

Host plant shifts and the emergence of new pest insects: the case of the weevil genus *Phlyctinus*.

Léo Laborieux

5 *Under the supervision of R. Allio, J. Haran & C. Perrier (CBGP, France)*

Word count: 4028 (<101%)

Keywords: genomics, phytosanitary pests, host plant shifts, species delimitation, phylogenetics, weevil

10

15

20

Abstract | Although insect host-plant shifts are a major threat to modern agriculture, they generally consist in ancient or sporadic events that remain difficult to investigate and hence insufficiently understood. Herein, we explored the genomic consequences of recent and parallel shifts undergone by some populations in two closely related species in the pest weevil genus *Phlyctinus*, namely *Phlyctinus callosus* Schönherr and *Phlyctinus xerophilus* Haran, endemic to South Africa. We first used whole-genome resequencing (WGR) to investigate the species status of cryptic lineages, their level of admixture and surveyed potential genome-wide signatures of positive selection in the lineages that underwent a shift from wild hosts (Asteraceae) to non-native crops (vine and apples). Our results provide a first analysis of the complex genetic structure in these pests, and highlight the presence of both formal cryptic species and admixture between divergent lineages. We further identified loci showing important differentiation in populations that shifted from wild hosts. These results are discussed in light of ecologically relevant traits for these species.

I. Introduction

25

30

35

40

45

The Cape Floristic Region of the Republic of South Africa (henceforth referred to as the CFR and South Africa, respectively) is a well-recognized biodiversity hotspot (Myers et al., 2000; Mittermeier et al., 2005), with around 9 000 plant species, approximately 70% of which are endemic (Goldblatt & Manning, 2002). Recent work suggests that the CFR's arthropod biodiversity, which was shaped by a unique interplay of climatic and geological events, may be comparably high (Matenaar et al, 2018), calling for strong research efforts in the area. In particular, weevils (Coleoptera: Curculionidea) are remarkably diverse within the CFR, with a high abundance of endemic and often undescribed taxa. This is particularly true of subfamily Entiminae, which is well-represented in the region and has been shown to comprise numerous endemic lineages (Alonso-Zarazaga & Lyal, 1999). Accordingly, recent revision of the phytophagous entimine genus *Phlyctinus* Schönherr has shed new light on this formerly monotypic group, with the description of five new, endemic species (Haran et al., 2020), although the status of some lineages remains uncertain. This is notably the case for *P. callosus* Schönherr and *P. xerophilus* Haran, which have been theorized to represent species complexes on the basis of morphological and molecular data (Haran et al., 2020; Hévin et al., 2022; J. Haran, *in preparation*). Remarkably, some lineages within the *P. xerophilus* and *P. callosus* clades have recently (after the rise of large-scale agriculture in the area, i.e. <200 ybp, see Lounsbury (1896)) undergone a parallel host-plant shift from their natural hosts (Asteraceae) to non-native cultivated plants such as grapevine and deciduous fruit trees (Haran et al., 2020). The uncertain taxonomic makeup of both pest species poses a significant challenge for the development of management practices (Bortolus, 2008), and this situation constitutes a worrying prospect, both for local business owners and for the international agricultural industry. The economic impact of these beetles is significant, and while recent data are unavailable, the South African agricultural industry reported an estimated 3 million USD of yearly damage (corrected for inflation) during the 1990s (Barnes & Capatos,

1989). In this context, ascertaining the taxonomic status of the lineages comprising the *P. xerophilus* and *P. callosus* clades is critical to our understanding of this problematic genus.

Species delimitation still remains, however, a significant challenge to modern evolutionary biology; one made particularly difficult by the nature of the speciation process. Indeed, the separation of populations into separate species is a continuous phenomenon, and diverging biological entities may still show evidence of gene flow, both genome-wide and at specific loci (Nosil et al., 2009). This so-called “grey zone” of speciation (Roux et al., 2016), which lies at the interface between intra (population genetics) and inter(phylogenetics) species differentiation, has been the focus of much species delimitation research. In recent years, the ever-increasing availability of molecular data enabled new powerful frameworks for the delineation of species. This led to the successful identification of numerous cryptic species spanning from invertebrates to, perhaps more surprisingly, megafauna such as giraffes (Fennessy et al., 2016), great apes (Nater et al., 2017) or, more recently, felids (Allio et al., 2021). Hévin et al. (2022) attempted to delineate species within the genus *Phlyctinus* using a handful of mitochondrial and nuclear markers, and while the status of several lineages was clarified, their results remained partly inconclusive regarding *P. callosus* and *P. xerophilus*. New approaches using genomic data can help disentangle the taxonomic identity of separate lineages, which is a prerequisite to investigating the evolutionary context of hosts shifts in these species.

The parallel putative shifts observed in populations of two closely related species of *Phlyctinus* seems to be an exemplary case of “evolutionary replicates” providing an opportunity to investigate the genomic signature of these (human induced ?) evolutionary processes. Host-plant shifts are a well-documented phenomenon known to be a key mechanism of pest emergence (Bernal & Medina, 2018). Phylogenomics have also repeatedly shown plant-insect coevolution to be a key driver of insect diversification (e.g. Ehrlich & Raven, 1964; Edger et al., 2015; Sahoo et al., 2017; Condamine et al., 2018). This link between host shifts and insect diversification has been acknowledged for decades (Diehl & Bush, 1984), and previous studies on swallowtail butterflies (Lepidoptera: Papilionidae) have started to uncover genomic signatures of host plant shifts, such as a remarkable diversification of certain gene families (Berenbaum et al., 1996), gene and genome duplications (Edger et al., 2015) and, more recently, evidence for genome-wide positive selection (Allio et al., 2021). These major new insights are well consolidated by several studies in Lepidoptera and Hemiptera, but this research has primarily focused on ancient evolutionary events that span entire insect families, leaving the small-scale processes involved in incipient host plant shifts largely unexplored (but see Hood et al. (2020)).

The taxonomically complex *Phlyctinus* weevils represent both a critical problem to international agriculture and a valuable opportunity to explore the early signatures of host-plant shifts at the genomic scale. Herein, we use whole-genome resequencing (WGR) data of multiple *Phlyctinus* lineages currently recognized as *P. callosus* and *P. xerophilus* to assess their taxonomic level and, for the first time, conduct exploratory genome scans for signatures of positive selection in both congeneric species. We expect that both species may contain cryptic taxa and that, in shifted lineages, some ecologically-relevant regions will exhibit clear signatures of selection, reflecting the recent evolutionary pressures exerted by adaptation to new hosts.

100 II. Materials and methods.

110 II.1 | Available specimens

105 An extensive sampling campaign was conducted in the CFR between 2017 and 2020, resulting in the collection of ~1 000 *Phlyctinus* specimens, all stored within the CIRAD collection, Montpellier, France (CBGP <https://doi.org/10.15454/D6XAKL>). These specimens, collected in 60 localities throughout the CFR, add up to a representative sample of the diversity of the genus, later used in Hévin et al. (2020). Some of this material serves as the basis of our study design. One individual belonging to the shifted lineage of *P. callosus* was used for whole-genome sequencing (WGS). Individuals belonging to four geographically-delimited populations of each pest species —two wild, two shifted— were selected for whole-genome resequencing. In total, 41 individuals belonging to six putative lineages were resequenced (**Table 1**). These lineages were designated based on morphological, ecological and molecular characteristics (Hévin et al., 2022; J. Haran, *in preparation*).

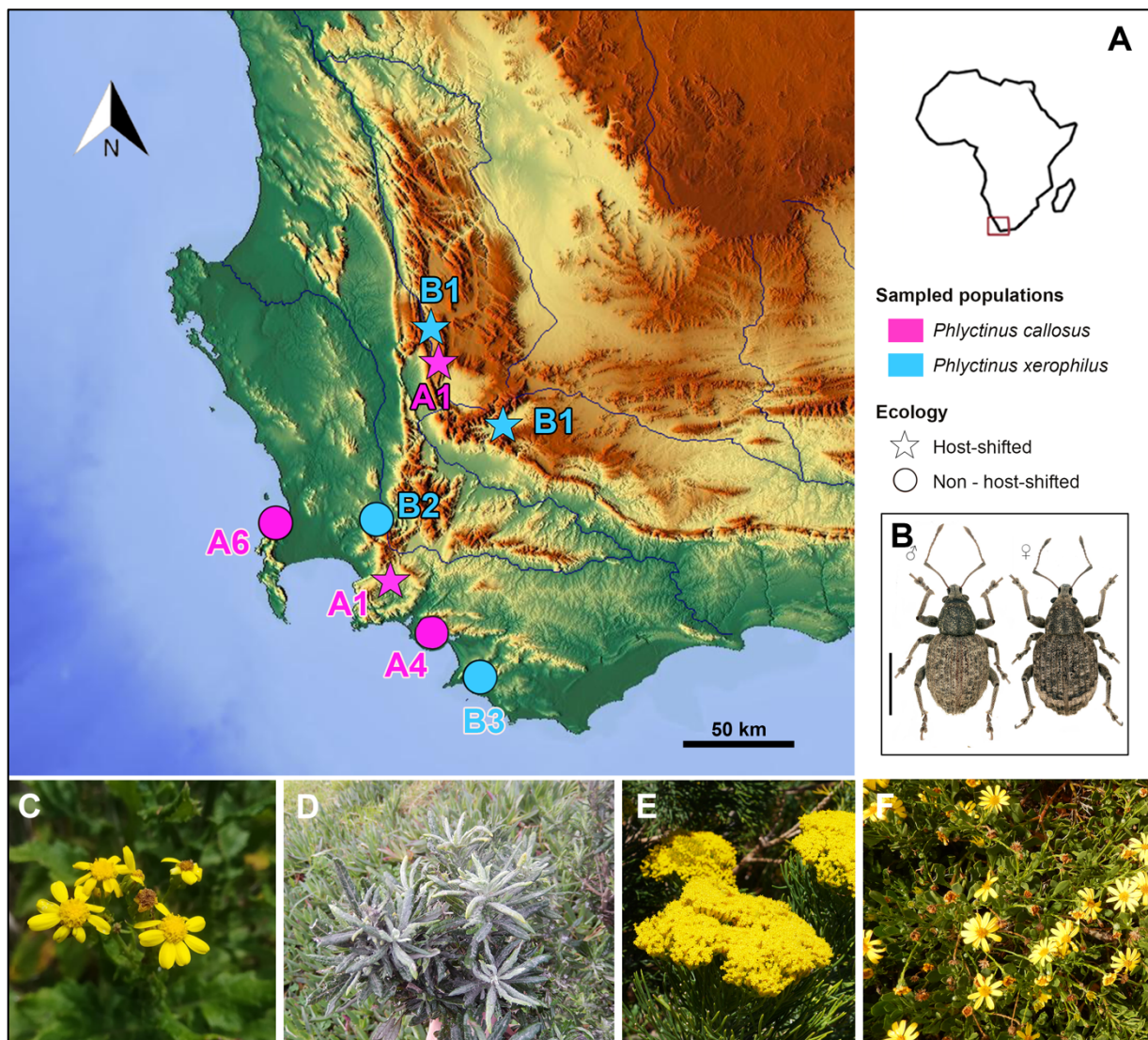


Figure 1 | Distribution map of sampled populations and corresponding host plants. A : topographic map of the CFR and sampled localities. B : *Phlyctinus callosus*, adult pair. C-F : host plants of non-shifted populations. *Phlyctinus xerophilus* : *Senecio rigidus* (C) and *Senecio pterophorus* (D). *Phlyctinus callosus* : *Hymenolepis* sp. (E) and *Osteospermum moniliferum* (F). Photos : J. Haran

120 Table 1 | Whole-genome resequenced specimens included in this study.. Lineages labeled "A" belong to *P. callosus*.
Lineages labeled "B" belong to *P. xerophilus*. DFT: deciduous fruit tree.

Species	Putative lineage	Number of individuals	Host plant (* shifted)
<i>P. callosus</i>	A1	9	Grapevine/DFT*
	A4	3	<i>Osteospermum moniliferum</i>
	A6	5	<i>Hymenolepis sp.</i>
	Total	17	
<i>P. xerophilus</i>	B1	10	Grapevine/DFT*
	B2	9	<i>Senecio pterophorus</i>
	B3	5	<i>Senecio rigidus</i>
	Total	24	

125 II.2 | Whole-genome resequencing and SNP genotyping of several lineages belonging to *Phlytinus callosus* and *Phlyctinus xerophilus*

The weevils to be resequenced underwent a non-destructive DNA extraction protocol (Biobasic kit) followed by Illumina library construction and sequencing (2x150bp). Preprocessing of reads was performed with Trimmomatic V.0.39 (Bolger et al., 2014) and assessed with FastQC V.0.12.0 (Andrews, 2010).

130 **Nuclear dataset.** Alignment to a *P. callosus* reference genome (de-novo assembly by Rémi Allio, CBGP, France) was performed using the Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). The cleaned files were then converted, sorted and indexed using SAMtools (Danecek et al., 2021). BCFtools V.1.17 (Danecek et al., 2021) was used for variant calling on all 41 individuals. Finally, SNPs were filtered using VCFtools (Danecek et al., 2011) with custom parameters (`--max-missing`, `--min-depth`, `--max-depth`, `--maf` and `--min-qual`) selected on a per-analysis basis (see section II.4 for details; information summarized in Table 2).

135 **Mitochondrial dataset.** The mitochondrial genomes of each specimen were assembled and annotated separately using MitoFinder V.1.4 (Allio et al., 2020) in its MEGAHIT mode (Li et al., 2016). The resulting sequences were then run through BLAST (Johnson et al., 2008) to check for contamination. Regions corresponding to the 15 mitochondrial genes were concatenated for all individuals and aligned using MAFFT (Katoh et al., 2002) and MACSE (Ranwez et al., 2018).

145 II.3 | Phylogenetic reconstruction, population structure and species delimitation analyses

Assessing the taxonomic level of independent lineages is a complex task that was carefully approached throughout this project. As such, we combined a diverse array of methods to avoid analysis-specific biases.

150 **Construction of molecular phylogenies.** The nuclear dataset (**Dataset A**) consisted in a VCF file filtered at low stringency (`--max-missing 0.75`, `--min-depth 5`, `--max-depth 25`, `--maf 0.05` and `--min-qual 32`) and randomly subset to 1.5M SNPs to decrease computational load. Maximum-likelihood phylogenies were generated with W-IQ-TREE V.2.2.2.6 (Trifinopoulos et al., 2016) for both the mitochondrial and nuclear datasets. The concatenated mitochondrial dataset was divided a priori into 41 partitions: three for each coding gene (one per codon position), and one for each non-coding gene. An automatic substitution model selection approach was adopted (best partition scheme while allowing for partition merges, `-m MFP+MERGE`). In both analyses, tree node supports were assessed with 1000 ultrafast bootstraps (`-bb 1000`).

Genetic divergence. Using the same filtered **Dataset A**, estimates of genetic divergence were calculated pairwise between putative lineages, at the intraspecific and interspecific levels. Species-level divergence has been investigated in past studies (~10% divergence between *P. callosus* and *P. xerophilus* using mitochondrial data, Hévin et al., 2022), and is expected to be significantly higher than intraspecific divergence. Leveraging this property can allow for the detection of overly divergent lineages that may warrant full species status. Because relative measures of divergence such as F_{st} are heavily impacted by within-population genetic diversity (Cruickshank & Hahn, 2014), absolute measures of divergence such as dXY were also computed. Calculations were performed in R using the Popgenome package (Pfeifer et al., 2014).

Population structure. In an exploratory step, a genetic principal components analysis (PCA) was conducted using Plink V.1.90 (Chang et al., 2015) in order to assess population structure based on a stringently filtered nuclear dataset to avoid biases induced by missing data. For this data, the raw VCF file was filtered with options `--max-missing 0.95`, `--min-depth 5`, `--max-depth 25`, `--maf 0.05` and `--min-qual 32` and one individual (SHAN00033_0133) had to be excluded entirely due to poor statistics (**Dataset B**). In order to further investigate structuration within the two clades, sNMF V.2.0 (Frichot et al., 2014) analyses were conducted for K values ranging from 2 to 8 on the same dataset, with default parameters. For each value of K, the quality of ancestry estimation was evaluated using a cross-entropy criterion as implemented in the package. Last, gene flow between putative populations was estimated using TreeMix V.1.13 (Pickrell & Pritchard, 2012).

Sensu-stricto species delimitation analyses. Making use of several, methodologically distinct analyses is crucial to the success of species delimitation, since agreement between multiple analyses is the best way to ensure robustness (Carstens et al., 2013). We therefore used bPTP (Zhang et al., 2013, <https://species.h-its.org/>) on the inferred nuclear phylogeny, as a final means of assessing the taxonomic level of the six lineages. bPTP is a variant of the Poisson Tree Processes model that implements the assignment of bayesian support values to the delineated taxa.

II.4 | Genome scans and search for signatures of positive selection

In an exploratory approach, F_{st} values between the shifted and non-shifted populations were computed genome-wide, on a per-site basis, for each species (VCFtools `--weir-fst-pop` tool) (Weir & Cockerham, 1984). For this purpose, the unfiltered VCF file containing all variant information was split into two species files, which subsequently underwent medium-stringency filtering with equal parameters (`--max-missing 0.85`, `--min-depth 5`, `--max-depth 25`, `--maf 0.05` and `--min-qual 32`, **Dataset C**). SNPs whose F_{st} values placed in the 95-percentile of the distribution were considered outliers and retained for both datasets. The positions of outlier SNPs that appeared in both species were then extracted. Analysis of linkage-disequilibrium decay patterns (**Fig.S1**) led us to extract nucleotide sequences in 10kb windows, centered around these positions, from the reference genome. This resulted in 10kb sequences that were later run through BlastX V.2.14.0 (Camacho et al., 2009) to check for homology to known proteins.

III. Results

III.1 | Whole-genome sequence alignments and genetic variant information

Summary information for the genomic alignments and the generated datasets is available in **Table 2**.

Table 2 : Summary information for all datasets used in this study. Filters are reported as VCFtools command-line arguments.

	Alignments	Raw VCF	Analysis-ready VCF		
			Dataset A	Dataset B	Dataset C*
Individual count	41	41	41	40	40
Subset from raw	—	—	Yes	No	No
Filters	—	—	--max-missing 0.75, --min-depth 5, --max-depth 25, --maf 0.05, --min-qual 32	--max-missing 0.95, --min-depth 5, --max-depth 25, --maf 0.05, --min-qual 32	--max-missing 0.85, --min-depth 5, --max-depth 25, --maf 0.05, --min-qual 32
Number of variants	—	143 109 837	282 476	95 595	22 393 793
Number of contigs	15392	—	—	—	—
N50	183 335	—	—	—	—

III.2 | Updated molecular phylogenies for *Phlyctinus callosus* and *Phlyctinus xerophilus*

Phylogenetic analysis of the nuclear and mitochondrial datasets returned well-supported maximum-likelihood trees with slightly different topologies (**Fig.2**, phylograms available in **Fig.S2**). The mitochondrial phylogeny recovers the monophyly of all the *a priori* defined putative lineages in *P. callosus* with maximum or near-maximum support (100%, 100% and 98% for lineages A1, A6 and A4 respectively). Lower-level structuration within these clades seems to be low, with short branches and more poorly-resolved nodes. Interestingly, this is not the case in *P. xerophilus*, in which lineage B1 is recovered as polyphyletic, forming a complete clade with the B2 lineage, while B3 is recovered as monophyletic with maximum support. The clade formed by lineages B1 and B2, however, is very poorly supported (two nodes with 59% and 73% support, respectively). Lower-level structuration within lineages is slightly more prevalent than what can be observed in the *P. callosus* lineages, though often moderately supported. One B2 individual (JHAR02057_0106) displays an artifactually large genetic distance, likely due to poor sequencing. The nuclear phylogeny, on the other hand, displays a different and better-supported topology that recovers all putative lineages as monophyletic with maximum support (100%). In the *P. xerophilus* clade, lineage B3 is the most divergent with the longest branch length, and displayed as a sister group to lineage B1. In the *P. callosus* clade, lineage A4 is the most divergent, occupying the most basal node, and with the longest branch length. Lineage B2 is subdivided into two previously undetected and distinct sub-groups, hereafter referred to as B2.1 and B2.2, with maximal support.

III.3 | Population structure, gene flow and species delimitation on nuclear data

Within both species, analysis of pairwise genetic divergence hints at high differentiation between putative lineages. One noteworthy observation is the clear contrast in the distribution patterns of F_{st} and d_{XY} values across all pairwise comparisons (**Fig.3**). F_{st} values are notably high between most populations, even at the intraspecific level, and interspecific F_{st} values are the highest ($F_{st} > 0.70$). Conversely, absolute divergence varies across lineage pairs by half an order of magnitude, with true intermediates (e.g. B1/B2, with $d_{XY} = 3,24 \cdot 10^{-9}$ and $F_{st} = 0,397$). This discrepancy may be due to demography, considering that F_{st} is sensitive to effective population size and that some lineages may have been heavily impacted by founder effects. Overall, all relative and absolute divergence measures calculated between *P. callosus* and *P. xerophilus* are high, while between-lineage divergence inside the clades is variable. Interestingly, lineage B3 is heavily differentiated from lineages B1 and B2 (B1/B3: $d_{XY} = 6,26 \cdot 10^{-9}$, $F_{st} = 0,689$; B1/B2: $d_{XY} = 4,48 \cdot 10^{-9}$, $F_{st} = 0,668$), a result consolidated by the concordance of both metrics.

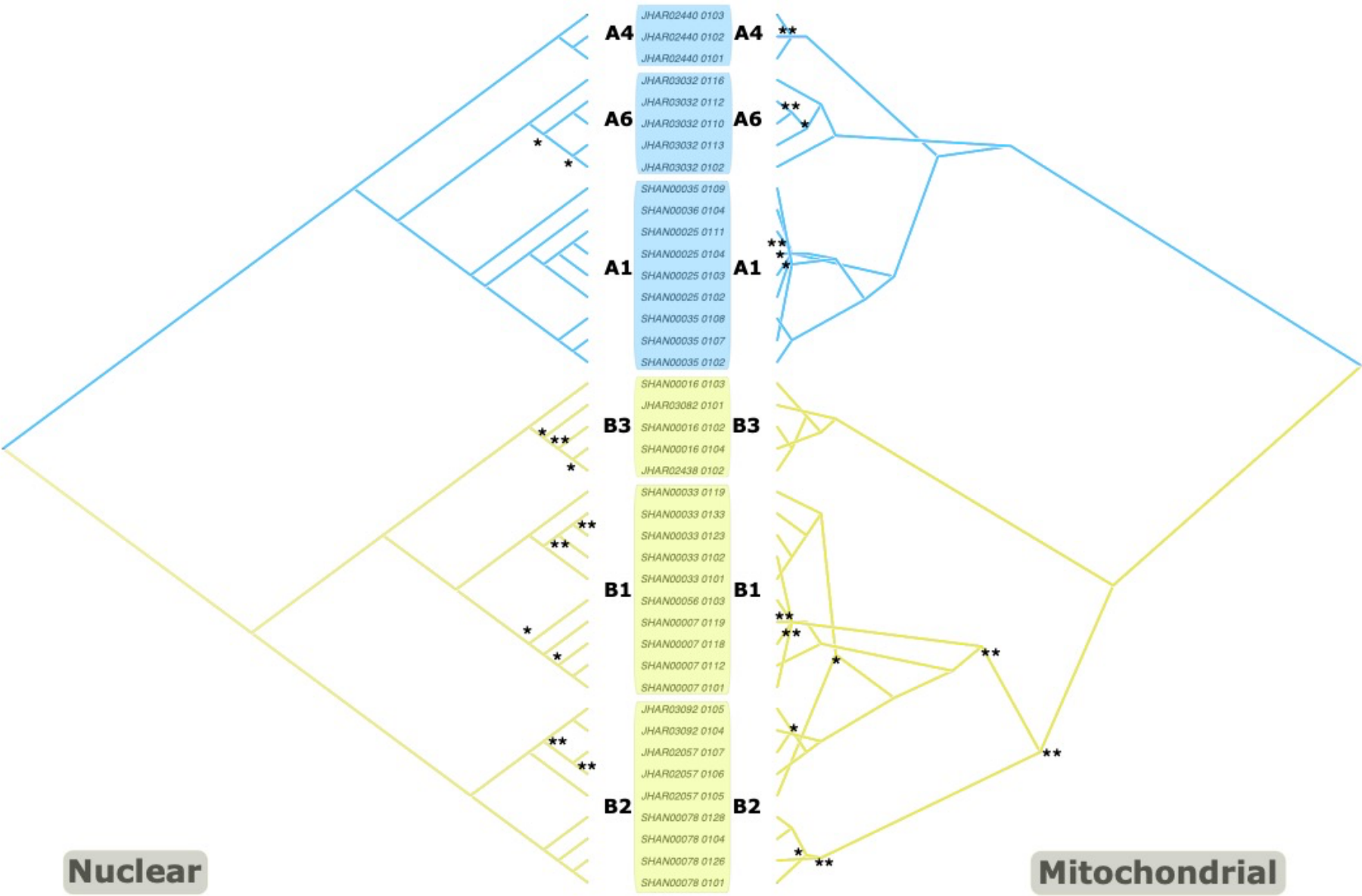


Figure 2 | Tanglegram of the maximum-likelihood phylogenies constructed from both the nuclear (left) and mitochondrial (right) datasets. Blue: *Phlyctinus callosus*. Yellow: *Phlyctinus xerophilus*. Individuals from each lineage are encapsulated in separate boxes. Nodes with ultrafast bootstrap values below 95% and 80% are coded * and **, respectively. Tanglegram with Phytools in R.

Notably, this divergence exhibits a level comparable to interspecific values (e.g. A1/B1: $dXY=7,75.10^{-9}$, $Fst=0,708$). A4 also appears differentiated from the other *P. callosus* lineages, though to a lesser extent, with high Fst (A1/A4: $Fst=0,640$; A4/A6: $Fst=0,678$) but low absolute divergence (A1/A4: $dXY=3,31.10^{-9}$; A4/A6: $dXY=3,24.10^{-9}$). A1 and A6 are not markedly discriminated ($dXY=1,89.10^{-9}$, $Fst=0,023$).

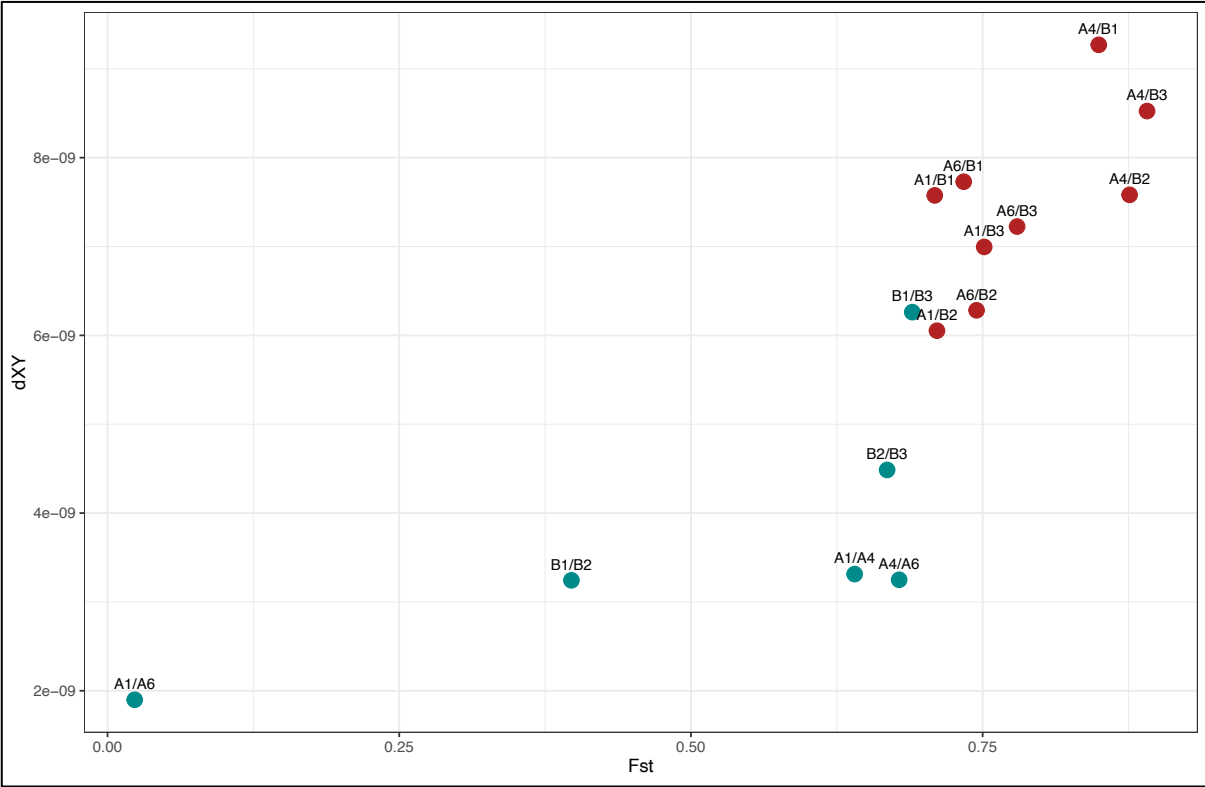


Figure 3 | Absolute (dXY) and relative (Fst) genetic divergence, computed pairwise between all putative lineages. Blue: intraspecific comparisons. Red: interspecific comparisons.

The genetic PCA conducted on **Dataset B** is consistent with the previous lines of evidence in favor of a high differentiation of lineages A4 and B3 (**Fig.4**). The first principal component separates the two established species, explaining 58.4 % of the variability in the dataset, while subsequent PCs corresponds to intra-clade differentiation (11.4%, 8.89% and 7,75% of variance explained for PC2, 3 and 4 respectively). Interestingly, PC2 strongly separates lineage B3 from the rest of the *P. xerophilus* clade, while PC4 differentiates lineage A4 from the remainder of *P. callosus*. In both species, other lineages are remarkably homogeneous on all principal components, except lineage B2, whose substructure is discriminated along PC3 (**Fig.S3**).

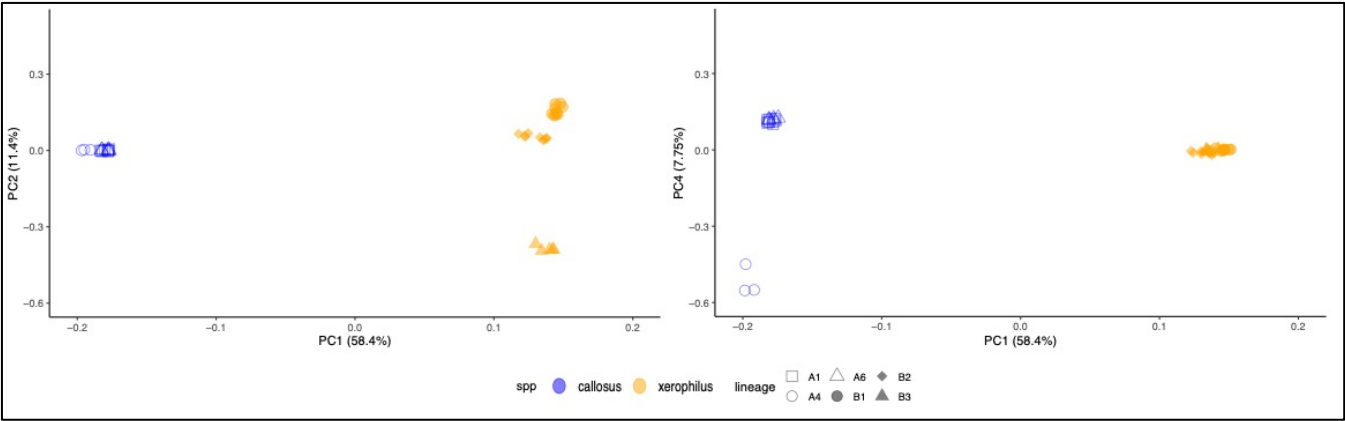


Figure 4 | Genetic principal components analysis conducted on 95595 SNPs (Dataset B). Blue: *P. callosus*. Yellow: *P. xerophilus*. Plots showing other principal component combinations are available in Fig.S3.

Patterns of population structure inferred by sNMF further consolidate the status of lineage B3 (*P. xerophilus*) as an evolutionarily independent unit (**Fig.5**). Comparison of cross-entropy criteria for all seven runs indicates K=5 to be the best value of K (cross-entropy = 0.367, see **Fig.S4**). First, the analysis of admixture for K=2 successfully recovers the dichotomy between both currently established species, albeit with slightly mixed ancestries for clades B2, B3 and A4. K=3 introduces a new cluster corresponding to lineage B1, while K=4 finally splits *P. xerophilus* into the three putative lineages. The admixture analysis performed with the optimal value of K, K=5, adds a new cluster corresponding to lineage A4, with faint signatures of admixture. The other four clusters are identical to those recovered with K=4. Higher values of K immediately separate lineages A1 and A6, while progressively introducing sub-lineage structure consistent with the overall phylogenetic topology for K=7 and K=8. Once recovered, lineage B3 shows no sign of admixture, except for one individual (SHAN00016_0103). Interestingly, signals of asymmetrical admixture between B1 and B2 appear to be pervasive from K=2 to K=6, which may indicate the existence of gene flow between lineages.

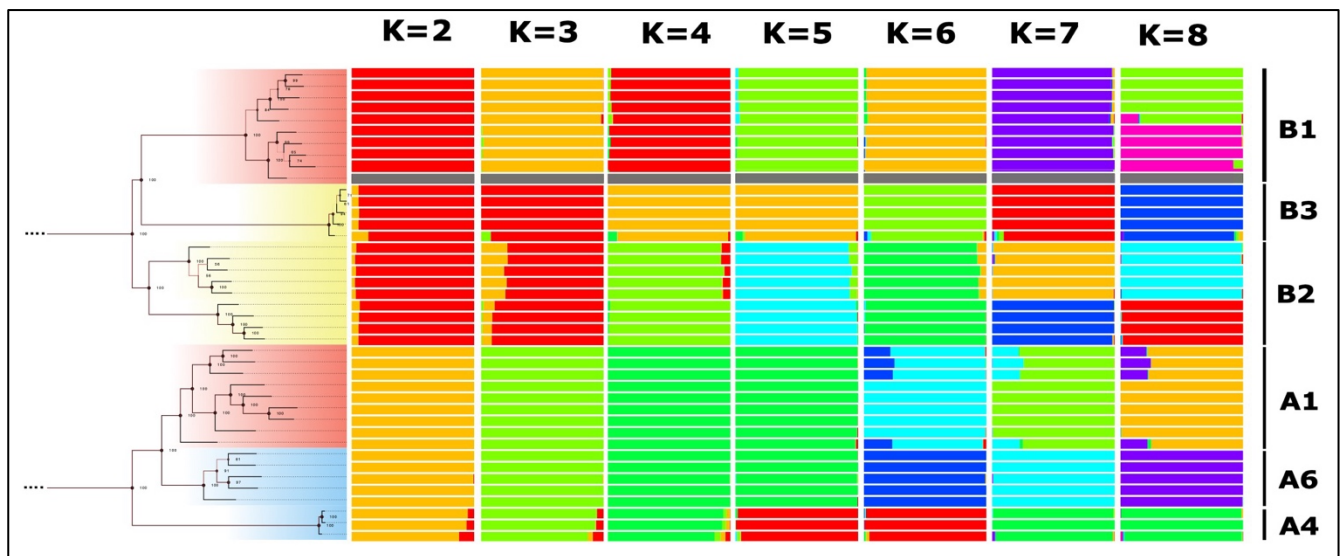


Figure 5 | Population structure analysis performed on the approx. 95 000 SNPs of Dataset B using sNMF. Left: Nuclear phylogeny based on Dataset A. Right: ancestry proportions for all 40 considered individuals. K indicates the number of putative clusters for each run. The grey bars symbolize SHAN00033_0133, which was removed from Dataset B.

For all migration edge counts (N=1 to N=5), however, the migration weights inferred by TreeMix were low (<10%) (**Fig.S5**). The contrasted results yielded by sNMF and TreeMix may be explained by biases specific to the inference methods of both programs, leaving the biological reality of the admixture signal in doubt.

Last, the species delimitation analysis bPTP provided convincing results in favor of the elevation of lineages A4 and B3 to species status. Convergence was checked for and acceptable (see **Fig. S6**). Both the maximum-likelihood solution and the highest Bayesian supported solution included six species corresponding to lineages (A1+A6) (p=0,657), A4 (p=0,973), B1 (p=0,892), B2.1 (p=0,693), B2.2 (p=0,661) & B3 (p=0,978). The number of inferred species ranged from 4 to 14 (mean=6.66).

III.4 | Genome-wide signatures of selection

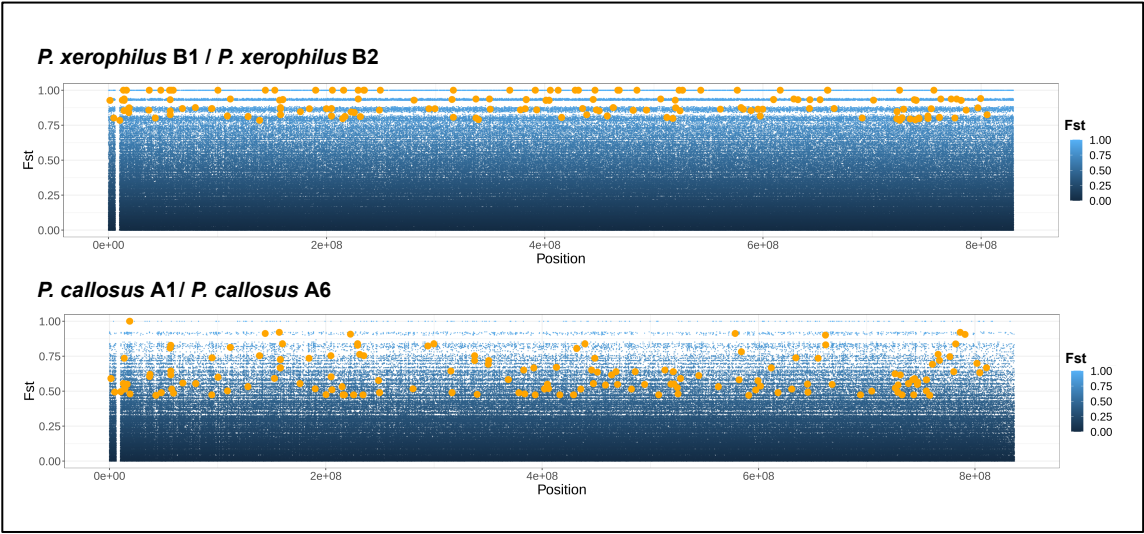


Figure 6 | Genome-wide per-site F_{st} between the shifted and closest non-shifted lineage in both species. Top: *Phlyctinus xerophilus* (B1 against B2). Bottom: *Phlyctinus callosus* (A1 against A6). Data points corresponding to common outlier loci are displayed in orange.

A total of 1 014 208 (*P. callosus*) and 118 822 (*P. xerophilus*) SNPs were retained based on high F_{st} values between shifted and non-shifted lineages, and considered outliers. The intersection of these two lists amounted to 150 SNPs common to both clades which were hypothesized, in the framework of this exploratory approach, to be tied to traits of potential relevance to the focal host shifts (**Fig. 6**). BlastX queries on all 150 10kb sequences extracted from the reference genome returned a total of 124 762 hits. The 5 best-scoring hits were investigated for each query, and high-scoring hits of potential ecological relevance are summarized in **Table 3**.

Table 3 | Summary information for BlastX hits of putative ecological relevance. Genomic location given with respect to the *Phlyctinus callosus* reference genome.

NCBI Accession Number	Hit taxon	Genomic location	Bitscore	Size of translated protein (AA)	Available annotation
XP_052750997.1	<i>Galleria mellonella</i>	Contig 10734	575	1431	Cellular and retroviral pepsin-like aspartate protease
XP_052750997.1	<i>Galleria mellonella</i>	Contig 10415	488	1431	Cellular and retroviral pepsin-like aspartate protease
XP_030765491.1	<i>Sitophilus oryzae</i>	Contig 10022	134	326	Stretch-activated Ca ²⁺ -permeable channel component
XP_050302448.1	<i>Anthonomus grandis</i>	Contig 10022	114	379	Stretch-activated Ca ²⁺ -permeable channel component
XP_045500068.1	<i>Colias croceus</i>	Contig 8696	75	591	Reverse-transcriptase-like

XP_048520651.1	<i>Dendroctonus ponderosa</i>	Contig 9402	152	3025	Cadherin tandem repeat domain
----------------	-------------------------------	-------------	-----	------	-------------------------------

IV. Discussion

IV.1 | Lineages A4 and B3 epitomize the obstacles faced by species delimitation

The present nuclear data enable the monophyletic recovery of all the lineages previously proposed based on morphology and mitochondrial phylogenies. Nevertheless, it is apparent that genetic differentiation between these groups is highly variable. In each species, the shifted lineage shares a close phylogenetic relationship with another, non-shifted lineage (A1 and A6; B1 and B2 for *P. callosus* and *P. xerophilus*, respectively), suggesting that the latter is a separate population closely related to the original group that shifted from Asteraceae to crops. On the other hand, lineages A4 and B3 both clearly show substantial differentiation from the rest of their respective clades, although species delimitation results from various approaches indicate that their respective evolutionary histories may be radically different. There appears to be reasonable agreement in favor of the delimitation of B3 as a separate taxon, as evidenced by high absolute and relative divergence from the rest of *P. xerophilus* —comparable to interspecific differences— and the apparent lack of admixture between it and the other lineages. Conversely, while lineage A4 is recovered as distinct in population structure analyses and displays high relative divergence from the rest of *P. callosus*, the net sequence divergence between these groups remains relatively low and comparable to known intraspecific levels. This would indicate that the differentiation observed between A4 and the other *P. callosus* groups is a consequence of local demographics while, on the contrary, B3 appears to be the product of a recent speciation event. This interpretation is consistent with the distinctive biogeography of the CFR and the distribution patterns of both species. With the exception of artificially displaced inland populations (e.g. lineage A1) *P. callosus* is a mostly coastal species, while *P. xerophilus* consistently occurs further inland, in drier areas exhibiting more complex topography. The extant genus-wide diversity of *Phlyctinus* was likely shaped by a sequence of pocket speciation events caused by oscillating oceanic levels during the Miocene (Hévin et al., 2022). However, the intraspecific processes that led to the differentiation of separate lineages in both focal species discussed herein are more recent, and were likely governed by the current geography of the area. In that respect, we hypothesize that the coastal environment allowed low gene flow to be maintained between A4 and the rest of *P. callosus*, giving a reasonable explanation for its high relative divergence and moderate absolute divergence. B3, on the other hand, is largely isolated from the rest of *P. xerophilus* by part of the cape fold belt, which effectively severed gene flow between the lineages in a case of standard vicariance. It is important to note, however, that preliminary work on morphometrics (S. Hansen, *in preparation*) appears to significantly discriminate both lineages B3 and A4 as separate taxa, corroborating the results of bPTP. This discrepancy between DNA-based species delimitation and morphology-based species inference reasserts the importance of an integrative approach to taxonomy, and may warrant further investigation of the status of *P. callosus* A4.

IV.2 | Unclear signals of admixture between shifted and non-shifted clades

The analysis of population structure by sNMF reveals asymmetrical signals of admixture between shifted and non-shifted lineages in both species. In *P. callosus*, some A1 individuals are consistently assigned as admixed with lineage A6, while the ancestry of A6 individuals is recovered as pure. Conversely, *P. xerophilus* B1 and B2 display relatively pure ancestries, with evidence for two-way, unbalanced background admixture (biased towards B2). Interestingly, the directions of these putative introgression-like events are opposite. Adaptive introgression of specific genes from pests into non-pest species in the face of anthropogenic pressure has been documented in the past

(Valencia-Montoya et al., 2020), but the case of lineages A1 and A6 appears to contradict this pattern. Nonetheless, local density imbalances between admixing populations are expected to generate asymmetrical admixture patterns (Currat et al., 2008), which may be a more parsimonious explanation for the signals observed in both the *Phlyctinus* species. In the case of *P. callosus* B1 and B2, this putative admixture is particularly coherent with the known movements of freight in the region. The Rawsonville area sits on the path of a major agricultural shipping lane for products headed towards the coast, likely creating a contact zone between displaced B1 individuals and the local B2 lineage. Perhaps more intriguingly, lower values of K yield signals of admixture between *P. callosus* A4 and *P. xerophilus*. While no major prezygotic barriers to hybridization are known for genus *Phlyctinus* (S. Hansen pers. obs.), the reduction of this signal for higher values of K (K=6-8) makes the recovery of such hybrids a likely artifact of sNMF's inference method.

IV. 3 | Genome scans point to regions of potential ecological relevance despite poor genome annotation

Although an apparent lack of annotation for the majority of the BlastX sequence hits limited the usability of the outlier SNP dataset, six loci of potential interest could be identified. First, the two pepsin-like proteins identified on contigs 10734 and 10415 could indicate adaptation to a new trophic resource. Interestingly, pepsin is a gut endopeptidase, consistent with a drastic shift in trophic resource at larval stages: grapevine and deciduous fruit trees have significantly more sclerified root systems than Asteraceae, likely requiring a suite of adaptations, including from the digestive tract. Further, two regions on contig 10022 aligned to sequences coding for mechanoreceptor subunits, albeit with lower scores. Mechanoreceptors have been demonstrated to play a critical role in oviposition and mating in other insects (see, e.g., Acebes et al. (2003), Maher & Thiery (2004), Romani et al. (2005)), and this may be especially important in *Phlyctinus* where males, remarkably, display erect setae on ventrite 5 that are likely involved in reproduction (Haran et al., 2020). Should this signal correspond to an actual signature of selection, this may be consistent with the pronounced differences in substrate between cultivated localities and areas where the non-shifted lineages occur, a critical factor for oviposition and, perhaps, the subsequent development of the ectophagous *Phlyctinus* larvae. Interpreting the possible implications of differences in the reproductive structures of shifted and non-shifted lineages in the context of host shifts, however, would be too speculative at this stage. Further, a region on contig 9402 is associated with a cadherin tandem repeat domain. Cadherin has previously been associated with resistance to *Bacillus thuringiensis* Cry toxins (Bel & Escriche, 2006), insecticidal compounds used in transgenic crops to increase resistance to pests (Bravo et al., 2013). This may be a particularly interesting result, considering that extensive use of pesticides was repeatedly shown to induce resistance in insects (Hemingway et al., 2002). Whether local crops can secrete such toxins, however, is unclear. Last, a lower-confidence hit was annotated as a possible reverse-transcriptase-like protein, likely indicating the presence of a retrotransposon. Transposable elements are pervasive in weevil genomes (72% in *Sitophilus*, see Parisot et al. (2021)) and may confer high evolvability to the group. Further investigation of the possible relevance of this region, however, is not possible at this stage.

V. Acknowledgements

I would like to sincerely thank my three supervisors, Rémi Allio, Julien Haran and Charles Perrier (CBGP, France), for their continued support, a number of stimulating discussions, as well as their valuable and extensive comments on earlier versions of this manuscript. I also thank Laure Benoit (CBGP, France) for the precious insight she provided regarding DNA extraction protocols, and Pierre Nouhaud (CBGP, France) for his help and expertise regarding genome scan practices. Special thanks to Steffan Hansen (Stellenbosch University, South Africa) for his knowledge of the *Phlyctinus* system and his availability in the later stages of this project. Finally, my gratitude goes to Julie Gauzere (INRAE,

France) and Pierre-Alexandre Gagnaire (ISEM, France) for their important comments on a preliminary report that served as the basis for this manuscript.

VI. References

- Acebes, A., Cobb, M., & Ferveur, J.-F. (2003). Species-specific effects of single sensillum ablation on mating position in *Drosophila*. *Journal of Experimental Biology*, 206(17), 3095–3100. <https://doi.org/10.1242/jeb.00522>
- Allio, R., Nabholz, B., Wanke, S., Chomici, G., Pérez-Escobar, O. A., Cotton, A. M., Clamens, A. L., Kergoat, G. J., Sperling, F. A. H., & Condamine, F. L. (2021). Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nature Communications*, 12(1). <https://doi.org/10.1038/s41467-020-20507-3>
- Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdociimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, 20(4), 892–905. <https://doi.org/10.1111/1755-0998.13160>
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Barnes, B. N., & Capatos, D. (1989). Evidence for an aggregation pheromone in adult frass of banded fruit weevil, *Phlyctinus callosus* (Schoenherr) (Col., Curculionidae). *Journal of Applied Entomology*, 108(1–5), 512–518. <https://doi.org/10.1111/j.1439-0418.1989.tb00485.x>
- Berenbaum, M. R., Favret, C., & Schuler, M. A. (1996). On Defining “Key Innovations” in an Adaptive Radiation: Cytochrome P450s and Papilionidae. *The American Naturalist*, 148, S139–S155.
- Bernal, J. S., & Medina, R. F. (2018). Agriculture sows pests: how crop domestication, host shifts, and agricultural intensification can create insect pests from herbivores. *Current Opinion in Insect Science*, 26, 76–81. <https://doi.org/10.1016/j.cois.2018.01.008>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Boncristiani, H., Li, J., Evans, J., Pettis, J., & Chen, Y. (2011). Scientific note on PCR inhibitors in the compound eyes of honey bees, *Apis mellifera*. *Apidologie*, 42, 457–460. <https://doi.org/10.1007/s13592-011-0009-9>
- Bortolus, A. (2008). Error cascades in the biological sciences: The unwanted consequences of using bad taxonomy in ecology. *Ambio*, 37(2), 114–118. [https://doi.org/10.1579/0044-7447\(2008\)37\[114:ecitbs\]2.0.co;2](https://doi.org/10.1579/0044-7447(2008)37[114:ecitbs]2.0.co;2)
- Bragard, C., Di Serio, F., Gonthier, P., Jaques Miret, J. A., Justesen, A. F., Magnusson, C. S., Milonas, P., Navas-Cortes, J. A., Parnell, S., Potting, R., Reignault, P. L., Thulke, H., Van der Werf, W., Vicent Civera, A., Yuen, J., Zappalà, L., Gregoire, J., Malumphy, C., Campese, C., . . . MacLeod, A. (2021). Pest categorisation of *Phlyctinus callosus*. *EFSA Journal*, 19(8). <https://doi.org/10.2903/j.efsa.2021.6800>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. H., & Sanmartín, I. (2018). Testing the Role of the Red Queen and Court Jester as Drivers of the Macroevolution of Apollo Butterflies. *Systematic Biology*, 67(6), 940–964. <https://doi.org/10.1093/sysbio/syy009>
- Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), 3133–3157. <https://doi.org/10.1111/mec.12796>

- 455 Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The hidden side of invasions: Massive
introgression by local genes. *Evolution; International Journal of Organic Evolution*, 62(8),
1908–1920. <https://doi.org/10.1111/j.1558-5646.2008.00413.x>
- 460 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter,
G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000 Genomes Project Analysis.
(2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158.
<https://doi.org/10.1093/bioinformatics/btr330>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T.,
McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools.
GigaScience, 10(2). <https://doi.org/10.1093/gigascience/giab008>
- 465 Diehl, S. R., & Bush, G. L. (1984). An Evolutionary and Applied Perspective of Insect Biotypes. *Annual*
Review of Entomology, 29(1), 471–504.
<https://doi.org/10.1146/annurev.en.29.010184.002351>
- Di Franco, A., Poujol, R., Baurain, D., & Philippe, H. (2019). Evaluating the usefulness of alignment
filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary*
Biology, 19(1), 21. <https://doi.org/10.1186/s12862-019-1350-2>
- 470 Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., Heckel, D. G., Der, J.
P., Wafula, E. K., Tang, M., Hofberger, J. A., Smithson, A., Hall, J. C., Blanchette, M., Bureau, T.
E., Wright, S. I., dePamphilis, C. W., Eric Schranz, M., Barker, M. S., . . . Wheat, C. W. (2015).
The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the*
National Academy of Sciences, 112(27), 8362–8366.
475 <https://doi.org/10.1073/pnas.1503926112>
- Ehrlich, P. R., & Raven, P. H. (1964). BUTTERFLIES AND PLANTS: A STUDY IN COEVOLUTION. *Evolution*,
18(4), 586–608. <https://doi.org/10.1111/j.1558-5646.1964.tb01674.x>
- 480 Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient
estimation of individual ancestry coefficients. *Genetics*, 196(4), 973–983.
<https://doi.org/10.1534/genetics.113.160572>
- Goldblatt, P., & Manning, J. C. (2002). Plant Diversity of the Cape Region of Southern Africa. *Annals of*
the Missouri Botanical Garden, 89(2), 281. <https://doi.org/10.2307/3298566>
- 485 Guillot, G., Mortier, F., & Estoup, A. (2005). Geneland: A computer package for landscape genetics.
Molecular Ecology Notes, 5(3), 712–715. <https://doi.org/10.1111/j.1471-8286.2005.01031.x>
- Haran, J. M., Hansen, S., Benoit, L., & Addison, P. (2020). Description of five new species in the genus
Phlyctinus Schoenherr (Coleoptera, Curculionidae): a first step in deciphering the *P. callosus*
complex. *European Journal of Taxonomy*, 669. <https://doi.org/10.5852/ejt.2020.669>
- 490 Hévin, N. M., Hansen, S., Addison, P., Benoit, L., Kergoat, G. J., & Haran, J. (2022). Late Cenozoic
environmental changes drove the diversification of a weevil genus endemic to the Cape
Floristic Region. *Zoologica Scripta*, 51(6), 724–740. <https://doi.org/10.1111/zsc.12563>
- Hood, G.R., Powell, T.H.Q., Doellman, M.M., Sim, S.B., Glover, M., Yee, W.L., Goughnour, R.B.,
Mattsson, M., Schwarz, D. and Feder, J.L. (2020), Rapid and repeatable host plant shifts drive
reproductive isolation following a recent human-mediated introduction of the apple maggot
495 fly, *Rhagoletis pomonella*. *Evolution*, 74: 156–168. <https://doi.org/10.1111/evo.13882>
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI
BLAST: a better web interface. *Nucleic Acids Research*, 36(suppl_2), W5–W9.
<https://doi.org/10.1093/nar/gkn201>
- 500 Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple
sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–
3066. <https://doi.org/10.1093/nar/gkf436>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform.
Bioinformatics, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>

- Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016).
505 MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
methodologies and community practices. *Methods* (San Diego, Calif.), 102, 3–11.
<https://doi.org/10.1016/j.ymeth.2016.02.020>
- Lounsbury, C. P. (1896). The calandra (*Phlyctinus callosus*, Bohem). *Agricultural Journal of the Cape
of Good Hope*, 9, 63-64.
- 510 Maher, N., & Thiery, D. (2004). Distribution of chemo- and mechanoreceptors on the tarsi and
ovipositor of female European grapevine moth, *Lobesia botrana*. *Entomologia
Experimentalis et Applicata*, 110(2), 135–143. [https://doi.org/10.1111/j.0013-
8703.2004.00131.x](https://doi.org/10.1111/j.0013-8703.2004.00131.x)
- Matenaar, D., Bazelet, C. S., & Hochkirch, A. (2015). Simple tools for the evaluation of protected areas
515 for the conservation of grasshoppers. *Biological Conservation*, 192, 192–199.
<https://doi.org/10.1016/j.biocon.2015.09.023>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler,
D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce
framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9),
520 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Mittermeier, R. A., Gil, P. R., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., Lamoreux, J.,
Fonseca, D. G. A. B., Seligmann, P. A., & Ford, H. (2005). *Hotspots Revisited: Earth's Biologically
Richest and Most Endangered Terrestrial Ecoregions* (Revised ed.). Conservation
International.
- 525 Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000b). Biodiversity
hotspots for conservation priorities. *Nature*, 403(6772), 853–858.
<https://doi.org/10.1038/35002501>
- Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic
divergence. *Molecular Ecology*, 18(3), 375–402. [https://doi.org/10.1111/j.1365-
294X.2008.03946.x](https://doi.org/10.1111/j.1365-294X.2008.03946.x)
- 530 O'Meara, B. C. (2009). New Heuristic Methods for Joint Species Delimitation and Species Tree
Inference. *Systematic Biology*, 59(1), 59–73. <https://doi.org/10.1093/sysbio/syp077>
- Parisot, N., Vargas-Chávez, C., Goubert, C., Baa-Puyoulet, P., Balmand, S., Beranger, L., Blanc, C.,
Bonnamour, A., Boulesteix, M., Burlet, N., Calevro, F., Callaerts, P., Chancy, T., Charles, H.,
535 Colella, S., Da Silva Barbosa, A., Dell'Aglio, E., Di Genova, A., Febvay, G., ... Heddi, A. (2021).
The transposable element-rich genome of the cereal pest *Sitophilus oryzae*. *BMC Biology*,
19(1), 241. <https://doi.org/10.1186/s12915-021-01158-2>
- Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient
Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, 31(7),
540 1929–1936. <https://doi.org/10.1093/molbev/msu136>
- Pickrell, J. K. & Pritchard, J. K. (2012) Inference of Population Splits and Mixtures from Genome-Wide
Allele Frequency Data. *PLoS Genet* 8(11): e1002967.
<https://doi.org/10.1371/journal.pgen.1002967>
- Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin,
W. D., & Vogler, A. P. (2006). Sequence-Based Species Delimitation for the DNA Taxonomy of
545 Undescribed Insects. *Systematic Biology*, 55(4), 595–609.
<https://doi.org/10.1080/10635150600852011>
- Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the
Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular
Biology and Evolution*, 35(10), 2582–2584. <https://doi.org/10.1093/molbev/msy159>
- 550 Romani, R., Salerno, G., Frati, F., Conti, E., Isidoro, N., & Bin, F. (2005). Oviposition behaviour in *Lygus
rugulipennis*: A morpho-functional study. *Entomologia Experimentalis et Applicata*, 115(1),
17–25. <https://doi.org/10.1111/j.1570-7458.2005.00268.x>

- 555 Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12), 1–22. <https://doi.org/10.1371/journal.pbio.2000234>
- Russo, A., Mayjonade, B., Frei, D., Potente, G., Kellenberger, R. T., Frachon, L., Copetti, D., Studer, B., Frey, J. E., Grossniklaus, U., & Schlüter, P. M. (2022). Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Frontiers in Plant Science*, 13. <https://doi.org/10.3389/fpls.2022.883897>
- 560 Sahoo, R. K., Warren, A. D., Collins, S. C., & Kodandaramaiah, U. (2017). Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: HesperIIDae). *BMC Evolutionary Biology*, 17(1). <https://doi.org/10.1186/s12862-017-1016-x>
- 565 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44(W1), W232–W235. <https://doi.org/10.1093/nar/gkw256>
- 570 Valencia-Montoya, W. A., Elfekih, S., North, H. L., Meier, J. I., Warren, I. A., Tay, W. T., Gordon, K. H. J., Specht, A., Paula-Moraes, S. V., Rane, R., Walsh, T. K., & Jiggins, C. D. (2020). Adaptive Introgression across Semipermeable Species Boundaries between Local *Helicoverpa zea* and Invasive *Helicoverpa armigera* Moths. *Molecular Biology and Evolution*, 37(9), 2568–2583. <https://doi.org/10.1093/molbev/msaa108>
- 575 Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6), 1358–1370. <https://doi.org/10.2307/2408641>.
- Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22), 2869–2876. <https://doi.org/10.1093/bioinformatics/btt499>
- 580