diptera: Tephritidae) in strains with different mating times. *Heredity*, 104, 387-392. doi:10.1038/hdy.2009.167
- Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, 201(4), 1555-1579.

- doi:10.1534/genetics.115.181453
- Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., . . . Estoup, A. (2013). Estimation of population allele frequencies from next-generation data: pooled versus individual genotyping. *Molecular Ecology*, 22(14), 3766–3779. doi:10.1111/mec.12360
- Gautier, M., & Vitalis, R. (2013). Inferring population histories using genome-wide allele frequency data. *Molecular Biology and Evolution*, 30(3), 654–668. doi:10.1093/molbev/mss257
- Godefroid, M., Rocha, S., Santos, H., Paiva, M.-R., Burban, C., Kerdelhué, C., . . . Rossi, J.-P. (2016). Climate constrains range expansion of an allochronic population of the pine processionary moth. *Diversity and Distributions*, 22(12), 1288–1300. doi:10.1111/ddi.12494
- Gschloessl, B., Dorkeld, F., Berges, H., Beydon, G., Bouchez, O., Branco, M., . . . Kerdelhué, C. (Submitted). Draft genome and reference transcriptomic resources for the urticating pine defoliator *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Molecular Ecology Resources*, MER-17-0372.
- Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1), 205–220. doi:10.1534/genetics.113.152462
- Guo, B., DeFaveri, J., Sotelo, G., Nair, A., & Merilä, J. (2015). Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology*, 13(1), 19. doi:10.1186/s12915-015-0130-8
- Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature*, 445, 82–85. doi:10.1038/nature05388
- Hasselmann, M., Ferretti, L., & Zayed, A. (2015). Beyond fruit-flies: population genomic advances in non-*Drosophila* arthropods. *Briefings in Functional Genomics*, 14(6), 424–431. doi:10.1093/bfpg/elv010
- Hendry, A. P., & Day, T. (2005). Population structure attributable to reproductive time: isolation by time and adaptation by time. *Molecular Ecology*, 14(4), 901–916. doi:10.1111/j.1365-294X.2005.02480.x
- Hohenlohe, P. A. (2014). Ecological genomics in full colour. *Molecular Ecology*, 23(21), 5129–5131. doi:10.1111/mec.12945
- Huchon, H., & Démolin, G. (1970). La bioécologie de la processionnaire du pin. Dispersion potentielle - Dispersion actuelle. *Revue Forestière Française*, 22, 220–234.
- Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., . . . Jiggins, C. D. (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution*, 32(1), 239–243. doi:10.1093/molbev/msu302
- Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1, 177–232.
- Levy, R. C., Kozak, G. M., Wadsworth, C. B., Coates, B. S., & Dopman, E. B. (2015). Explaining the sawtooth: latitudinal periodicity in a circadian gene correlates with shifts in generation number. *Journal of Evolutionary Biology*, 28(1), 40–53. doi:10.1111/jeb.12562
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25, 1754–1760. doi:10.1093/bioinformatics/btp324
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and

- SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352
- Limborg, M. T., Waples, R. K., Seeb, J. E., & Seeb, L. W. (2014). Temporally isolated lineages of pink salmon reveal unique signatures of selection on distinct pools of standing genetic variation. *Journal of Heredity*, 105(6), 835-845. doi:10.1093/jhered/esu063
- Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142-152. doi:10.1111/1755-0998.12635
- Lozier, J. D., Jackson, J. M., Dillon, M. E., & Strange, J. P. (2016). Population genomics of divergence among extreme and intermediate color forms in a polymorphic insect. *Ecology and Evolution*, 6(4), 1075-1091. doi:10.1002/ece3.1928
- McCoy, R. C., Garud, N. R., Kelley, J. L., Boggs, C. L., & Petrov, D. A. (2014). Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular Ecology*, 23(1), 136-150. doi:10.1111/mec.12591
- McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, 17(3), 356-361. doi:10.1111/1755-0998.12649
- Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154, 931-942.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3), 1065-1093. doi:10.1534/genetics.112.145037
- 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. *Forest Ecology and Management*, 233(1), 108-115. doi:10.1016/j.foreco.2006.06.005
- Powell, T. H. Q., Forbes, A. A., Hood, G. R., & Feder, J. L. (2014). Ecological adaptation and reproductive isolation in sympatry: genetic and phenotypic evidence for native host races of *Rhagoletis pomonella*. *Molecular Ecology*, 23(3), 688-704. doi:10.1111/mec.12635
- Ragland, G. J., Egan, S. P., Feder, J. L., Berlocher, S. H., & Hahn, D. A. (2011). Developmental trajectories of gene expression reveal candidates for diapause termination: A key life-history transition in the apple maggot fly *Rhagoletis pomonella*. *Journal of Experimental Biology*, 214, 3948-3960. doi:10.1242/jeb.061085
- Ragland, G. J., & Keep, E. (2017). Comparative transcriptomics support evolutionary convergence of diapause responses across Insecta. *Physiological Entomology*, 42(3), 246-256. doi:10.1111/phen.12193
- Rocha, S., Kerdelhué, C., Ben Jamaa, M. L., Dhahri, S., Burban, C., & Branco, M. (2017). Effect of heat waves on embryo mortality in the pine processionary moth. *Bulletin of Entomological Research*, 107(5), 583-591. doi:10.1017/S0007485317000104
- Rodríguez-Mahillo, A. I., González-Muñoz, M., Vega, J. M., López, J. A., Yart, A., Kerdelhué, C., . . . Moneo, I. (2012). Setae from the pine processionary moth (*Thaumetopoea pityocampa*) contain several relevant allergens. *Contact Dermatitis*, 67(6), 367-374. doi:10.1111/j.1600-0536.2012.02107.x
- Rosser, N. L. (2015). Asynchronous spawning in sympatric populations of a hard coral reveals cryptic species and ancient genetic lineages. *Molecular Ecology*, 24, 5006-5019. doi:10.1111/mec.13372

- Rosser, N. L. (2016). Demographic history and asynchronous spawning shape genetic differentiation among populations of the hard coral *Acropora tenuis* in Western Australia. *Molecular Phylogenetics and Evolution*, 98, 89-96. doi:10.1016/j.ympev.2016.02.004
- Rousselet, J., Zhao, R., Argal, D., Simonato, M., Battisti, A., Roques, A., & Kerdelhué, C. (2010). The role of topography in structuring the demographic history of the pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae). *Journal of Biogeography*, 37(8), 1478-1490. doi:10.1111/j.1365-2699.2010.02289.x
- Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, 8, 336-352. doi:10.1111/j.1461-0248.2004.00715.x
- Santos, H., Burban, C., Rousselet, J., Rossi, J.-P., Branco, M., & Kerdelhué, C. (2011). Incipient allochronic speciation in the pine processionary moth *Thaumetopoea pityocampa* (Lepidoptera, Notodontidae). *Journal of Evolutionary Biology*, 24(1), 146-158. doi:10.1111/j.1420-9101.2010.02147.x
- Santos, H., Paiva, M.-R., Rocha, S., Kerdelhué, C., & Branco, M. (2013). Phenotypic divergence in reproductive traits of a moth population experiencing a phenological shift. *Ecology and Evolution*, 3(15), 5098-5108. doi:10.1002/ece3.865
- Santos, H., Paiva, M.-R., Tavares, C., Kerdelhué, C., & Branco, M. (2011). Temperature niche shift observed in a Lepidoptera population under allochronic divergence. *Journal of Evolutionary Biology*, 24(9), 1897-1905. doi:10.1111/j.1420-9101.2011.02318.x
- 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. *Proceedings of the Royal Society of London Series B*, 274(1612), 935-941. doi:10.1098/rspb.2006.3767
- Savolainen, V., Anstett, M.-C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., . . . Baker, W. J. (2006). Sympatric speciation in palms on an oceanic island. *Nature*, 441, 210-213. doi:10.1038/nature04566
- Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323, 737-741. doi:10.1126/science.1160006
- Servedio, M. R., Doorn, G. S. V., Kopp, M., Frame, A. M., & Nosil, P. (2011). Magic traits in speciation: magic but not rare? *Trends in Ecology & Evolution*, 26(8), 389-397. doi:10.1016/j.tree.2011.04.005
- Smadja, C. M., & Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, 20(24), 5123-5140. doi:10.1111/j.1365-294X.2011.05350.x
- Sota, T., Yamamoto, S., Cooley, J. R., Hill, K. B. R., Simon, C., & Yoshimura, J. (2013). Independent divergence of 13- and 17-y life cycles among three periodical cicada lineages. *Proceedings of the National Academy of Sciences of the USA*, 110(17), 6919-6924. doi:10.1073/pnas.1220060110
- Szulkin, M., Gagnaire, P.-A., Bierne, N., & Charmantier, A. (2016). Population genomic footprints of fine-scale differentiation between habitats in Mediterranean blue tits. *Molecular Ecology*, 25(2), 542-558. doi:10.1111/mec.13486
- Taylor, R. S., & Friesen, V. L. (2017). The role of allochrony in speciation. *Molecular Ecology* 26(13), 3330-3342. doi:10.1111/mec.14126
- Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, 196, 799-817. doi:10.1534/genetics.113.152991
- Wadsworth, C. B., & Dopman, E. B. (2015). Transcriptome profiling reveals mechanisms for the evolution of insect seasonality. *The Journal of Experimental Biology*, 218(22), 3611-

3622. doi:10.1242/jeb.126136  
Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358-1370.  
Weis, A. E., Winterer, C., Vacher, C., Kossler, T. M., Young, C. A., & LeBuhn, G. L. (2005). Phenological assortative mating in flowering plants: the nature and consequences of its frequency-dependence. *Evolutionary Ecology Research*, 7, 161–181. doi:10.1111/j.1420-9101.2005.00891.x  
Yamamoto, S., Beljaev, E. A., & Sota, T. (2016). Phylogenetic analysis of the winter geometrid genus *Inurois* reveals repeated reproductive season shifts. *Molecular Phylogenetics and Evolution*, 94(Pt A), 47-54. doi:10.1016/j.ympev.2015.08.016  
Yamamoto, S., & Sota, T. (2009). Incipient allochronic speciation by climatic disruption of the reproductive period. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1668), 2711-2719. doi:10.1098/rspb.2009.0349  
Yamamoto, S., & Sota, T. (2012). Parallel allochronic divergence in a winter moth due to disruption of reproductive period by winter harshness. *Molecular Ecology*, 21(1), 174-183. doi:10.1111/j.1365-294X.2011.05371.x

**DATA ACCESSIBILITY**

Read count data for the five pool samples and individual genotyping datasets are provided in Supplementary Table S7

**AUTHOR CONTRIBUTIONS**

*Conceived and designed the study: M.Gau., C.K., J.F., M.B., K.G. Performed the wet lab experiments: J.F., M.Gal., L.S., A.L., C.B. Analyzed the data: A.R., M.Gau., R.V., R.L., J.F. Wrote the paper: M.Gau., C.K., R.L. All authors read and approved the final manuscript.*

## FIGURE LEGENDS

**Figure 1:** Principal component analysis of gene frequencies across the AWP, LSP1, LSP2, LWP1 and LWP2 pool samples. A) Plot of the pool sample coordinates on the first two axes of variation of  $\Omega$ , the scaled covariance matrix of allele frequencies across the five pool samples. The matrix  $\Omega$  was estimated with BAYPASS (Gautier, 2015) using read count data (rPS) available on 58 210 SNPs. B) First factorial plan of the joint PCA performed on the projected allele count data (pPS) for the five pool samples together with genotyping data for 28 LSP and 12 LWP individuals. The combined dataset consisted of 742 SNPs.

**Figure 2:** Comparisons under a pure-drift divergence model of three bifurcating tree (A-C) and a star phylogeny (D), relating the AWP, LSP (represented by the LSP2 pool sample) and LWP (represented by the LWP2 pool sample) using KIMTREE (Gautier & Vitalis, 2013). The tree with the highest support (smallest DIC) is represented in red and corresponds to ((AWP,LWP),LSP). For that tree, the posterior mean of the divergence time (measured on a diffusion time scale) is provided for each branch.

**Figure 3:** Graphical representation of the best supported demographic and historical model inferred using the SFS analyses for the three populations AWP, LWP and LSP. Both time and population sizes are represented on a log scale. Inferred parameter values are given in Table 3. SP is the ancestral population of LSP; WP is the ancestral population of AWP and LWP; P is the ancestral population of SP and WP; ANC is the ancestral population of P. All populations have undergone past exponential variations in size, except the ANC population that had a constant size through time. Because of the logarithm representation of time and population sizes, these past exponential population size variations appear linear on the graphic. Arrows represent migration from one population to another, their thickness being proportional to the inferred migration rates.

**Figure 4:** Results of the genome scans for adaptive differentiation performed with SELESTIM and BAYPASS. For each SNP the KLD estimated with SELESTIM that quantifies to which extent the locus-specific coefficient of selection is extreme is plotted against the  $XtX$  measure of differentiation estimated with BAYPASS. The horizontal (resp. vertical) dotted line indicates the 0.1% significance thresholds for the KLD (resp.  $XtX$ ) analysis that was determined as the 99.9% quantile of an empirical distribution obtained after analyzing a pseudo observed dataset simulated under the SELESTIM (resp. BAYPASS) null model. According to these thresholds, SNPs are displayed in red (outliers based on both the KLD and  $XtX$ ), in blue (outlier based on the KLD only), in green (outlier based on the  $XtX$  only) or in black (not outlier).



**SUPPORTING INFORMATION**

**Table S1:** Parameter ranges explored in the SFS analyses for the DivDrift and DivDriftVarMig models.

**Table S2:** Results of the  $F_3$ -statistics for the 30 possible configurations among the five pool samples.

**Table S3:** Comparisons of the three models of pure divergence based on their AIC computed from SFS analyses.  $d$  is the number of parameters of the model,  $L$  is the likelihood,  $AIC$  is  $2d - 2\ln(L)$ ,  $\Delta i$  is  $AIC_i - \min(AIC_i)$ , and  $w_i$  is the model normalized relative likelihood computed as  $\exp(-0.5\Delta i) / \sum_k \exp(-0.5\Delta k)$ . Parentheses indicate the most recent branching in the tree going backward in time.

**Table S4:** Comparisons of the four demographic models analyzed (DivDrift, DivDriftMig, DivDriftVar, DivDriftVarMig) based on their AIC computed from SFS analyses. For this table, the population tree considered is always (LSP,(LWP,AWP)). See main text for details about the different models and Table S3 for details about the notations.

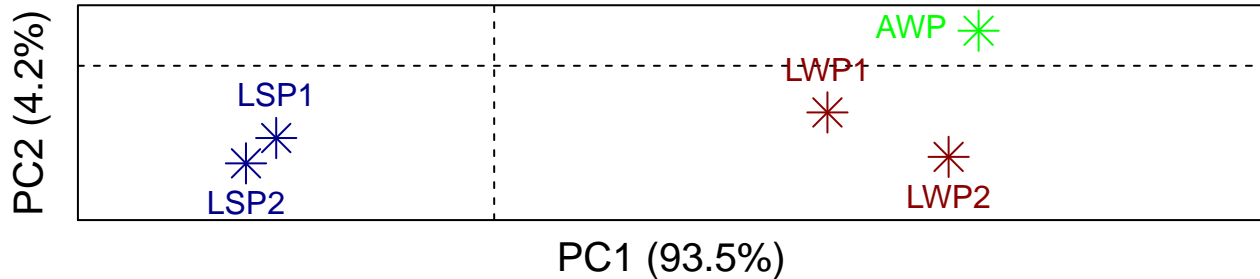
**Table S5:** Details of the 74 SNPs identified as outliers in SELESTIM and/or BAYPASS analyses. For each SNP, the scaffold and position are given together with the KLD (SELESTIM analysis) and XtX (BAYPASS analysis) values. Only values exceeding the 0.1% significance threshold are reported for the latter.

**Table S6:** Details of the 63 scaffolds containing the 74 SNPs identified as outliers in SELESTIM and/or BAYPASS analyses, position from potentially identified gene when relevant (within intron, within exon, < 2000 bp) and annotation when available.

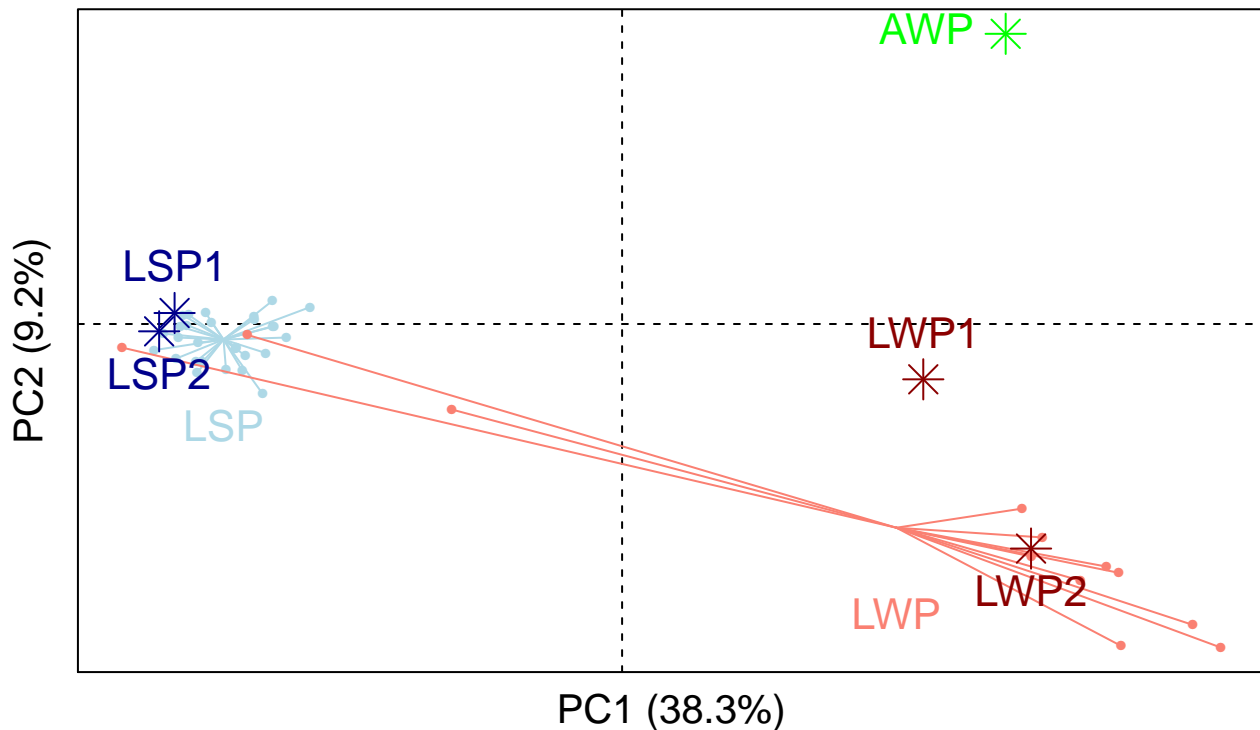
**Table S7:** Read count data for the five pool samples and individual genotyping datasets

**Figure S1:** SNP population-specific coefficient of selection and standardized allele frequencies. For each SNP and population,  $\sigma_{\max}$  that corresponds to the largest coefficient of selection among the two estimated by SELESTIM (one for each allele) is plotted against the standardized allele frequencies for the reference allele (given in absolute) as estimated by BAYPASS. For the latter, the vertical dotted line indicates the 99.9% quantile of the corresponding empirical distribution (from the pseudo-observed dataset). The colour code used to represent the SNPs is the same as in Fig. 4: SNPs are displayed in red (outliers based on both the KLD and XtX), in blue (outlier based on the KLD only), in green (outlier based on the XtX only) or in black (not outlier).

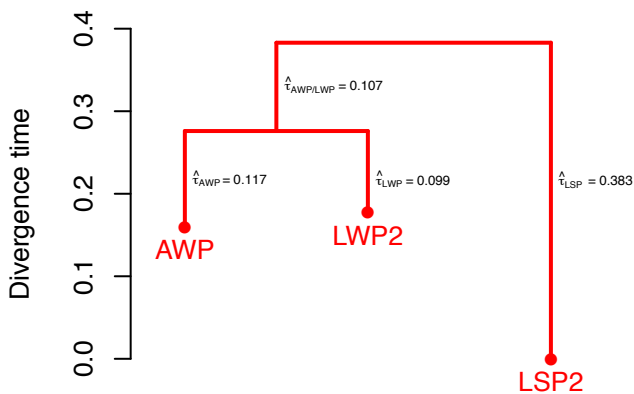
**Appendix S1:** Inference settings for the SFS analyses.

A) Eigen-Analysis of  $\hat{\Omega}$ 

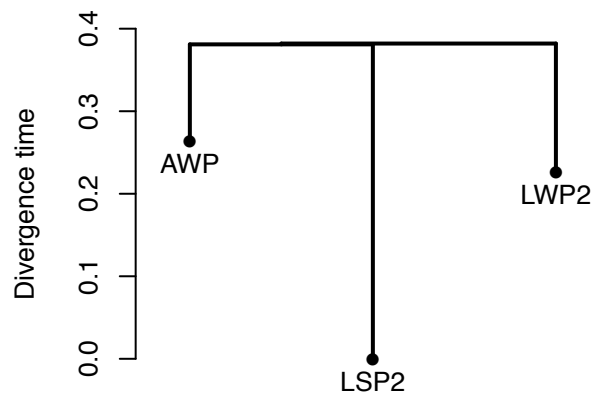
## B) Joint PCA of ind. and pool data



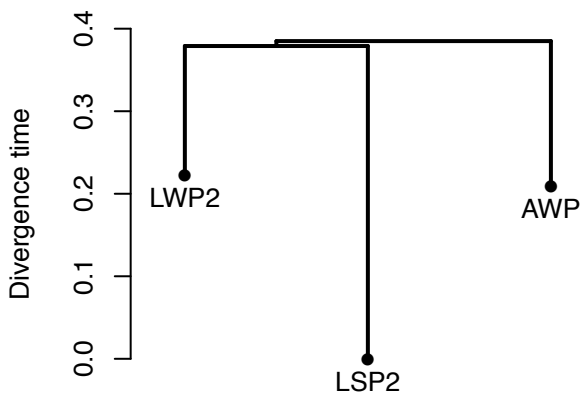
A. Topology T1

 $DIC_{T1} = 560,684$ 

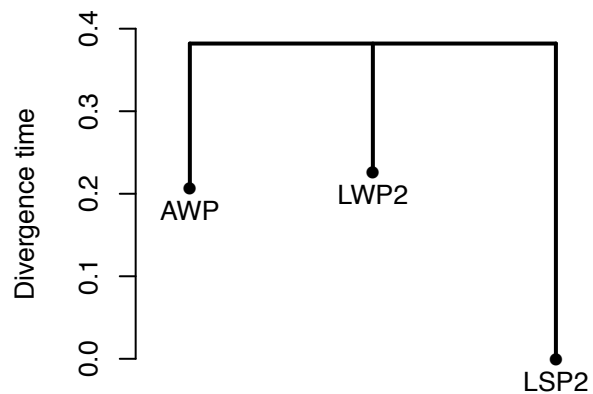
B. Topology T2

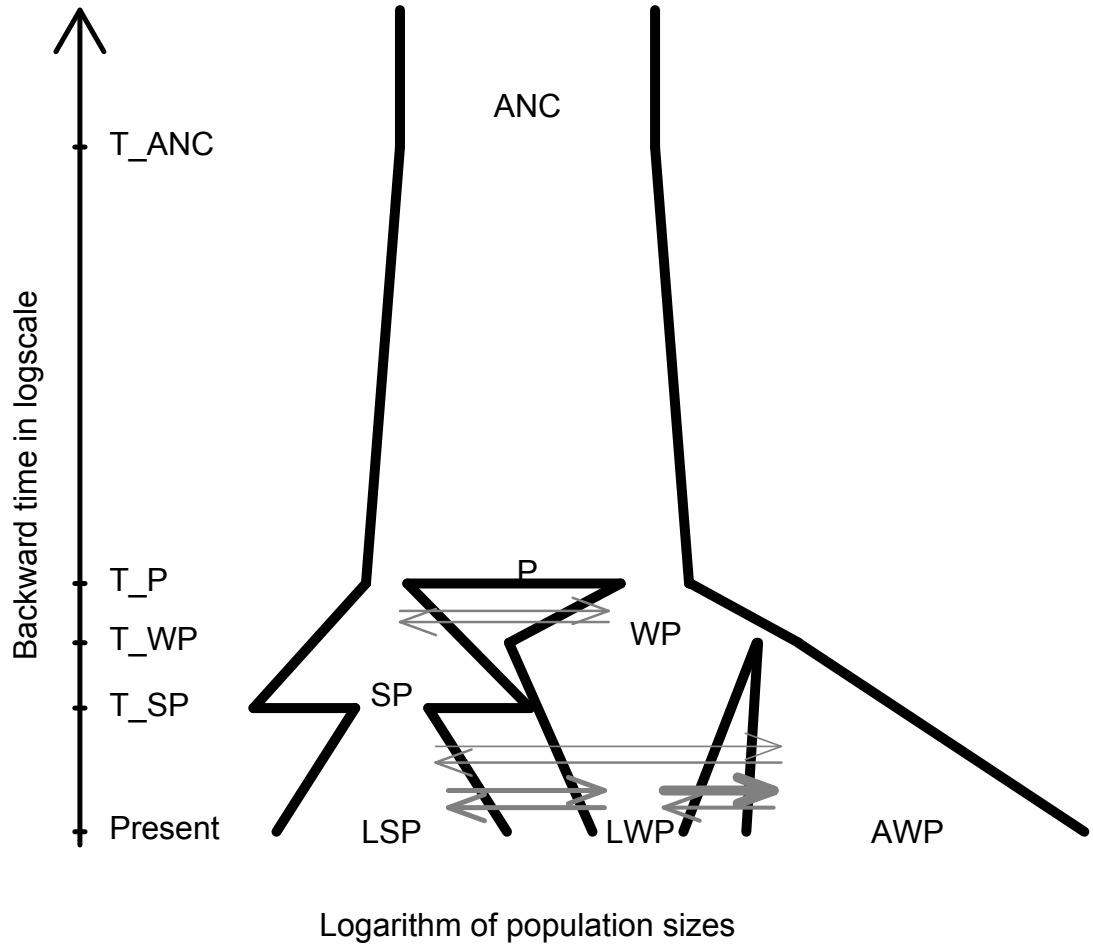
 $DIC_{T2} - DIC_{T1} = 209$ 

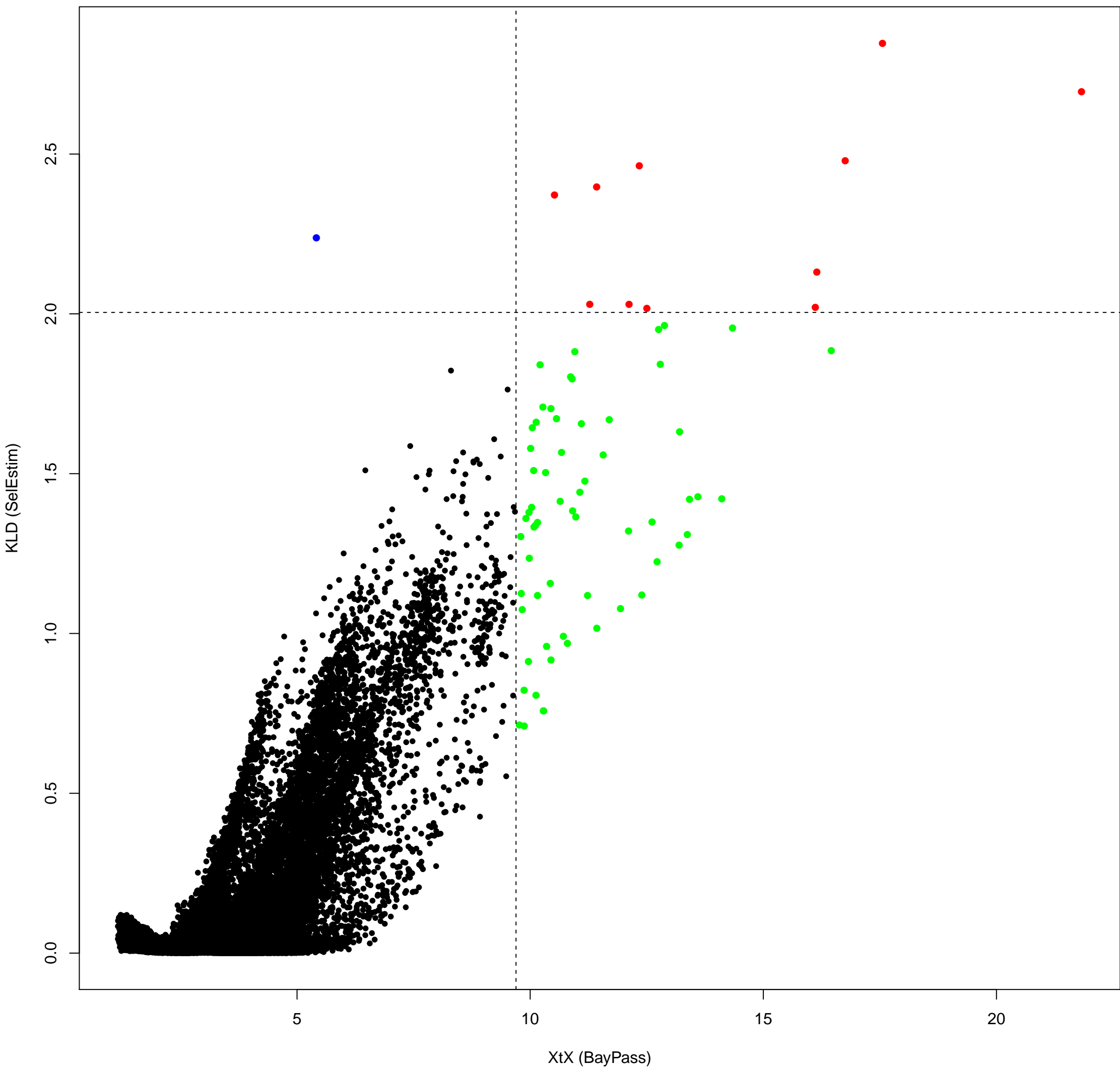
C. Topology T3

 $DIC_{T3} - DIC_{T1} = 236$ 

D. Topology S

 $DIC_S - DIC_{T1} = 283$





**Table 1:** Sampling details for the 5 batches of individuals used in the different experiments.

Code	Population	# Individuals from each stage	Dates	Experiment
LSP2	LSP	20 males	2008 – 2010	Individual and Pool RAD-seq
	LSP	10 males	2008 – 2010	Pool RAD-seq
LSP1	LSP	10 males and 10 females	2010	Individual and Pool RAD-seq
	LSP	20 males	2010	Pool RAD-seq
LWP2	LWP	20 males	2008 – 2010	Individual and Pool RAD-seq
	LWP	10 males	2008 – 2010	Pool RAD-seq
LWP1	LWP	10 males and 10 females	2010	Individual and Pool RAD-seq
	LWP	20 L5 larvae	2010	Pool RAD-seq
AWP	Apostiça	40 L5 larvae	2010	Pool RAD-seq

**Table 2:** pairwise  $F_{ST}$  estimates between the analyzed pools.

	AWP	LSP1	LSP2	LWP1
AWP				
LSP1	0.369			
LSP2	0.374	0.038		
LWP1	0.095	0.302	0.307	
LWP2	0.125	0.368	0.362	0.068

**Table 3:** Parameter point estimates and associated confidence intervals obtained for the best-supported demographic history (DivDriftVarMig with the population tree (LSP,(AWP,LWP))) from the SFS analysis. See Fig. 3 for details.

Type of parameter	Demographic Parameter	“DivDrift” Model	“DivDriftVarMig” Model
Effective population size	$N_{AWP}(0)$	2330 [2240;1.64x10 <sup>4</sup> ]	4.40x10 <sup>6</sup> [2.85x10 <sup>6</sup> ;5.16x10 <sup>6</sup> ]
	$N_{LSP}(0)$	2.16x10 <sup>5</sup> [2.07x10 <sup>4</sup> ;4.05x10 <sup>5</sup> ]	6.63x10 <sup>4</sup> [2.48x10 <sup>4</sup> ;1.63x10 <sup>5</sup> ]
	$N_{LWP}(0)$	3790 [2930;1.55x10 <sup>4</sup> ]	293 [129;828]
	$N_{SP}(T_{SP})$	n.a.	3.97x10 <sup>5</sup> [1.23x10 <sup>5</sup> ;8.62x10 <sup>5</sup> ]
	$N_{WP}(T_{WP})$	7.16x10 <sup>5</sup> [6.64x10 <sup>4</sup> ;1.38x10 <sup>6</sup> ]	4.97x10 <sup>5</sup> [1.33x10 <sup>5</sup> ;1.28x10 <sup>6</sup> ]
	$N_P(T_P)$	5.60x10 <sup>6</sup> [5.43x10 <sup>5</sup> ;8.53x10 <sup>6</sup> ]	2.44x10 <sup>6</sup> [2.17x10 <sup>6</sup> ;2.53x10 <sup>6</sup> ]
	$N_P(T_{ANC})$	n.a.	1.85x10 <sup>5</sup> [9.83x10 <sup>4</sup> ;2.17x10 <sup>5</sup> ]
	$N_{AWP}(T_{WP})$	n.a.	43 [17;121]
	$N_{LSP}(T_{SP})$	n.a.	147 [52;741]
	$N_{LWP}(T_{WP})$	n.a.	1.42.10 <sup>5</sup> [2.65x10 <sup>4</sup> ;4.85x10 <sup>5</sup> ]
	$N_{SP}(T_P)$	n.a.	43 [35;257]
	$N_{WP}(T_P)$	n.a.	129 [103;974]
Divergence time (in generation)	$T_{SP}$	n.a.	69 [35;216]
	$T_{WP}$	216 [190;1380]	207 [95;526]
	$T_P$	7.59x10 <sup>4</sup> [6.90x10 <sup>3</sup> ;1.38x10 <sup>5</sup> ]	560 [448;2.28x10 <sup>3</sup> ]
	$T_{ANC}$	n.a.	9.05.10 <sup>5</sup> [8.79x10 <sup>5</sup> ;1.10x10 <sup>6</sup> ]
Migration rate	$m(LSP \rightarrow AWP)$	n.a.	5.36x10 <sup>-7</sup> [2.95x10 <sup>-8</sup> ;9.38x10 <sup>-6</sup> ]
	$m(LWP \rightarrow AWP)$	n.a.	5.15x10 <sup>-3</sup> [1.62x10 <sup>-3</sup> ;1.04x10 <sup>-2</sup> ]
	$m(AWP \rightarrow LSP)$	n.a.	2.40x10 <sup>-8</sup> [2.67x10 <sup>-8</sup> ;9.86x10 <sup>-7</sup> ]
	$m(LWP \rightarrow LSP)$	n.a.	9.69x10 <sup>-4</sup> [3.10x10 <sup>-4</sup> ;1.91 x10 <sup>-3</sup> ]
	$m(AWP \rightarrow LWP)$	n.a.	4.29x10 <sup>-4</sup> [4.91x10 <sup>-6</sup> ;1.10x10 <sup>-3</sup> ]
	$m(LSP \rightarrow LWP)$	n.a.	1.11x10 <sup>-3</sup> [3.46x10 <sup>-4</sup> ;2.53x10 <sup>-3</sup> ]
	$m(WP \rightarrow SP)$	n.a.	2.29x10 <sup>-5</sup> [2.34x10 <sup>-8</sup> ;2.55x10 <sup>-5</sup> ]
	$m(SP \rightarrow WP)$	n.a.	8.72x10 <sup>-8</sup> [2.99x10 <sup>-8</sup> ;8.50x10 <sup>-6</sup> ]
Growth rate	$G_{AWP}$	n.a.	-0.480 [-1.10;-0.174]
	$G_{LSP}$	n.a.	-0.764 [-1.89;-0.175]
	$G_{LWP}$	n.a.	0.258 [8.60x10 <sup>-2</sup> ;0.551]
	$G_{SP}$	n.a.	-0.160 [-0.201;-2.79x10 <sup>-2</sup> ]
	$G_{WP}$	n.a.	-0.201 [-0.247;-0.029]
	$G_P$	n.a.	-2.46x10 <sup>-5</sup> [-2.56x10 <sup>-5</sup> ;-2.22x10 <sup>-5</sup> ]