Population genomics of the mostly thelytokous *Diplolepis rosae* (Linnaeus, 1758) (Hymenoptera: Cynipidae) reveals population-specific selection for sex.

Ksenia Mozhaitseva[1], Zoé Tourrain[1], and Antoine Branca[1]

[1] Laboratoire Evolution, Génomes, Comportement, Ecologie, l'Institut Diversité, Ecologie et Evolution du Vivant, Université Paris-Saclay, Gif-sur-Yvette, France

[1] *E-mail: ksenia.mozhaitseva@universite-paris-saclay.fr*

ABSTRACT

Arrhenotoky is the favored reproductive mode used by Hymenoptera though thelytoky, the production of females by virgin females, is observed in some taxa. *Diplolepis rosae* is a gall wasp producing bedeguars when parasitizing wild dog roses and reproduces mainly by thelytoky. In this study, we assess the effect of facultative thelytoky on the genetic structure, the genomic structure of polymorphism, and the efficacy of recombination in *D. rosae*. We assembled a high-quality reference genome using Oxford Nanopore long-read technology and sequenced 17 samples collected in France with high-coverage Illumina reads. We found two *D. rosae* peripatric populations that split recently and diverged quickly probably due to *Wolbachia*-induced thelytoky. However, we demonstrated a switch from thelytoky to arrhenotoky in both populations. One population ('arrhenotokous') shows a 6.7 times higher recombination rate and localized runs of homozygosity (ROHs) compared with the other population ('thelytokous') showing widespread ROHs. We showed that genes involved in male traits in the 'arrhenotokous' population are under negative selection while under balancing or relaxed selection in the 'thelytokous' population. Thus, although *D. rosae* favors thelytoky, strong selection acts on maintaining male function and provides a striking example of the advantage of recombination in natural populations.

Key words: *Diplolepis rosae,* thelytoky, selection, population genomics, demographic history, *Wolbachia.*

INTRODUCTION

Many organisms can alternate between sexual and asexual reproduction in different ways. Sexual reproduction is essential to create genetic diversity for adaptation to changing environments, whereas asexual reproduction is important in short term and in stable environments providing rapid population growth.

The Hymenoptera use a variety of reproductive modes: complementary sex determination (CSD), arrhenotokous haplodiploidy (arrhenotoky), and thelytoky (Vorburger 2014). Complementary sex determination has been described in *Habrobracon* sp. (Whiting 1933) and is based on the genotype at one or several sex-determination loci (Asplen et al. 2009). In the case of single locus CSD, heterozygotes at this locus develop into females while homozygotes develop into males. Diploid males can be either sterile like in *Habrobracon* sp. (Holloway et al. 1999), or, in some cases, such as *Cotesia* spp. (de Boer et al. 2007; Elias et al. 2009) and *Euodynerus foraminatus* (Cowan and Stahlhut 2004), viable and fertile. Arrhenotokous haplodiploidy is a special case of CSD and represents the classical haplodiploid sex determination when diploid females develop from fertilized eggs, while unfertilized eggs produce haploid hemizygous males. Thelytoky is a less frequent reproductive mode when virgin females produce only females (Heimpel and De Boer 2008; Rabeling and Kronauer 2013). Thelytoky can be encoded in the genomes of hymenopterans (Belshaw and Quicke 2003; Capdevielle-Dulac et al. 2022; Engelstädter et al. 2011; Foray et al. 2013; Wenseleers and Billen 2000) or induced by endosymbionts (Stouthamer et al. 1990; Stouthamer and Kazmer 1994). Genetically based thelytoky exists in the form of automixis, i. e. gamete fusion or gamete duplication after meiosis, and in the form of apomixis, i. e.

mitotic division of ootids with no meiosis (Heimpel and De Boer 2008; Queffelec et al. 2021; Schön et al, 2009). Notably, automictic Hymenopteran can still produce rare diploid males (the mechanism is described in detail in Stenberg and Saura 2009). The presence of males in thelytokous populations can also be explained by a failure of the mechanism inducing thelytoky or by occasional gene flow between arrhenotokous and thelytokous populations (Stouthamer and Kazmer 1994; Engelstädter et al. 2011). Endosymbiont-induced thelytoky is provoked by endocytoplasmic bacteria like *Rickettsia, Cardinium,* or *Wolbachia* (Adachi-Hagimori et al. 2011; Giorgini et al. 2009; Werren et al. 2008). *Wolbachia* is the most widespread intracellular parasite infecting from 25% to 70% of all insect species (Kozek and Rao 2007), as well as other arthropods and nematodes (Hilgenboecker et al. 2008). *Wolbachia* is known for its effects on arthropod reproduction. In the Hymenoptera, it can cause cytoplasmic incompatibility or thelytokous parthenogenesis (Vavre et al. 2002). *Wolbachia*-induced thelytoky has been extensively studied in *Trichogramma* spp. (Stouthamer et al. 1990; Stouthamer and Kazmer 1994), where thelytokous females exposed to antibiotic treatment or high temperatures constantly produced arrhenotokous offspring (Stouthamer et al. 1990). *Wolbachia*-mediated thelytoky is induced via gamete duplication, i. e. the failure of chromosome segregation in unfertilized eggs during anaphase I and subsequent duplication of terminal meiosis products. Hence, this leads to completely homozygous females (Stouthamer and Kazmer 1994).

Consequently, thelytoky may result in the rapid development of parthenogenetic lineages and fine population structure, i.e. 1) allopatric (Plantard et al. 1998; Young 1990) or sympatric (Adachi-Hagimori et al. 2011; Schneider et al. 2002; Stouthamer and Kazmer 1994) coexistence of the arrhenotokous and thelytokous populations, and/or 2) several locally adapted thelytokous populations (Schilder et al. 1999).

The coexistence of different modes of reproduction, showing variability in their frequency of recombination and heterozygosity, is deeply related to the cost of sex in both short- and long-term perspectives (Hamilton 1980; Maynard-Smith 1978). Thelytokous lineages provide a rapid increase in population size and maintain adaptability to a constant environment. At the same time, recombination and heterozygosity provide genetic variability in the face of parasitism (Hamilton 1980; Van Valen 1973) and changing environments (Amat et al. 2006). Thus, this may lead to ecological niche differentiation and reduce competition between two modes of reproduction (Case and Taper 1986). For instance, in the parasitic wasp *Venturia canescens,* arrhenotokous and thelytokous individuals sampled from the same geographical location were shown to be adapted to different habitats (Amat et al. 2006). Arrhenotokous wasps are found in field (spatially heterogeneous) conditions and are more sensitive to temperature changes. Thelytokous (genome-encoded thelytoky) wasps are found only in indoor (more constant) conditions.

Thus, from a long-term perspective, thelytoky might lead to reproductive isolation between populations and further speciation (Adachi-Hagimori et al. 2011; Bordenstein 2003; Werren 1998).

In this study, we aimed at assessing the effect of facultative thelytoky on the genomic structure of polymorphism and the efficacy of recombination in the mostly thelytokous *Diplolepis rosae* (Hymenoptera: Cynipidae).

Diplolepis rosae is a holarctic cynipid wasp parasitizing the wild dog roses (*Rosa* spp. section *caninae*) and causing specific plant outgrowths called rose bedeguar galls (Giron et al. 2016; Shorthouse and Floate 2010).

Cytological examination of *Diplolepis rosae* showed that they can reproduce by thelytokous parthenogenesis via gamete duplication (Stille and Dävring 1980). After anaphase of the second meiotic division, three nuclei formed polar bodies, later dissolved, while one of ootids

became the pronucleus entering the mitotic division. Diploidy was restored by the fusion of the two daughter products resulting in homozygous females. Indeed, *D. rosae* individuals collected from several European locations consisted of almost only females, with the proportion of males varying from 1% to 4% (Nordlander 1973). A further study (Stille 1985) based on the electrophoresis of 27 isozymes showed that *D. rosae* females sampled in Sweden, Germany, and Greece were completely homozygous. Additionally, the females of *D. rosae* were shown to be infected with *Wolbachia* (Plantard et al. 1999; Schilthuizen and Stouthamer 1998; Van Meer et al. 1995). For instance, in French territory, the percentage of *Wolbachia*-infected females varied from 71% to 100% depending on the location (Plantard et al. 1999). Finally, *Wolbachia*-induced thelytoky was indirectly confirmed in a closely related species, *D. spinosissimae* (Plantard et al. 1998; Zhang et al. 2020): in *Wolbachia*-infected populations, males comprised about 1.3% of individuals and no heterozygous genotypes were found when examining three microsatellite loci. Conversely, in *Wolbachia*-free populations, males represented 21 - 29% of individuals and 78 - 96% of females were heterozygous (Plantard et al. 1998).

Diplolepis rosae has one generation per year. The emergence of adults takes place from May to June and is synchronized with the development of host plant tissues to a specific stage (Shorthouse and Floate 2010). After emergence, the females immediately oviposit into epidermal plant cells located between the developing leaflets of an expanding bud (Bronner 1985). The galls of *D. rosae* undergo three stages: initiation, growth, and maturation. Once eggs are laid, gall tissue begins its development. The feeding larvae are surrounded by gall cells and spend at this stage before the next spring (Shorthouse and Floate 2010).

Rosa canina is one of the wild roses from the section *Caninae.* The roses from this section are considered to be the product of allopolyploid hybridization. They are distributed in central Europe and occupy similar habitats. They differ in phenological and morphological

characteristics. All these species are successfully attacked by *D. rosae* (Kohnen et al. 2011; Stille 1985).

In our study, we (i) sampled *D. rosae* galls in metropolitan France, (ii) *de novo* assembled one genome using long and short-read sequencing technology, (iii) sequenced the whole genome of 17 samples using short-read, (iv) reconstructed population structure and demographic history, and (v) performed a genomic scan of polymorphism and recombination across the genome in relation to reproductive modes.

MATERIALS AND METHODS

Sampling. The dog rose (*Rosa canina*) bud galls with *Diplolepis rosae* for Illumina sequencing were collected from April 2020 to December 2020 in France (**Table S1**). The dog rose bud galls with *Diplolepis rosae* for Nanopore sequencing were collected on 17/10/2021 in Le Hom, France (48.993463, -0.470689). After sampling, the galls were kept in plastic bags at room temperature. Insect material was removed by dissection of gall tissue for further DNA extraction. One tube sample contained emerged adults or larvae from the same gall.

DNA extraction for Illumina sequencing. Prior to DNA extraction, the insects were frozen at -80 °C overnight. For 17 samples sent to Illumina sequencing DNA quantity and quality were estimated. The initial mass of the material provided for DNA extraction varied from 13.7 mg to 56.0 mg. Insect material was homogenized by TissueLyser (Qiagen) using a metallic bead. Insect DNA was extracted using the DNeasy Blood and Tissue Kit (QIAGEN, Germany). After extraction, the purity of the samples was evaluated by estimating the 260/280 and the 260/230 ratios measured by the Thermo Scientific NanoDrop 2000 Spectrophotometer. The 260/280 ratio varied from 1.69 to 2.08, the 260/230 ratio varied from 0.79 to 2.08. DNA quantity was estimated using Qubit dsDNA BR Assay Kit (Invitrogen) and the Qubit

Fluorometer and varied from 0.98 to 6.9 µg. The Illumina sequencing was performed by Genotul, Toulouse, France (https://get.genotoul.fr). The Illumina samples are available in the NCBI database under the BioSample accessions from SAMN32903234 to SAMN32903250.

DNA extraction for Nanopore sequencing. Prior to DNA extraction, the fresh insect material was ground by hand with a sterile pestle. The DNA was extracted using the NucleoBond Buffer Set IV kit (MACHEREY-NAGEL, Germany) and the NucleoBond AXG 20 columns (MACHEREY-NAGEL, Germany). The long-read sequencing was performed using Flow Cell Wall Wash Kit (EXP-WSH004) (Oxford Nanopore Technologies, UK) and Oxford Nanopore MinION Flow Cell R10 (Oxford Nanopore Technologies, UK).

Genome assembly. Oxford Nanopore Technology (ONT) raw reads were base called using Guppy basecaller v. 6.1.2+e0556ff (Oxford Nanopore Technologies, UK) with accurate mode and dna_r9.4.1_450bps_sup.cfg config. Resulting *.fastq* reads were assembled with Flye assembler (Kolmogorov et al. 2019) using --nano-hq flag, an estimated coverage of 18x, and an error rate of 0.1. The input genome size was 580 Mbp based on a k-mer estimate using GenomeScope (Vurture et al. 2017). Because of the high error rate of ONT reads, the resulting assembly was polished using NextPolish (Hu et al. 2020) with high-coverage Illumina reads (119x) using a sample from the same population (sample ESE-219, **Table S1**). Repeats were discovered in the de novo assembly using RepeatModeler v. 2.0.3 (Flynn et al. 2020) using RMBlast v. 2.11.0. *De novo*-detected repeats were assessed for potential wrong assignment to repeats of highly duplicated genes by blasting each consensus repeat sequence onto the NCBI nr database. Repeats matching known protein-coding genes not related to known repeats were removed and filtered repeats were then used to build the repeat database. *De novo* assembly was then masked using the newly built repeat database. The assembly quality was evaluated by computing different metrics using QUAST v. 5.1.0rc1 (Mikheenko et al. 2018). Protein-coding genes were predicted using Braker with ab-initio mode (Hoff et

al. 2019) using protein homology from OrthoDB Metazoa, Fungi, Plants, and Bacteria for the ProtHint (Bruna et al. 2020) step. The completeness of annotated genes in terms of their expected gene content was evaluated using BUSCO v. 5.3.2 (Manni et al. 2021). Predicted genes were functionally annotated using eggNOG-mapper v. 2.1.7 web server (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2019).

SNP calling. All reads of *Diplolepis rosae* were firstly aligned to the reference using Bowtie 2 v. 2.3.5.1 (Langmead and Salzberg 2012). The alignments were processed for further manipulations using Samtools 1.7 (view, sort, index, depth, and stat options) (Danecek et al. 2021). The quality of each alignment was estimated by calculating the average percentage of the mapped reads (varied from 57.3% to 98.1%), the average quality (35.4 - 35.7), the average coverage (25x - 119x), and the error rate (0.0078 to 0.015). After that, the alignments were converted to a file in variant call format by using an algorithm proposed by GenomeAnalysisToolkit (GATK) v. 4.0 (AddOrReplaceReadGroups, HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs tools) (https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode). Indels and non-biallelic sites were removed from the *.vcf* file using bcftools v. 1.7 (filter command) (Danecek et al. 2021). The total number of examined SNPs was 3,281,648.

Population structure. The population structure of *Diplolepis rosae* was inferred according to the fastStructure algorithm (Raj et al. 2014) based on the calculation of the allele frequency spectra from SNP data. Prior to fastStructure, indels and sites linkage disequilibrium (r^2>0.2) were filtered out using bcftools v. 1.7 (filer and prune commands) (Danecek et al. 2021) from the initial *.vcf* file that was then converted to a *.bed* file by plink v. 1.9 (Purcell et al. 2007). After filtering, the total number of examined SNPs was 12,309. The population structure was

assessed with the number of populations, K, varying from 1 to 4. Admixture statistics were calculated by ADMIXTOOLS 2.0.0 (Patterson et al. 2012).

DNA extraction for genotyping. Prior to DNA extraction, the insects were frozen at -20 °C overnight. The DNA was extracted using either DNeasy Blood and Tissue Kit (QIAGEN, Germany), or Chelex 100 Resin (Bio-Rad).

Genotyping. A genetic marker helping to distinguish the lineages of *D. rosae* was searched along the genome by choosing a 15 kbp-window showing Fst value closest to 1 and showing substantial polymorphism. A 706-bp sequence containing 9 SNPs was selected and the primers were designed using Primer3web 4.1.0 (Untergasser et al. 2012**).** The marker was then amplified using polymerase chain reaction (PCR) (**Protocol S1**). The presence of the PCR product was verified by performing the electrophoresis in 3.5% agarose gel and then sent for sequencing at Eurofins Genomics. The sequences were assessed visually on the trace file using SnapGene Viewer 6.0.2 ("SnapGene software" www.snapgene.com) and quality trimmed. After that, the sequences were used for cladogram construction (Maximum likelihood statistical method, Tamura-Nei substitution model (default)) by MEGA 11 (Tamura et al. 2021). One hundred twenty-three *D. rosae* individuals from 61 different locations were genotyped for the marker. Samples were collected from different habitats in France, from different host plant individuals from the same habitat, and from different galls from the same host plant.

Population genomic statistics. Absolute divergence D_{xy} (Nei 1987) and the fixation index Fst (Weir and Cockerham 1984) were calculated in the genome 10-kbp windows using pixy v. 1.2.6.beta1 (Korunes and Samuk 2021). Tajima's D was calculated in 10-kbp windows by VCF-kit 0.2.9 (Cook and Andersen 2017). Nucleotide diversity π (in 10-kbp windows) and per-individual heterozygosity H were calculated by vcftools 0.1.17 (Danecek et al. 2021).

Watterson estimator θ*w* was calculated as the number of segregating sites (provided in the output table using vcftools 0.1.17 when calculating π) in the genome 10-kbp windows divided by a sum of the (n-1) first harmonic means, where n is the number of haplotypes (Watterson 1975). Fst, D_{xy}, Tajima's D, π, H, and θ*w* were calculated taking into account all sites of the assembly (--all-sites tag in GATK). Because deleterious mutations arise more frequently in coding sequences, background selection is expected to cause a negative correlation between gene density and diversity. Therefore, the effect of background selection on the genome was assessed by checking the correlation between gene density (the proportion of nucleotides assigned to a coding sequence in a genome fragment) and π within each scaffold. Population-scaled per-bp recombination rate ρ and r^2 statistic of linkage disequilibrium (LD) were calculated within each scaffold having a length of more than 1 Mbp. The total number of examined scaffolds was 138 having 710.0 Mbp of the total length (93.3% of the assembly). In order to detect putative recombining scaffolds, 'the median value of r^2 per 50 bp bins' was fitted as a linear function to the log(distance between two SNPs)' on the 0 - 2000 bp region (lm function in R 4.2.2) (R Core Team 2022 https://www.R-project.org/). The scaffolds with a slope not significantly deviating from 0 (i.e. showing a constant r^2 value) at any distance between examined alleles were considered to reflect the absence of recombination, i.e. all alleles were transmitted together from one generation to another. The LD was calculated by vcftools 0.1.17 (Danecek et al. 2011). Population scaled recombination rate ρ was estimated in the filtered-from-minor-allele-frequency phased *.vcf* file (vcftools tag --maf 0.2, phased by beagle v. 5.4 (Browning et al. 2021)) by ReLERNN (Adrion et al. 2020) using ReLERNN_SIMULATE - ReLERNN_TRAIN - ReLERNN_PREDICT - ReLERNN_BSCORRECT pipeline. In ReLERNN_TRAIN, --nEpochs (time) was estimated

at 500 corresponding to a minimal convergence of loss (mean squared error) between the training set and the validation set.

Demographic scenarios. Scenarios describing the demographic history of *D. rosae* were examined using *dadi* software based on the diffusion (continuous approximation) approach (Gutenkunst et al., 2009). The following standard two-population models were examined: bottleneck followed by exponential growth then split without (i) and with migration (ii), split into two populations of specified size (iii), isolation with exponential population growth (iv), isolation with exponential population growth, and a size change prior to split (v) (**Code S1**). A set of joint allele frequency spectra (AFS) was generated to compare the model with the data. Each model was examined by varying population sizes, time of split/isolation, and migration parameter (0, symmetric, or asymmetric). The models were ranked by calculating the Akaike Information Criterion (AIC). A likelihood-ratio test was performed to compare the best models according to AIC.

Estimation of model parameters. To confirm the demographic scenario found by *dadi* and estimate parameters describing the demography of *D. rosae* (**Code S2**), the approximate Bayesian computation (ABC) was applied. One million simulations of the population parameters (effective population size Ne, mutation rate μ, bottleneck time, split time) were performed, and resulting summary statistics were calculated by msprime simulator v. 1.1.1 (Baumdicker et al. 2022). After that, the goodness-of-fit and validation of the model were performed using the *abc* R package v. 2.2.1 (Csilléry et al. 2012).

Wolbachia strain comparison. To test the hypothesis of whether different *D. rosae* lineages are infected with different strains of *Wolbachia*, its sequences were removed from initial Illumina *D. rosae* reads by aligning them to A- and B-supergroup *Wolbachia* references obtained from the NCBI (accessions are CP042445.1 and CP016430.1, respectively) using

Bowtie 2. After that, the sequences were processed using Samtools, GATK, and fastStructure similarly to *Diplolepis rosae*. To calculate the coverage of each *Wolbachia* strain the reads were aligned to an A- and B-supergroup combined reference with further identification of reads unique to A- or B-supergroup (samtools view and samtools depth). The mean coverage of *Wolbachia* was normalized by the mean coverage of corresponding nuclear *D. rosae* nuclear scaffolds.

Detection of genome regions under selection. In order to find regions under selection in the *D. rosae* genome, we calculated Fst and π in 10-kbp windows using pixy v. 1.2.6.beta1 (Korunes and Samuk 2021) and detected the runs of homozygosity (ROH) using bcftools v. 1.7 (roh command) with the option -G, the phred-scaled genotype likelihoods, set to 30 (Danecek et al. 2021). After that, we performed the scan of a composite score equal to (1 - Fst)*2(F(π) - 0.5), where F(π) is the cumulative distribution function, in each population. The composite score values above 0.5 reflected genome regions showing putative signatures of balancing selection (low Fst and high π), and the values below -0.5 reflected regions with putative signatures of purifying selection (low Fst and low π). After that, annotated gene sets (eggNOG-mapper) found in these regions were used in the Gene Set Enrichment Analysis (GSEA). The GSEA was performed using the *topGO* R package v. 2.50.0 using Fisher Exact test (Alexa and Rahnenfuhrer 2022).

Statistics. All statistical analyses were performed by R v 4.2.2 (R Core Team 2022 https://www.R-project.org/). All results are presented as mean ± standard error of mean (SE). The significance level was set to 0.05. The figures were produced by R v. 4.2.2 or Microsoft Excel 2010.

RESULTS

Genome structure. The total length of the genome assembly is estimated at 760.7 Mbp, the total number of contigs is 757, N50 is 7,663,408 bp, the largest scaffold is 33,454,033 bp long, and L50 is 25 (**Table S2**). Repetitive sequences represented 69% of the genome assembly: 48.50% of the assembly length are unclassified repeats, 11.40% are retroelements, and 8.66% are DNA transposons (**Table S3**). Regarding expected gene content, the completeness of annotated gene set is estimated at 91.8% of complete and single-copy, 0.5% of complete and duplicated, 1.6% of fragmented, and 6.1% of missing BUSCOs when using hymenoptera_odb10 dataset. The total number of predicted genes was 20,301, of which 14559 were partially or fully annotated. The assembled genome of *Diplolepis rosae* is available at DDBJ/ENA/GenBank under the accession JAPYXD000000000. The version described in this paper is version JAPYXD010000000.

Population structure. Two populations of *Diplolepis rosae* have been inferred in metropolitan France (**Fig. 1**). One individual, *D. rosae*-652, shows admixed ancestry with 86.2% of its genome assigned to population 2 and the other 13.8% to population 1. The negative value of f3-admixture statistics ($f3 = -0.0797 \pm 0.0150$, $z = -5.29$, $p < 0.0001$) proves this individual to be admixed between lineages related to population 1 and population 2. The additional genotyping confirms two lineages distributed in the French territory (**Fig. 2, Fig. S1**).

Population genomic statistics. The mean Fst value between the two populations is 0.82 ± 0.00070 (**Table 1**). Population 1 is characterized by 6.7 times higher population-scaled recombination rate ρ ($U = 184,850$, $z = 8.40$, $p < 0.0001$), and 1.7 times higher heterozygosity H ($U = 9$, $z = 2.55$, $p = 0.0108$) compared with population 2. The recombination rate in both populations significantly differs from zero (one-sample Wilcoxon test, comparison with the median $u = 0$; pop 1: $W = 249,570$, $p < 0.0001$; pop 2: $W = 249,570$, $p < 0.0001$). In population 1, 83.8% of the assembly length represents scaffolds with detected recombination (**Fig. 3**), and linkage disequilibrium decays as the distance between SNP increases (**Fig. 4A,**

Fig. 4B). In population 2, evidence of recombination was found for 47.6% of the assembly length (**Fig. 3**). In 45.7% of the sequence, the median value of r2 varies from 0.22 to 0.77 depending on the scaffold and does not change with the distance between two SNPs (**Fig. 4C, Fig. 4D**). In population 1, the distribution of Tajima's D shows a slight excess of positive values (one-sample Wilcoxon test, comparison with the median u = 0: W = 1.5325e+09, $p < 0.0001$) (**Table 1**, **Fig. 5**). In population 2, the distribution of Tajima's D shows an excess of negative values (W = 2.1419e+09, $p < 0.0001$). There is no correlation between the gene density and nucleotide diversity in population 1 (Pearson's correlation r = -0.0142, t = -0.218, p = 0.827) nor population 2 (r = -0.0237, t = -0.365, p = 0.715) (**Fig. S2**).

Demographic scenarios. Following *dadi* inference, the model with the lowest AIC describing the demographic history of *Diplolepis rosae* (**Table 2**) is a bottleneck of an ancestral population followed by exponential growth then split into two populations with no gene flow between them (**Fig. 6**). The likelihood ratio test shows no significant difference between the model without or with the migration parameter (Dadj = 0.37, p = 0.28). Migration rate m is estimated at 0.81/(2*Neff) migrants per generation, where Neff is the number of diploid individuals. Additional examination by the Bayesian approach confirms the demographic model to provide a good fit to the empirical summary statistics (dist. = 3365.0, p = 0.21) (see details in **Table S4** and **Fig. S3**).

Wolbachia strain comparison. The population structure of *Wolbachia* does not follow that of *D. rosae* (**Fig. 7**). Almost all examined individuals are infected by a mix of two *Wolbachia* strains representing the supergroups A and B. There is no significant difference between *D. rosae* populations in terms of the number of *Wolbachia* copies neither from supergroup A (Mann–Whitney U test: U = 31, z = 0.433, p = 0.665), nor from supergroup B (U = 25, z = 1.01, p = 0.312), nor their ratio (U = 35, z = 0.048, p = 0.961) (**Fig. S4**).

Detection of genome regions under selection. Population 2 shows widespread runs of homozygosity (ROH) across the genome (**Fig. 8**). On the contrary, population 1 shows localized ROH that can extend to several Mbp (**Fig. 9**). In several regions, the ROH overlaps with a decrease in Fst value with a simultaneous change in π in one or both populations (**Fig. 8, Fig. 9**). When considering regions outside the interval]-0.5;0.5[for the composite score, the total length of the regions assumed under putative selection was estimated at 32,38 Mbp (4.23% of the genome). The GSEA of the genome regions showing the composite score less than -0.5 (negative selection) or more than 0.5 (balancing selection) found several significantly enriched terms of gene ontologies (GO) (**Table 3, Table S5**): negative selection in both populations, balancing selection in both populations (GO 'ribosome'), negative selection in population 1 but balancing selection in population 2 (GO 'sperm competition', GO 'metalloendopeptidase activity', GO 'extracellular space'), and balancing selection in population 1 but negative selection in population 2 (GO 'commissural neuron axon guidance') **(**for more detail, see **Fig. S5-S8**).

DISCUSSION

In this study, we aimed to determine how thelytoky affects the genomic structure of polymorphism in *D. rosae*. Two peripatric highly differentiated lineages of *D. rosae* have been found in France, with no clear geographical structure but each of them differing in how thelytoky have shaped their genome.

Demographic history of D. rosae. In the Île-de-France region, population 1 is more frequent than population 2 while in the south of France, population 2 seems to be more frequent than population 1. Regarding their geographical distribution, current abiotic factors, such as current climate conditions or topography, do not seem to be the structuring factors. Indeed, both populations have been recovered all across France. The host plants could also be a

structuring factor when it is a limiting resource, but dog roses are widespread and very common plants. In addition, it has been found no correlation between host plant genotypes and gall wasp genotypes (Kohnen et al. 2011; Stille 1985) so it is unlikely as well that specialization at the genotype level is at play. Furthermore, in two instances among 61 localities sampled, individuals from different populations were sampled onto the exact same host plant on the same branch. So it seems unlikely that the host availability or genotype plays a role in structuring the populations of *D. rosae* in France. Gall wasps have been shown to be exposed to intense attacks by natural enemies notably hymenopteran parasitoids and cynipid inquilines (Laszlo et al. 2014; Rizzo and Massa 2006; Stille 1984; Todorov et al. 2012). For instance, the study of Rizzo and Massa (2006) shows that on average only 5% of the progeny per gall survives in each generation (Rizzo and Massa 2006). Therefore, avoiding parasitism would be highly selected and it is assumed to be the main factor driving gall morphology (Stone et al. 2003). We could then hypothesize that each population is attacked by/defended against a specific parasite community that controls population density. Local variation in the presence of parasite species will then determine which population prevails. However, we cannot exclude the existence of microhabitats in the plant such as the different elevations of parasitized leaf buds or the orientation of the galls toward the sun.

Diplolepis rosae demographical history follows the scenario "bottleneck of an ancestral population then split into two populations with no migration". Although mathematically the most probable model does not comprise a migration parameter, we observed a recently admixed individual and the negative value of admixture f3 statistics that gene flow has taken place between two lineages. Whereas the two methods show the same demographic scenario, the estimation of model parameters by ABC lacks accuracy. Indeed, first, ABC estimates the prediction error for all model parameters to be close to 1, and, second, the posteriors have the same or broader distributions than the priors and do not allow to properly estimate the

parameters. The ABC inferences seem to be adequate only for effective population sizes, estimated at 500 diploid individuals in each lineage, which is typical for hymenopterans (Antolin et al. 2003; Schmack et al. 2019; Zayed 2004; Zayed and Packer 2001).

In population 2, the excess of negative values of Tajima's D statistic could be due to either background selection or recent population expansion. However, in Hymenoptera, since deleterious mutations in haploid males or homozygous females are immediately removed from the population, the genetic load is expected to be low. In addition, negative selection would lead to a correlation between genetic diversity and gene content, which we did not observe. Thus, in population 2, the excess of the negative Tajima's D values is probably due to recent population expansion. In population 1, the peak corresponding to negative values could be explained by the demography such as in population 2. The peak corresponding to positive values could be explained by the effect of localized balancing selection (see below).

Origin of thelytoky in Diplolepis rosae. In accordance with previous studies (Plantard et al. 1999; Schilthuizen and Stouthamer 1998; Van Meer et al. 1995), all sampled *D. rosae* were found to be infected with *Wolbachia.* The two same strains bi-infect individuals from both populations. This observation can be the result of two scenarios. First, *Wolbachia* strains would have infected the *D. rosae* ancestral population and because of thelytoky induction, resulted in two lineages diverging quickly due to asexual reproduction. However, following that scenario, *Wolbachia* was expected to match the population structure of *D. rosae* and no structure was found for the *Wolbachia* strains. Therefore, second, we can hypothesize that the two populations pre-existed before *Wolbachia* infection and that they have acquired these endocytoplasmic bacteria independently by horizontal transmission either via parasitoids (Werren et al. 1995) or via the same host plants through the phloem (Chrostek et al. 2017; Schilthuizen and Stouthamer 1998). This is also consistent with the observation that different

European *Diplolepis* species have been infected with the same *Wolbachia* strain (Plantard et al. 1999).

Diplolepis rosae was shown cytologically to reproduce by thelytokous parthenogenesis via gamete duplication giving almost only homozygous female offspring (Nordlander 1973, Plantard et al. 1999; Stille 1985; Stille and Dävring 1980). Although two populations show widespread patterns of ROH, in all individuals, they do not cover all the genome and they are very localized in population 1. Our observations can be interpreted in two ways. We can first assume that having been infected with a parthenogenesis-inducing (PI) *Wolbachia* and reproducing by thelytokous parthenogenesis, populations could develop genetic or cytological resistance and restore arrhenotoky or a different thelytoky mechanism such as central or terminal fusion (López-Madrigal and Duarte 2019). Apparently, population 1 developed resistance longer ago than population 2 because it presents much fewer ROHs and more recombining regions. The remaining homozygous blocks could reflect ancestral thelytoky via gamete duplication. The second hypothesis is that both populations are contaminated with non-PI *Wolbachia* and *D. rosae* is capable of restoring genome-encoded thelytoky via central or terminal fusion of gametes. However, it disagrees with the cytological data (Stille and Dävring 1980). In this case, we could suppose that the studied Sweden *D. rosae* population (Stille and Dävring 1980) had not yet developed resistance to PI-*Wolbachia* provoking thelytoky via gamete duplication.

Change in selective regime in D. rosae. According to previous data, it was expected that *D. rosae* reproduced mainly by *Wolbachia*-induced thelytoky through gamete duplication. In such a situation, the efficacy of recombination is strikingly reduced leading to selection by clonal interference rather than selective sweep. Therefore selection would act on the whole genome rather than being localized. However, we found that recombination does take place frequently although at a much higher level in population 1 than in population 2. So we

perform scans of negative and balancing selection in both populations to assess how selection changed according to the mode of reproduction which differed between the two populations. Balancing selection such as negative selection are both expected to reduce Fst but balancing selection maintains a high level of diversity (high π) while negative selection tends to maintain a low level of diversity (low π). To the contrary, positive selection would lead to high Fst and low diversity but in *D. rosae* we expect it to be too challenging to assess as Fst between the two populations is already close to one and clonal interference may have been at play, especially in population 2, leading to whole genome background being hitchhiked. Therefore, to properly detect positive selection further data would be needed notably an outgroup to perform tests such as the McDonald-Kreitman test (McDonald and Kreitman 1991) which should not be affected by demography or the mode of reproduction (Eyre-Walker 2002; Parsch et al. 2009). Regarding balancing and negative selection, our results demonstrate a difference in the selective regime between the two populations in localized genomic regions. The GSEA shows that in the more recombining population 1, the GO terms such as 'sperm competition' ('Biological process' annotation), 'metalloendopeptidase activity' ('Molecular function' annotation), and 'extracellular space' ('Cellular component' annotation) are under negative selection, while in less recombining population 2 they are putatively under balancing selection. The genes from 'commissural neuron axon guidance' ('Biological process' annotation) are under balancing selection in population 1 and under negative selection in population 2. What is striking is that it seems that selection is opposite between the two populations, switching between balancing selection and negative selection. It is possible that the detected balancing selection would be actually relaxed selection where Fst would be low because of negative ancestral selection and π high because of no more constraint on mutation. Even so, there seems to be a potential link between most of the GO terms and ecological, physiological, or morphological differences in the two *D. rosae* populations. Indeed, the

changes in the selective regime of 'sperm competition' and 'commissural neuron axon guidance' candidate genes could give some insights into different selective processes acting on arrhenotokous and thelytokous populations. In population 1 having restored arrhenotokous reproduction longer ago, genes involved in sexual reproduction such as male traits would be under strong selection while in a strictly thelytokous population, male traits would be expected not to be kept. Such a strong selection for the male function could be explained by the importance of recombining, i.e. producing males, to generate genetic diversity in face of parasitism. Indeed, Rizzo and Massa (2006) showed a possible association between average parasitization and the percentage of emerged *D. rosae* males: parasitism rate was 30.5% and 34.3% in two *D. rosae* populations consisting of 21% (Sicilian galls collected in 1965) and 15.6% (Hungarian galls collected in 2001) of males, respectively, while in others (Sicilian galls collected from 1992 to 2005) there were no males and average parasitization reached to 57.6% (varied from 12.5% to 100%). In population 2 showing widespread ROHs and recombining less, the candidate genes from the 'sperm competition' GO term are under putative balancing selection or relaxed selection. Relaxed selection is easier to interpret because it is expected under the loss of male function but balancing selection could also be found if there is a selection for maintaining polymorphism in the capacity to produce males. Contrary to 'sperm competition', the genes involved in 'commissural neuron axon guidance' were found to be under purifying selection in population 2 and in balancing selection or relaxed selection in population 1. We could assume that thelytokous populations harbor a special host-searching behavioral pattern because we expect the asexual lineage to maximize progeny size. For instance, Ramirez-Romero et al. (2012) showed differences in host-searching behavior patterns between thelytokous and arrhenotokous populations of *Odontosema anastrephae* (Hymenoptera: Cynopoidea: Cynipidae *sensu lato*), the figitid parasitoid of the fruit fly (*Anastrepha ludens* feeding on guavas) (Blaimer et al. 2020; Hearn

et al. 2022). In this study, thelytokous females showed a basic behavioral sequence when exploring the odor source and foraged longer in non-infested fruits than arrhenotokous wasps.

CONCLUSION

Our study presents a unique example of how the change in the mode of reproduction affects the genome and the selective regime. Indeed, we could demonstrate the coexistence of two peripatric populations in *Diplolepis rosae* that differ in their mode of reproduction. This difference results in a contrasting pattern of recombination and homozygosity across the genome. But even more strikingly, we could show that selection acts upon radically different traits in both populations, in direct relation to their mode of reproduction and/or ecology. However, it is necessary to better assess the difference in the natural history of the populations. Does each population really differ in progeny sex ratio and size? Can we find differences in the parasite community attacking them, qualitatively and quantitatively? Do they show the same phenology? These are key questions to answer in order to figure out what is selective pressure leading to the maintenance of arrhenotoky in *D. rosae*. In addition, we need to resolve how *Wolbachia* interacts with *D. rosae* notably if any genetic resistance to *Wolbachia* can have arisen and be selected in *D. rosae*.

AUTHOR CONTRIBUTIONS

KM performed DNA extraction (genotyping), simulations, and data analysis, and wrote the manuscript. AB conceived and designed the study, wrote the manuscript, performed genome assembly, described the *D. rosae* genome, and performed GSEA. AB, KM, and ZT collected gall samples. ZT kept the gall samples and performed DNA extraction (Illumina). All authors discussed the results and contributed to the final manuscript.

ACKNOWLEDGMENTS

We would like to thank David Ogereau (CNRS, EGCE) for providing the DNA for Nanopore sequencing. We also thank Jacqui Shykoff (Paris-Saclay University, ESE), Olivier Plantard (INRAE), Amir Yassin (CNRS, EGCE), Florence Mougel (CNRS, EGCE), and Thierry Robert (Paris-Saclay University, ESE). The study received a grant from the National Agency for Research (France) (Project ANR-19-CE02-0008 "Tracing back the history of an adaptive trait: genetic basis of plant host manipulation by gall wasps - BETAGALL").

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA ARCHIVING

Tables, Figures, and Supporting Information (S) are available at https://src.koda.cnrs.fr/g686mjly/mozhaitseva-2022-drosae.

SUPPORTING INFORMATION

Protocol S1. Polymerase chain reaction protocol using *Diplolepis rosae* genetic marker.

Code S1. Scripts used to find the best scenario describing the demography of *Diplolepis rosae* by continuous approximation approach.

Code S2. Script used to estimate model parameters of *Diplolepis rosae* demography (approximate Bayesian computation approach).

Table S1. Quality of *Diplolepis rosae* genome assembly.

Table S2. Classification and percentage of repetitive sequences in the genome of *Diplolepis rosae*.

Table S3. Parameter estimations obtained by approximate Bayesian computation for the demographic model "Bottleneck_growth_split" in *Diplolepis rosae.*

Table S4. Candidate sets of genes under selection in two *Diplolepis rosae* populations detected by the gene set enrichment analysis in the 'Molecular function' and 'Cellular component' Gene Ontologies.

Fig. S1. Phyloogram of *Diplolepis rosae* genotypes collected in France based on one marker. Two branches correspond to two populations.

Fig. S2. Relation between gene density (the proportion of nucleotides assigned to a coding sequence in a genome fragment) and nucleotide diversity (π) in *Diplolepis rosae*.

Fig. S3. Parameter inference of simulated demographic model "Bottleneck_growth_split" in *Diplolepis rosae* using approximate Bayesian computation (ABC). Grey line: given distribution of simulated model parameters (priors). Red line: distribution accepted model parameters (posteriors). N1, N2: effective population sizes of the first and the second population, respectively; proportion: ratio between a number of individuals of *D. rosae* from a population after a bottleneck to an ancestral one; tbot, tsplit: bottleneck and split time, respectively; mu: mutation rate.

Fig. S4. Coverage and ratio of *Wolbachia* Super-group A and B reads recovered from *Diplolepis rosae* sequences. A: *Wolbachia* Super-group A; B: *Wolbachia* Super-group B reference.

Fig. S5. Relation in the 'Biological process' ontology between gene terms under negative selection in population 1 *Diplolepis rosae*. Results are presented as Gene ontology (GO) term, annotation, p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

Fig. S6. Relation 'Biological process' ontology between gene terms under balancing selection in population 2 *Diplolepis rosae*. Results are presented as Gene ontology (GO) term,

annotation, p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

Fig. S7. Relation in the 'Biological process' ontology between gene terms under balancing selection in population 1 *Diplolepis rosae*. Results are presented as Gene ontology (GO) term, annotation, p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

Fig. S8. Relation in the 'Biological process' ontology between gene terms under negative selection in population 2 *Diplolepis rosae*. Results are presented as Gene ontology (GO) term, annotation, p-value (Fisher's exact test), number of detected genes in *D. rosae* / total number of genes in the full gene set.

LITERATURE CITED

Adachi-Hagimori, T., Miura, K., & Abe, Y. (2011). Gene flow between sexual and asexual strains of parasitic wasps: a possible case of sympatric speciation caused by a parthenogenesis‐inducing bacterium. *Journal of Evolutionary Biology*, *24*(6), 1254-1262.

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular biology and evolution*, *37*(6), 1790-1808.

Alexa, A. , Rahnenfuhrer, J. (2022). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.50.0.

Amat, I., Castelo, M., Desouhant, E., & Bernstein, C. (2006). The influence of temperature and host availability on the host exploitation strategies of sexual and asexual parasitic wasps of the same species. *Oecologia*, *148*(1), 153-161.

Antolin, M. F., Ode, P. J., Heimpel, G. E., O'hara, R. B., & Strand, M. R. (2003). Population structure, mating system, and sex-determining allele diversity of the parasitoid wasp *Habrobracon hebetor. Heredity*, *91*(4), 373-381.

Asplen, M. K., Whitfield, J. B., De Boer, J. G., & Heimpel, G. E. (2009). Ancestral state reconstruction analysis of hymenopteran sex determination mechanisms. *Journal of Evolutionary Biology*, *22*(8), 1762-1769.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., ... & Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229.

Belshaw, R., & Quicke, D. L. (2003). The cytogenetics of thelytoky in a predominantly asexual parasitoid wasp with covert sex. *Genome*, *46*(1), 170-173.

Blaimer, B. B., Gotzek, D., Brady, S. G., & Buffington, M. L. (2020). Comprehensive phylogenomic analyses re-write the evolution of parasitism within cynipoid wasps. *BMC evolutionary biology*, *20*(1), 1-22.

Bordenstein, S. R. (2003). 17 Symbiosis and the Origin of Species. *Insect symbiosis*, 283.

Bronner, R. (1985). Anatomy of the ovipositor and oviposition behavior of the gall wasp *Diplolepis rosae* (Hymenoptera: Cynipidae). *The Canadian Entomologist*, *117*(7), 849-858.

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880-1890.

Bruna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP and-EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins. *bioRxiv*, *17*, 2019-12.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, *38*(12), 5825-5829.

Capdevielle-Dulac, C., Benoist, R., Paquet, S., Calatayud, P. A., Obonyo, J., Kaiser, L., & Mougel, F. (2022). Spontaneous parthenogenesis in the parasitoid wasp *Cotesia typhae*: low-frequency anomaly or evolving process? *Peer Community Journal*, *2*.

Case, T. J., & Taper, M. L. (1986). On the coexistence and coevolution of asexual and sexual competitors. *Evolution*, *40*(2), 366-387.

Chrostek, E., Pelz-Stelinski, K., Hurst, G. D., & Hughes, G. L. (2017). Horizontal transmission of intracellular insect symbionts via plants. *Frontiers in Microbiology*, *8*, 2237.

Cook, D. E., & Andersen, E. C. (2017). VCF-kit: assorted utilities for the variant call format. *Bioinformatics*, *33*(10), 1581-1582.

Cowan, D. P., & Stahlhut, J. K. (2004). Functionally reproductive diploid and haploid males in an inbreeding hymenopteran with complementary sex determination. *Proceedings of the National Academy of Sciences*, *101*(28), 10374-10379.

Csilléry, K., Lemaire, L., François, O., & Blum, M. G. B. (2012). Approximate Bayesian computation (ABC) in R: a Vignette. *Doi*, *202*, 217-531.

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.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008.

De Boer, J. G., Ode, P. J., Vet, L. E. M., Whitfield, J. B., & Heimpel, G. E. (2007). Diploid males sire triploid daughters and sons in the parasitoid wasp *Cotesia vestalis*. *Heredity*, *99*(3), 288-294.

Elias, J., Mazzi, D., & Dorn, S. (2009). No need to discriminate? Reproductive diploid males in a parasitoid with complementary sex determination. *PLoS One*, *4*(6), e6024.

Engelstädter, J., Sandrock, C., & Vorburger, C. (2011). Contagious parthenogenesis, automixis, and a sex determination meltdown. *Evolution: International Journal of Organic Evolution*, *65*(2), 501-511.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics*, *162*(4), 2017-2024.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451-9457.

Foray, V., Henri, H., Martinez, S., Gibert, P., & Desouhant, E. (2013). Occurrence of arrhenotoky and thelytoky in a parasitic wasp *Venturia canescens* (Hymenoptera: Ichneumonidae): Effect of endosymbionts or existence of two distinct reproductive modes?. *European Journal of Entomology*, *110*(1).

Giorgini, M., Monti, M. M., Caprio, E., Stouthamer, R., & Hunter, M. S. (2009). Feminization and the collapse of haplodiploidy in an asexual parasitoid wasp harboring the bacterial symbiont *Cardinium*. *Heredity*, *102*(4), 365-371.

Giron, D., Huguet, E., Stone, G. N., & Body, M. (2016). Insect-induced effects on plants and possible effectors used by galling and leaf-mining insects to manipulate their host-plant. *Journal of Insect Physiology*, *84*, 70-89.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, *5*(10), e1000695.

Hamilton, W. D. (1980). Sex versus non-sex versus parasite. *Oikos*, 282-290.

Hearn, J., Gobbo, E., Nieves-Aldrey, J. L., Branca, A., Nicholls, J. A., Koutsovoulos, G., ... & Ronquist, F. (2022). Phylogenomic Analysis of Protein-Coding Genes Resolves Complex Gall Wasp Relationships. *bioRxiv*.

Heimpel, G. E., & De Boer, J. G. (2008). Sex determination in the Hymenoptera. *Annu. Rev. Entomol.*, *53*, 209-230.

Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A., & Werren, J. H. (2008). How many species are infected with *Wolbachia*?–a statistical analysis of current data. *FEMS microbiology letters*, *281*(2), 215-220.

Hoff, K.J., Lomsadze, A., Borodovsky, M. and Stanke, M. (2019). Whole-Genome Annotation with BRAKER. Methods Mol Biol. 1962:65-95, doi: 10.1007/978-1-4939-9173-0_5.

Holloway, A. K., Heimpel, G. E., Strand, M. R., & Antolin, M. F. (1999). Survival of diploid males in *Bracon* sp. near *hebetor* (Hymenoptera: Braconidae). *Annals of the Entomological Society of America*, *92*(1), 110-116.

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309-D314.

Kohnen, A., Wissemann, V., & Brandl, R. (2011). No host-associated differentiation in the gall wasp *Diplolepis rosae* (Hymenoptera: Cynipidae) on three dog rose species. *Biological Journal of the Linnean Society*, *102*(2), 369-377.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, *37*(5), 540-546.

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular ecology resources*, *21*(4), 1359-1368.

Kozek, W. J., & Rao, R. U. (2007). The discovery of *Wolbachia* in arthropods and nematodes–A historical perspective. In *Wolbachia: a bug's life in another bug* (Vol. 5, pp. 1-14). Karger Publishers.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

Laszlo, Z., Solyom, K., Prazsmari, H., Barta, Z., & Tothmeresz, B. (2014). Predation on rose galls: parasitoids and predators determine gall size through directional selection. *PLoS One*, *9*(6), e99806.

López-Madrigal, S., & Duarte, E. H. (2019). Titer regulation in arthropod-*Wolbachia* symbioses. *FEMS microbiology letters*, *366*(23), fnz232.

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. Current Protocols, 1, e323. doi: 10.1002/cpz1.323

Maynard-Smith, J. (1978). *The evolution of sex* (Vol. 4). Cambridge: Cambridge University Press.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, *351*(6328), 652-654.

Mejias, J., Truong, N. M., Abad, P., Favery, B., & Quentin, M. (2019). Plant proteins and processes targeted by parasitic nematode effectors. *Frontiers in plant science*, *10*, 970.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150. doi: 10.1093/bioinformatics/bty266.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.

Nordlander, G. (1973). Parasitsteklar i galler av *Diplolepis rosae* (L.) och *D. mayri* Schlechtd. (Hym. Cynipidae) (Hym. Ichneumonoidea, Chalcidoidea, Cynipoidea). *Entomologisk Tidskrift*, *94*(3-4), 148-176.

Parsch, J., Zhang, Z., & Baines, J. F. (2009). The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in *Drosophila*. *Molecular Biology and Evolution*, *26*(3), 691-698.

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.

Plantard, O., Rasplus, J. Y., Mondor, G., Clainche, I. L., & Solignac, M. (1998). *Wolbachia*–induced thelytoky in the rose gall wasp *Diplolepis spinosissimae* (Giraud) (Hymenoptera: Cynipidae), and its consequences on the genetic structure of its host. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1401), 1075-1080.

Plantard, O., Rasplus, J. Y., Mondor, G., Le Clainche, I., & Solignac, M. (1999). Distribution and phylogeny of *Wolbachia* inducing thelytoky in Rhoditini and 'Aylacini'(Hymenoptera: Cynipidae). *Insect Molecular Biology*, *8*(2), 185-191.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Queffelec, J., Allison, J. D., Greeff, J. M., & Slippers, B. (2021). Influence of reproductive biology on establishment capacity in introduced Hymenoptera species. *Biological Invasions*, *23*(2), 387-406.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabeling, C., & Kronauer, D. J. (2013). Thelytokous parthenogenesis in eusocial Hymenoptera. *Annu Rev Entomol*, *58*(273), 92.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573-589.

Ramirez-Romero, R., Sivinski, J., Copeland, C. S., & Aluja, M. (2012). Are individuals from thelytokous and arrhenotokous populations equally adept as biocontrol agents? Orientation and host searching behavior of a fruit fly parasitoid. *BioControl*, *57*(3), 427-440.

Rizzo, M. C., & Massa, B. (2006). Parasitism and sex ratio of the Bedeguar gall wasp Diplolepis rosae (L.) (Hymenoptera: Cynipidae) in Sicily (Italy). *Journal of Hymenoptera Research*, *15*(2), 2779285.

Schilder, K., Heinze, J., Gross, R., & Hölldobler, B. (1999). Microsatellites reveal clonal structure of populations of the thelytokous ant *Platythyrea punctata* (F. Smith) (Hymenoptera; Formicidae). *Molecular Ecology*, *8*(9), 1497-1507.

Schilthuizen, M., & Stouthamer, R. (1998). Distribution of *Wolbachia* among the guild associated with the parthenogenetic gall wasp *Diplolepis rosae*. *Heredity*, *81*(3), 270-274.

Schmack, J. M., Brenton-Rule, E. C., Veldtman, R., Wenseleers, T., Beggs, J. R., Lester, P. J., & Bulgarella, M. (2019). Lack of genetic structuring, low effective population sizes, and major bottlenecks characterize common and German wasps in New Zealand. *Biological Invasions*, *21*(10), 3185-3201.

Schneider, M. V., Beukeboom, L. W., Driessen, G., Lapchin, L., Bernstein, C., & Van Alphen, J. J. M. (2002). Geographical distribution and genetic relatedness of sympatrical thelytokous and arrhenotokous populations of the parasitoid V*enturia canescens* (Hymenoptera). *Journal of Evolutionary Biology*, *15*(2), 191-200.

Schön, I., Martens, K., & van Dijk, P. (2009). Lost sex. *The evolutionary biology of parthenogenesis*, 1-615.

Shorthouse, J. D., & Floate, K. D. (2010). Galls induced by cynipid wasps of the genus *Diplolepis* (Hymenoptera: Cynipidae) on the roses of Canada's grasslands. *Arthropods of Canadian Grasslands. Biological Survey of Canada, Ottawa*, 251-279.

Stenberg, P., & Saura, A. (2009). Cytology of asexual animals. In *Lost sex* (pp. 63-74). Springer, Dordrecht.

Stille, B. (1984). The effect of hosptlant and parasitoids on the reproductive success of the parthenogenetic gall wasp *Diplolepis rosae* (Hymenoptera, Cynipidae). *Oecologia*, *63*(3), 364-369.

Stille, B. (1985). Population genetics of the parthenogenetic gall wasp *Diplolepis rosae* (Hymenoptera, Cynipidae). *Genetica*, *67*(2), 145-151.

Stille, B. O., & Dävring, L. (1980). Meiosis and reproductive strategy in the parthenogenetic gall wasp *Diplolepis rosae* (L.) (Hymenoptera, Cynipidae). *Hereditas*, *92*(2), 353-362.

Stone, G. N., Graham N., & Schönrogge, K. (2003). The adaptive significance of insect gall morphology. *Trends in Ecology & Evolution*, *18*(10), 512-522.

Stouthamer, R., & Kazmer, D. J. (1994). Cytogenetics of microbe-associated parthenogenesis and its consequences for gene flow in *Trichogramma* wasps. *Heredity*, *73*(3), 317-327.

Stouthamer, R., Luck, R. F., & Hamilton, W. D. (1990). Antibiotics cause parthenogenetic Trichogramma (Hymenoptera/Trichogrammatidae) to revert to sex. *Proceedings of the National Academy of Sciences*, *87*(7), 2424-2427.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution*, *38*(7), 3022-3027.

Todorov, I., Stojanova, A., Parvanov, D., & Boyadzhiev, P. (2012). Studies on the gall community of *Diplolepis rosae* (Hymenoptera: Cynipidae) in Vitosha Mountain, Bulgaria. *Acta Zool. Bulg*, *4*, 27-37.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic acids research*, *40*(15), e115.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1-30.

Van Meer, M. M. M., van Kan, F. J. P. M., Breeuwer, J. A. J., & Stouthamer, R. (1995). Identification of symbionts associated with parthenogenesis in *Encarsia formosa* (Hymenoptera: Aphelinidae) and *Diplolepis rosae* (Hymenoptera. Cynipidae). *Proc. Exp. Appl. Entomol. NE*, *6*.

Vavre, F., Fleury, F., Varaldi, J., Fouillet, P., & Bouletreau, M. (2002). Infection polymorphism and cytoplasmic incompatibility in Hymenoptera-*Wolbachia* associations. *Heredity*, *88*(5), 361-365.

Vorburger, C. (2014). Thelytoky and sex determination in the Hymenoptera: mutual constraints. *Sexual Development*, *8*(1-3), 50-58.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202-2204.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, *7*(2), 256-276.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *evolution*, 1358-1370.

Wenseleers, T., & Billen, J. (2000). No evidence for *Wolbachia*-induced parthenogenesis in the social Hymenoptera. *Journal of Evolutionary Biology*, *13*(2), 277-280.

Werren, J. H. (1998). *Wolbachia* and speciation. *Endless forms: species and speciation*, 245-260.

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia*: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*(10), 741-751.

Werren, J. H., Zhang, W., & Guo, L. R. (1995). Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1360), 55-63.

Whiting, P. W. (1933). Selective fertilization and sex-determination in Hymenoptera. *Science*, *78*(2032), 537-538.

Young, D. K. (1990). Distribution of *Pelecinus polyturator* in Wisconsin (Hymenoptera: Pelecinidae), with speculations regarding geographical parthenogenesis. *The Great Lakes Entomologist*, *23*(1), 1.

Zayed, A. (2004). Effective population size in Hymenoptera with complementary sex determination. *Heredity*, *93*(6), 627-630.

Zayed, A., & Packer, L. (2001). High levels of diploid male production in a primitively eusocial bee (Hymenoptera: Halictidae). *Heredity*, *87*(6), 631-636.

Zhang, Y. M., Buffington, M. L., Looney, C., László, Z., Shorthouse, J. D., Ide, T., & Lucky, A. (2020). UCE data reveal multiple origins of rose gallers in North America: Global phylogeny of *Diplolepis* Geoffroy (Hymenoptera: Cynipidae). *Molecular phylogenetics and evolution*, *153*, 106949.

TABLES

Table 1. Population genomic statistics (mean ± standard error of mean) for *D. rosae.*

population /statistic	π	ρ	H, %	θ*w*	TajimaD	D_{xy}	Fst
pop 1	0.00067 ± 2.5e-6	4.8e-09 ± 1.6e-10	0.0010 ± 3.6e-5	0.00059 ± 2.0e-6	0.28 ± 0.0041	0.0033 ± 8.8e-6	0.82 ± 0.00070
pop 2	0.00058 ± 2.6e-6	7.2e-10 ± 6.9e-11	0.00060 ± 0.00010	0.00060 ± 2.1e-6	-0.52 ± 0.0038		

π: nucleotide diversity; ρ: per-bp population-scaled recombination rate in units of 4*Neff generations, where Neff is effective population size (haploid individuals); H: per individual heterozygosity (% of heterozygous sites in the total assembly length); D_{xy}: absolute divergence; Fst: fixation index; θ*w*: Watterson estimator.

Table 2. Akaike Information Criterion (AIC) values for examined demographic scenarios for *D. rosae*.

Model	AIC
bottlegrowth_split_no_migration	74419.2
bottlegrowth_split_migration	79152.6
split_no_migration	144643.8
split_asymmetric_migration	145062.8
split_symmetric_migration	233881.6
isolation_no_migration	154600.6
isol_asymmetric_migration	146823.6
isol_symmetric_migration	159047.4
exp_pop_growth_size_change_split_no_migration	372553.8
exp_pop_growth_size_change_split_asymmetric_migration	371429.2
exp_pop_growth_size_change_split_symmetric_migration	371833.6

Table 3. Candidate sets of genes under selection in two *D. rosae* populations detected by the gene set enrichment analysis in the 'Biological process' Gene Ontology.

Selective regimes in *D. rosae* populations	Annotation
Negative population 1	GO:0007618 ('mating'): 9/160 ($p < 0.0001$) GO:0046692 ('sperm competition') †: 9/29 ($p < 0.0001$)
Negative population 2	GO:0000270 ('peptidoglycan metabolic process'): 5/11 ($p = 0.000113$) GO:0006418 ('tRNA aminoacylation for protein translation'): 10/55 ($p = 0.000358$) GO:0071679 ('commissural neuron axon guidance') ‡: 6/18 ($p = 0.000174$) GO:0038089 ('positive regulation of cell migration by vascular endothelial growth factor signaling pathway'): 5/12 ($p = 0.000186$) GO:2001028 ('positive regulation of endothelial cell chemotaxis'): 5/13 ($p = 0.000291$)
Balancing population 1	GO:0006412 ('translation'): 40/566 ($p < 0.0001$) GO:0071679 ('commissural neuron axon guidance') ‡: 6/18 ($p < 0.0001$) GO:0021794 ('talamus development'): 5/19 ($p = 0.000477$)
Balancing population 2	GO:0046692 ('sperm competition') †: 9/29 ($p < 0.0001$)

Results are presented as "Gene ontology (GO) term ('annotation'): number of detected genes in *D. rosae* / total number of genes in a full gene set (p-value, Fisher's exact test)". †: candidate genes

under negative selection in population 1 and in balancing selection in population 2; ‡: candidate genes under balancing selection in population 1 and in negative selection in population 2.

FIGURES

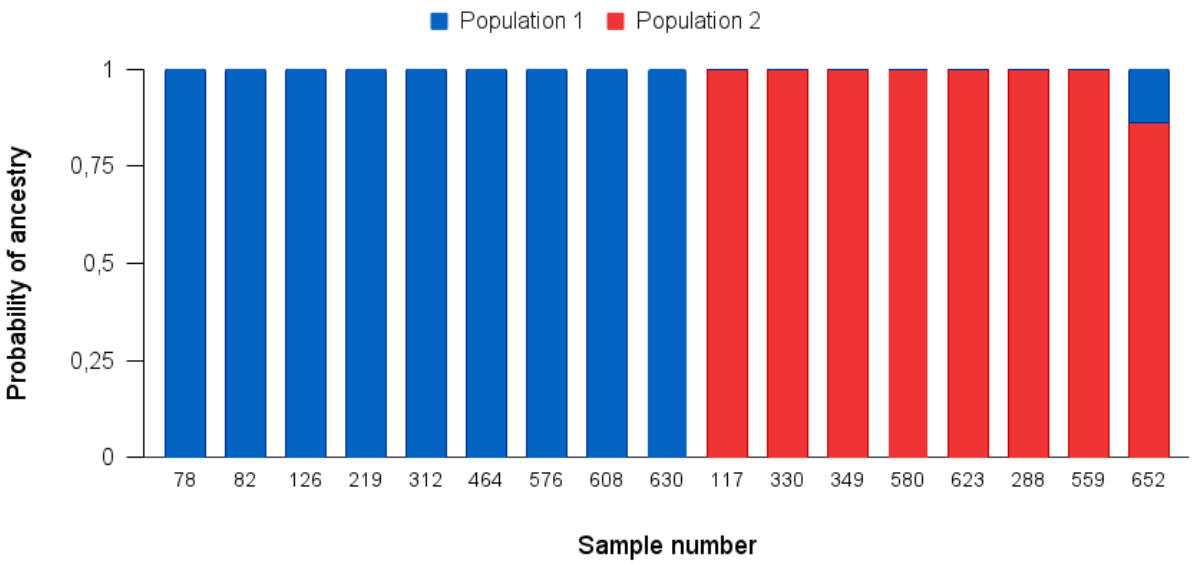

Fig 1. Probability of ancestry of each *Diplolepis rosae* sample using ancestry proportions inferred by fastStructure algorithm for K = 2.

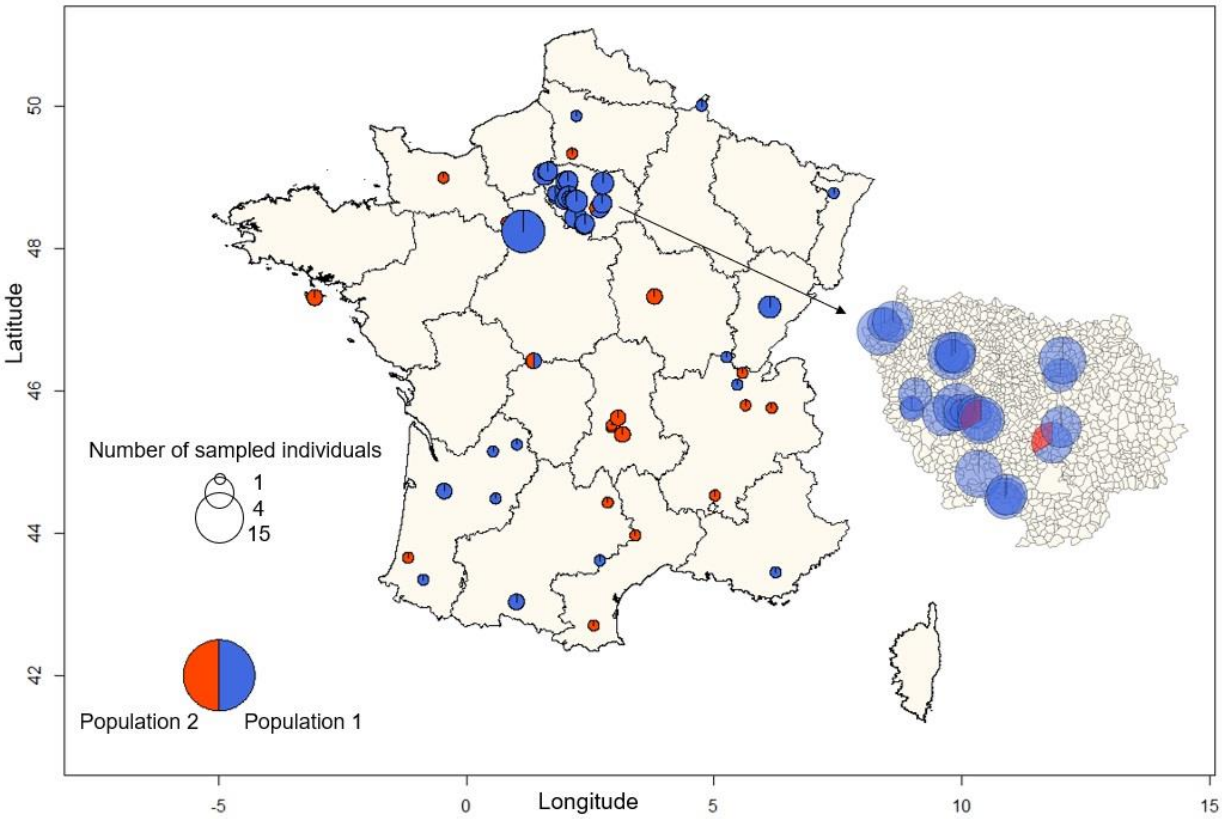

Fig 2. Geographical distribution of *Diplolepis rosae* populations in metropolitan France according to one population-specific genetic marker.

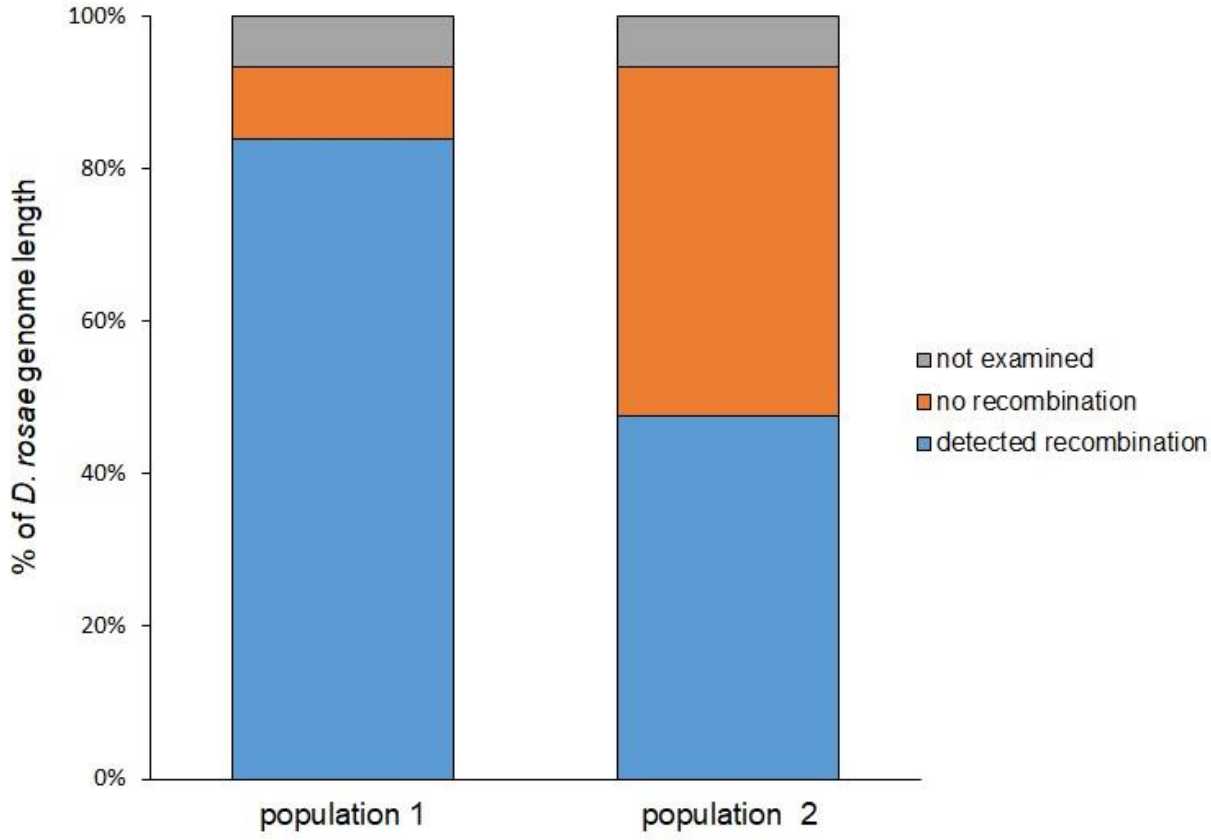

Fig. 3. Percentage of *Diplolepis rosae* genome length where the decay of linkage disequilibrium with distance has been observed.

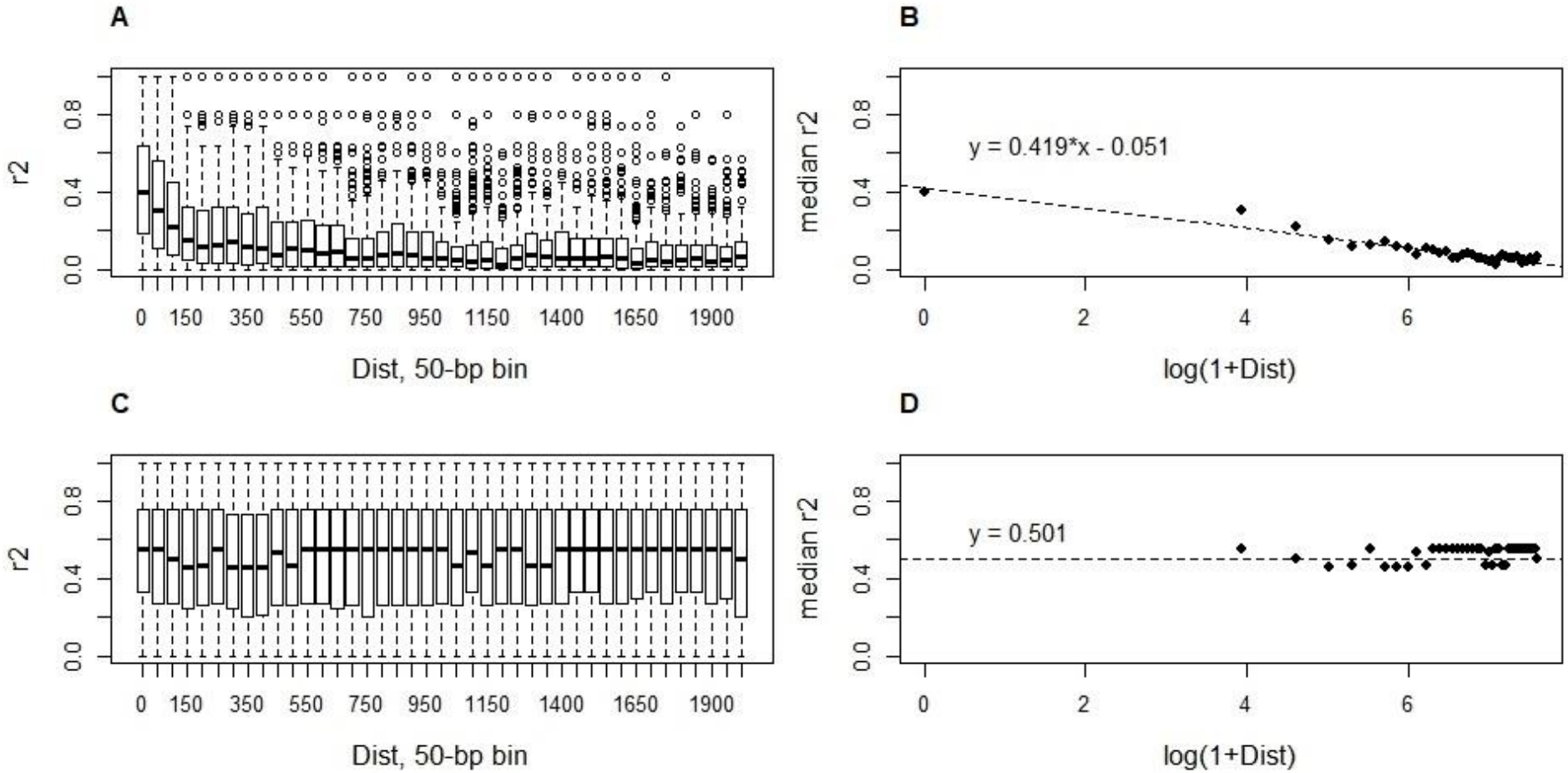

Fig. 4. **A**: relation between r^2 statistics of linkage disequilibrium (LD) and the distance between two SNPs calculated for the contig_264 of the *Diplolepis rosae* genome assembly. **B**: relation between the median r^2 and log(1+Dist). The dotted line represents the linear fit for y = a + b*x, with y = a + (median r^2)*b, b < 0 (*lm* R function): r^2 = 0.419 - 0.051*Dist (F-statistic = 367 on 1 and 39 degrees of freedom, p-value < 0.0001; a = 0.419 ± 0.018, p < 0.0001; b = -0.051 ± 0.003, p < 0.0001; multiple R-squared = 0.904, adjusted R-squared = 0.901). **C**: relation between r2 statistics of linkage disequilibrium (LD) and the distance between two SNPs calculated for the contig_90 in the *Diplolepis rosae* genome assembly. **D**: the dotted line represents the fit for y = a + b*x, with y = a + (median r^2)*b, b < 0 (*lm* R function): r^2 = 0.501 + 0.004*Dist (F-statistic = 0.817 on 1 and 39 degrees of freedom, p-value = 0.372; a = 0.501 ± 0.032, p < 0.0001; b = 0.004 ± 0.005, p = 0.372; multiple R-squared = 0.0205, adjusted R-squared = -0.00459). Dist: distance, 50-bp bins.

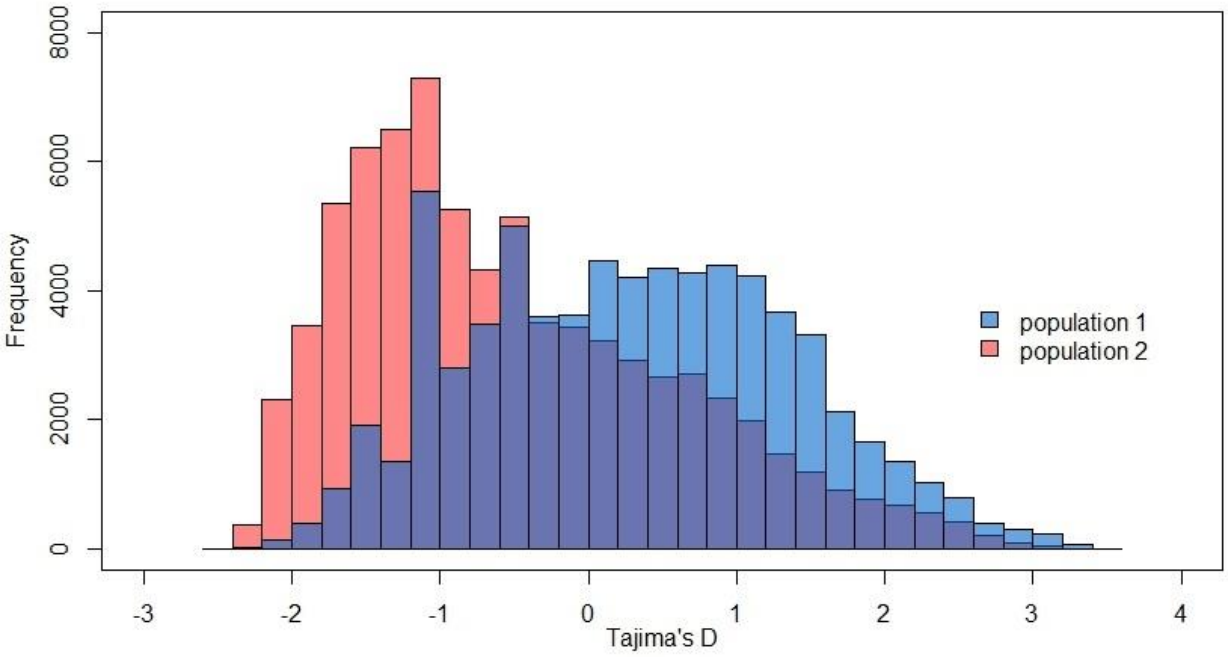

Fig. 5. Distribution of Tajima's D values in *Diplolepis rosae* for population 1 and population 2.

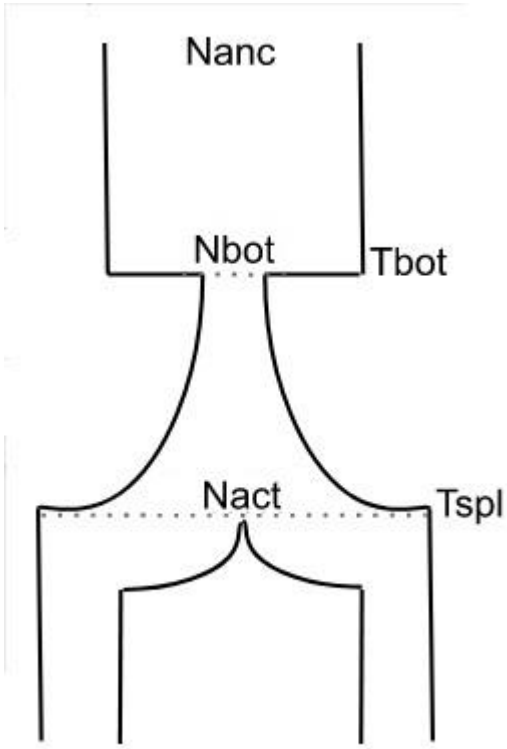

Fig 6. Schema describing the best-fitted model for the demographic scenario of *Diplolepis rosae*: a bottleneck of the ancestral population followed by exponential growth then split into two populations with no gene flow between them. Nanc, Nbot, Nact: population sizes (ancestral, after the bottleneck, just before the split, respectively); Tbot, Tspl: times (bottleneck, split, respectively). Line sizes are scaled relatively to *dadi* estimates: Nbot/Nanc = 0.13; Nact/Nanc = 2.4; bottleneck time = 0.22; split time = 0.20; time unit is 2*Neff generations ago; Neff: effective population size (diploid individuals).

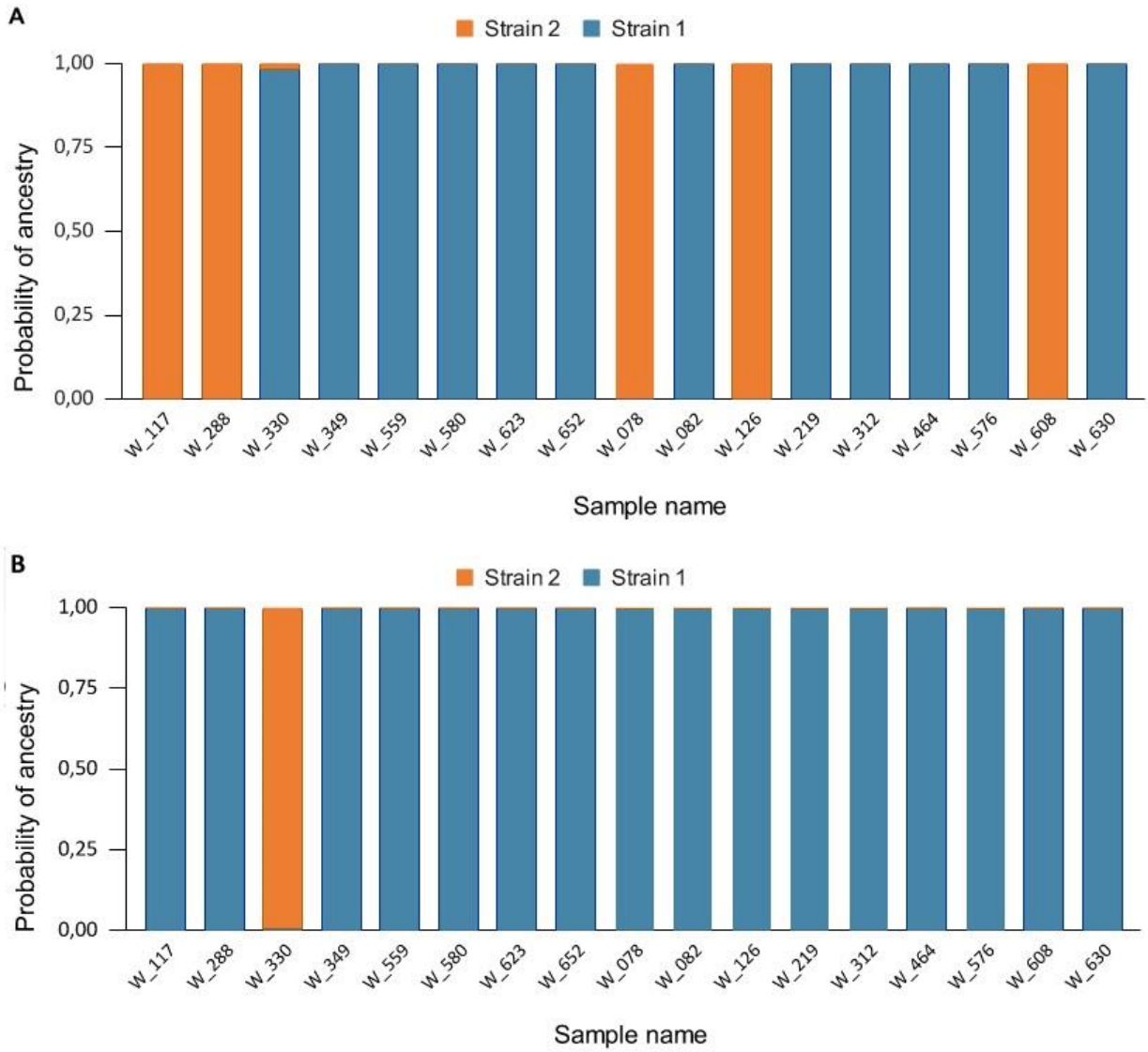

Fig 7. Probability of the ancestry for *Wolbachia* strains using ancestry proportions inferred by fastStructure algorithm for K = 2 strains. *Wolbachia* sequences were recovered from raw *.fastq Diplolepis rosae* reads and aligned to reference genomes of *Wolbachia* Super-group A (**A**) and *Wolbachia* Super-group B (**B**) obtained from the NCBI (accessions are CP042445.1 and CP016430.1, respectively). *Wolbachia* sample numbers correspond to *Diplolepis rosae* sample numbers in **Fig. 1**.

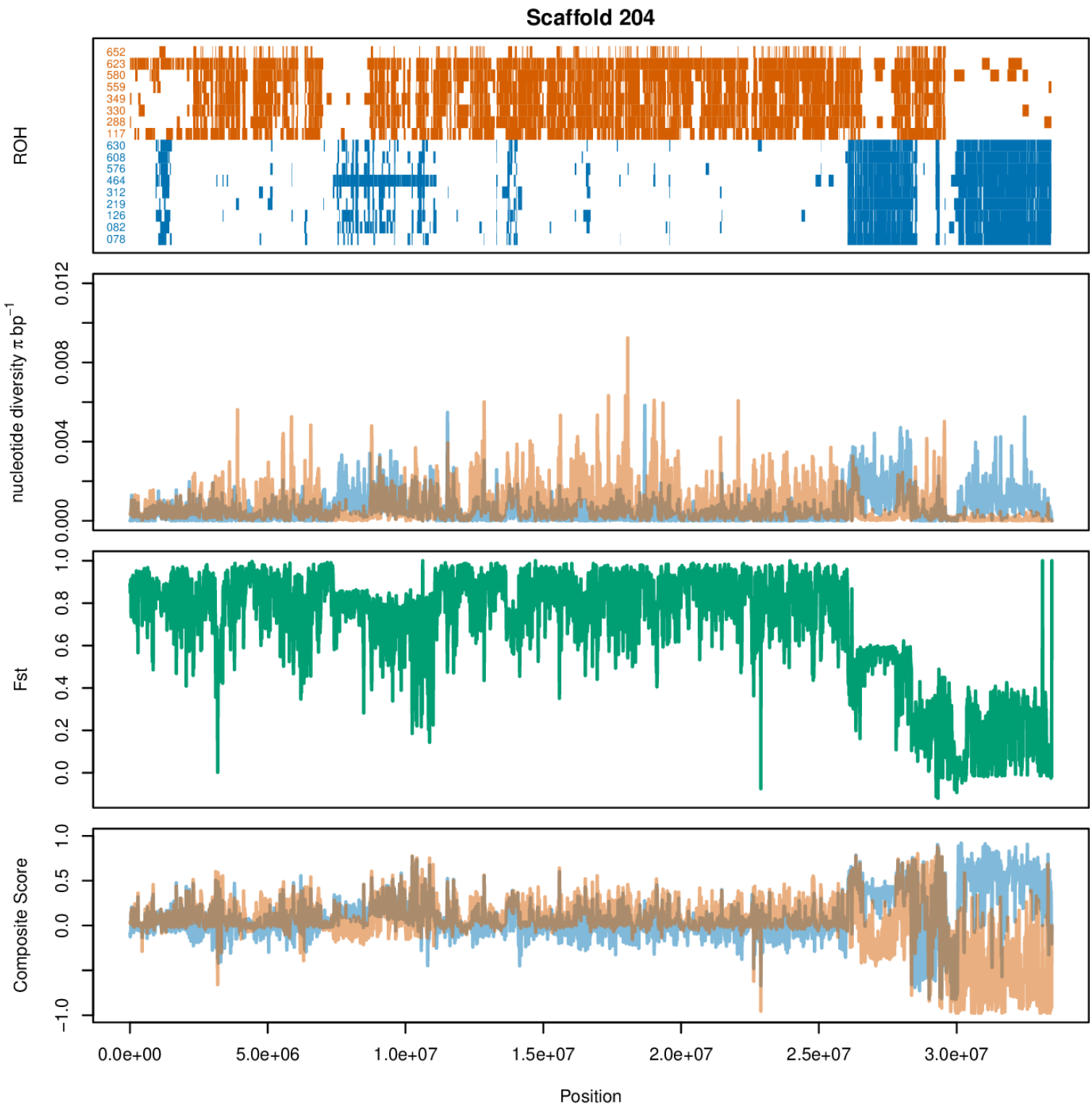

Fig. 8. Runs of homozygosity (ROH), scan of Fst, nucleotide diversity π, and composite score in scaffold_204 of the *Diplolepis rosae* genome. Composite score equal to (1 - Fst)*2(F(π) - 0.5), where F(π) is the cumulative distribution function. Values more than 0.5 show balancing selection and values less than -0.5 show negative selection. Genome position is presented in bp. Blue: population 1; orange: population 2. Numbers in the ROH plot (y-axis) show *Diplolepis rosae* individual names. The region between 30 Mbp and the end of the scaffold shows extended ROH, low Fst, and high π in population 1 (blue) typical of balancing selection (composite score < -0.5), and yet no ROH, low π in population 2 (orange) typical of negative selection.

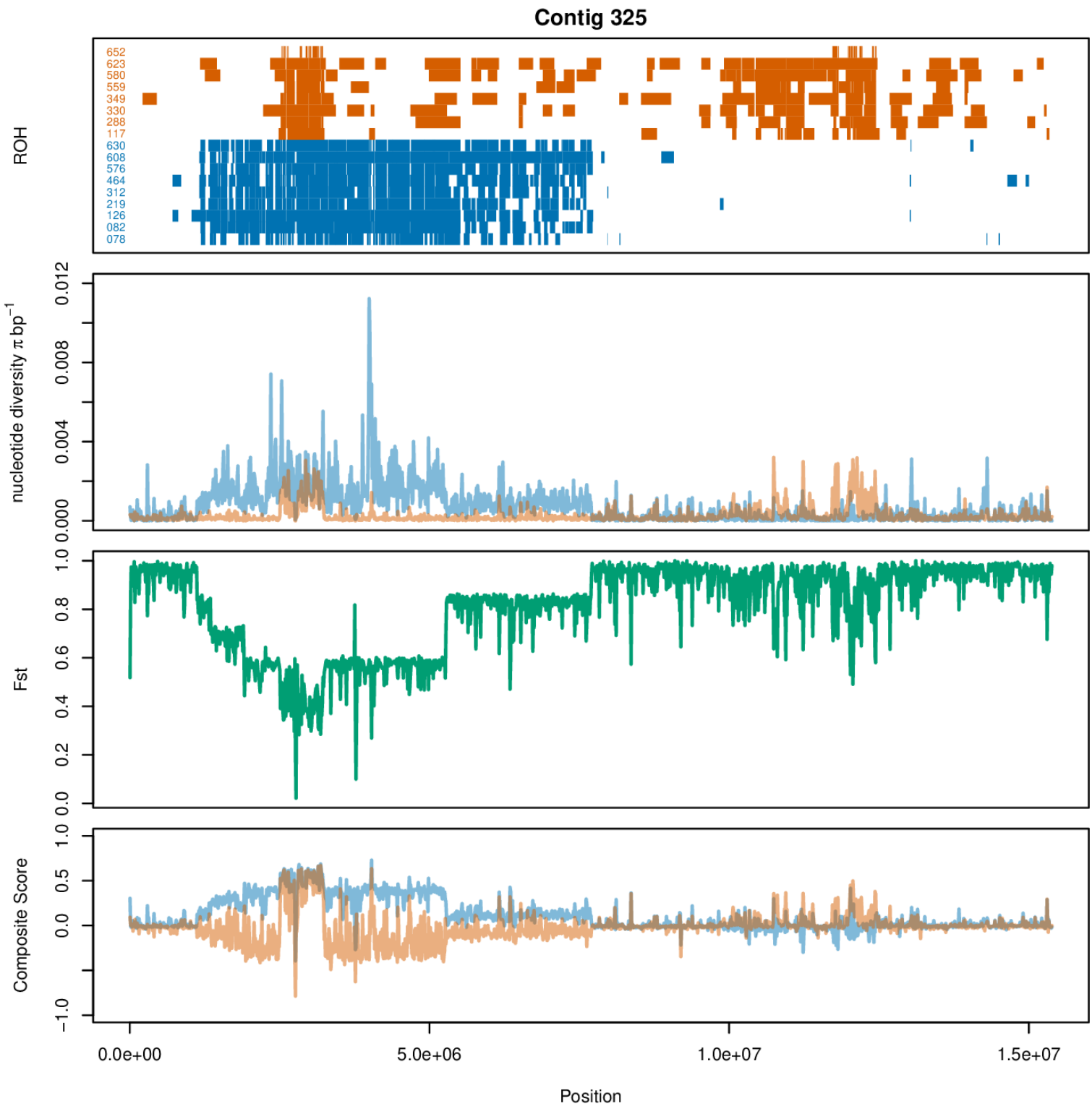

Fig. 9. Runs of homozygosity (ROH), scan of Fst, nucleotide diversity π, and composite score in contig_325 of the *Diplolepis rosae* genome. Composite score equal to (1 - Fst)*2(F(π) - 0.5), where F(π) is the cumulative distribution function. Values more than 0.5 show balancing selection and values less than -0.5 show negative selection. Genome position is presented in bp. Blue: population 1; orange: population 2. Numbers in the ROH plot (y-axis) show *Diplolepis rosae* individual names. The region between 3.1 and 3.8 Mb shows extended ROH with high π and low Fst in both populations, typical of balancing selection.