Data in Brief 36 (2021) 106967

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Aphis gossypii/*Aphis frangulae* collected worldwide: Microsatellite markers data and genetic cluster assignment

Pascale Mistral[a], Flavie Vanlerberghe-Masutti[b], Sonia Elbelt[a], Nathalie Boissot[a,*]

[a] *INRAE, GAFL, 84143 Montfavet, France*
[b] *CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France*

ARTICLE INFO

Article history:
Received 2 February 2021
Revised 11 March 2021
Accepted 12 March 2021
Available online 18 March 2021

Keywords:
Aphis gossypii
Aphis frangulae
Microsatellite
Population structure
Diversity
Host plant
Aphid

ABSTRACT

Aphis gossypii is a cosmopolitan aphid species able to colonize hundreds of plant species from various families [1]. It causes serious damage to a wide range of crops and it is considered a major pest of cucurbits and cotton [2]. It reproduces clonally, by obligate parthenogenesis, on secondary hosts present throughout the year in the intertropical area. At higher latitude, some lineages clonally overwinter but part of the population may have a sexual reproduction in autumn on primary host such as *Hibiscus syriacus*, to generate cold resistant overwintering eggs [3]. It is highly challenging to distinguish *A. gossypii* from its sister species *Aphis frangulae* as both are colonizing solanaceous plants as secondary hosts but the primary host of *A. frangulae* is *Frangula alnus* [4]. This paper describes a worldwide collection of both species from December 1989 to September 2019. Aphids were collected individually on plants (19 families) or in traps. The location, the morph type and the botanical family of the host plant were registered. DNA was extracted from each aphid and amplified at 8 microsatellite loci [5]. Amplicons were analysed with ABI technology and their size was defined with Genemapper software. We named each unique combination of alleles, called a multilocus genotype (MLG), and then each individual was given its MLG. The matrix of alleles of all MLGs was run

* Corresponding author.
E-mail address: nathalie.boissot@inrae.fr (N. Boissot).

https://doi.org/10.1016/j.dib.2021.106967

for a Bayesian analysis to describe the genetic structure of the diversity collected and then each MLG had a probability to belong to a genetic group [6,7]. Probability of assignation to each genetic group revealed by the analysis was reported to each individual according to its MLG.
This dataset can be used to analyze host plant specificities in *A. gossypii*, genetic diversity in *A. gossypii* and relative incidence of variants in diverse geographical regions, admixture between two sister species (*Aphis gossypii* and *Aphis frangulae*).

Specifications Table

Subject	Genetics
Specific subject area	Genetic diversity of a pest of crops on which it clonally reproduces, but has potentially one sexual generation per year. Some clones are observed on several continents.
Type of data	Table
How data were acquired	Microsatellites amplification by multiplex PCR Amplicon separation with ABI 3100 Genetic Analyser 3730XL, Size of amplicons determined with Genemapper software Check data quality with R-script Assignment of each individual to a genetic cluster via a Bayesian analysis with Structure software
Data format	Raw Analysed
Parameters for data collection	Date, host plant, Serial sampling in melon/cotton fields grown in areas with high density of cultivated host plant, Serial sampling over year in a specific geographical region, Sampling on a same host plant in very remote areas.
Description of data collection	The dataset is composed of: • Aphid collection information Sampling characteristics: date, country, GPS point, Host on which aphids were collected: Botanical Family, genus, species Sample characteristics: Morph, Sex • 16 microsatellite length (both alleles at 8 microsatellite markers (Ago126, Ago24, Ago53, Ago59, Ago66, Ago69, Ago84, Ago89) • The corresponding MultiLocus Genotype (MLG) name (given for an allelic combination) • The probabilities to belong to each of the 9 genetic clusters according to a Bayesian clustering (Structure software results)
Data source location	There were 129 sampling localities, distributed in 17 countries, see Supplemental 4 for details
Data accessibility	https://doi.org/10.15454/HNGGMX

Value of the Data

- Microsatellites data obtained from different teams cannot be pooled for analyses because data may be lab-dependent. We gathered a large set of microsatellites data obtained by a unique team for a very important pest, *Aphis gossypii*, attacking crops worldwide [1,2]. The sampling was done by different collaborative partners.

- The data can benefit to any researchers working on aphids, and to any professor/students looking for a large set of data for population genetics analysis in an organism with a complex reproduction system such as *A. gossypii* [3].
- The data set might be used to investigate i/ the sound of *Aphis gossypii*/*Aphis frangulae* differentiation [4], ii/ the role of worldwide clones as crop pest, iii/ the importance of sexual reproduction in *Aphis gossypii* population in a given area, iv/ the relationship between genetic group and specialization on host plant families.
- Moreover, we propose to send reference DNAs to any researcher who runs a genetic population analysis in *A. gossypii* with microsatellites and would like to integrate/use the large database given here. The reference MLGs are NM1, C6, C9, CUC1, GWD, Pot1, PsP4 and Burk1.

1. Data Description

The dataset gathers information on 16,016 individuals *Aphis gossypii*/*Aphis frangulae* collected in 17 countries from 1989 to 2019 from all continents. Aphids collected were either apterous (10,177), winged (5654) or eggs (15). They were collected on 19 plant families or in traps deposited in melon fields.

For each individual, primary data collected are country with GPS position (aphids were collected in the 10 km around this position, except in Australia where aphids were collected up to 70 km from this position), date, morph, host plant or trap. For each aphid collected the second type of data consists in the allelic composition at 8 microsatellites and its corresponding MultiLocus Genotype name, and last the probability of assignation to a genetic group given by a Bayesian analysis.

Table 1 summarizes the number of individuals collected according to their host plant species and their morph.

Table 2 gives the common MLGs between three geographical areas.

Fig. 1 pictures the sample locations on a world map and the number of individuals collected in each country.

Table 1
Number of aphids collected on 19 botanical families of host plant and their morph.

	Morph				
Botanical family / trap	Apterous	Winged	Egg	Unknown	Total
Asteraceae	186				186
Bignoniaceae	5				5
Brassicaceae	4				4
Chenopodiaceae	4				4
Convolvulaceae	15				15
Cucurbitaceae	8641	3668		70	12,403
Euphorbiaceae	14				14
Fabaceae	18				18
Lamiaceae	19	1			20
Liliaceae	5				5
Lythraceae	5				5
Malvaceae	843				843
Polygonaceae	5				5
Portulacaceae	5				5
Rhamnaceae	139	36	15	100	290
Rosaceae	10				10
Rutaceae	60				60
Solanaceae	182				182
Zygophyllaceae	9				9
Trap	8	1876			1884
Unknown		73			73
Total	10,177	5654	15	170	16,016

Table 2
MLGs shared by aphids collected in several geographic areas. 1/ Europe, Tunisia, Turkey, 2/ Benin, Burkina, Cameroon, Senegal, 3/ Brazil, California, West Indies and 4/ Vietnam, Thailand, Australia.

	Benin, Burkina, Cameroon, Senegal	Brazil, California, West Indies	Vietnam, Thailand, Australia
	26 MLGs	13 MLGs	24 MLGs
Europe, Tunisia, Turkey 2222 MLGs	Aub1, Aub2	Al11, Al11–27, Al12–12	
	Burk1, Burk2, Burk3,	Aub3, Aub4, Aub5	
	Burk4, Burk5, Burk7	Burk1	
	C1, C4, C5, C8, C9, C14	C5, C9	C4, C12, C13
		GWD	
	Hib3, Hib5, Hib6, Hib7,	Hib1	Hib4
	Hib8	M12–42	
	Ivo		NM1
		PsP2, PsP3	
	PsP1, PsP4		

		Netherlands	3		
		France	11024		
California	4	Italy	92		
		Spain	105		
		Tunisia	125		
		Turkey	101		
French West Indies	3485	Burkina Faso	25	Thailand	2
		Benin	6	Vietnam	12
		Cameroon	70		
		Senegal	12		
Brazil	6	Madagascar	1	Australia	936

Fig. 1. Distribution of aphids sampling (made from https://fr.batchgeo.com) and number of aphids collected within countries. Aphids were collected within a 10 km radius of the point (70 km for Australia) for the red points, locality unknown for Madagascar (yellow point). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

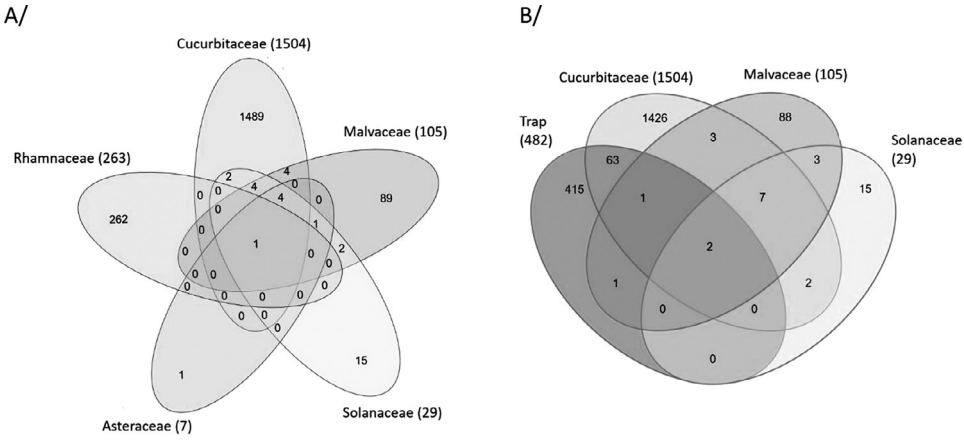

Fig. 2. A) Number of MLGs identified in aphids according to their host plant families (representing 88% of the collected aphids). **B)** Number of MLGs identified in aphids trapped in melon fields and aphids collected on host plants belonging to Cucurbitaceae, Malvaceae and Solanaceae families (representing 95,4% of the collected aphids). Aphids trapped had only common MLGs with aphids collected on plants belonging to these three families. Venn diagrams are made from http://www.interactivenn.net.

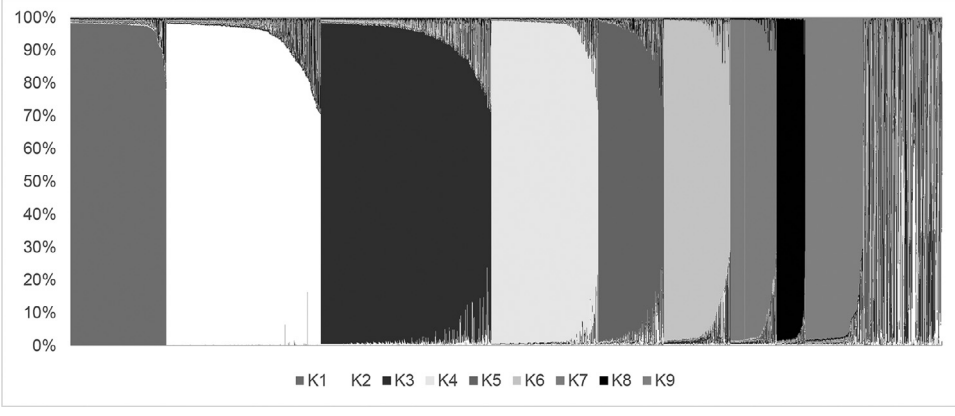

Fig. 3. Percentage of MLG assignment to nine clusters resulting from a Bayesian clustering analysis of the assignment of 2360 MLGs to 1 to 10 clusters (K), with 10 simulations per K. 260 out of the 261 expected MLGs for individuals belonging to *Aphis frangulae*, i.e. identified in aphids collected on *Frangula alnus*, were assigned to the first cluster with a probability over 0.77. These MLGs are characterized by null alleles at Ago24 and Ago84 and are homozygous for the 110 allele at Ago53.

Fig. 2 contains two Venn diagrams illustrating A) the number of MLG observed in individuals collected on five botanical families (98,7% of the collected aphids) and B) the number of MLG observed in individuals collected in traps and on Cucurbitaceae, Malvaceae and Solanaceae plants. These three host families were chosen because they hosted individuals with MLGs common with those in traps.

Fig. 3 gives for each MLG the percentage assignation to the nine clusters. It is the results of a Bayesian clustering after analyzing the assignment of 2358 MLGs to 1 to 10 clusters (K), with 10 simulations per K. 260 out of the 261 expected MLGs for *Aphis frangulae*, i.e. identified

in individuals collected on *Frangula alnus*, were assigned to the first cluster with a probability over 0.77. These MLGs are characterized by null alleles for Ago24 and Ago84 and an homozygous 110–110 alleles for Ago53.

Supplemental 1 Primer sequences for microsatellite amplification

Supplemental 2 R script for checking MLG consistency

Supplemental 3 Results of computation to determine the most likely number of genetic clusters in the data set according to a Bayesian analysis and corresponding graphics

Supplemental 4 List of sampling locations (GPS coordinates and countries)

2. Experimental Design, Materials and Methods

2.1. Aphids sampling

Aphids, winged or apterous, visually expected as belonging to the *Aphis gossypii* species group, were removed with a brush from their host plant and individually sunk in 70–90% ethanol in a numbered tube. On *Frangula alnus*, the primary host of *Aphis frangulae,* we also collected aphid eggs. Winged individuals collected on *Frangula alnus* were examined for their genitalia under a binocular microscope to separate males from females. Aphids were also collected in non-biased suction traps designed to sample winged insects daily at the crop height [8]. The traps were deposited in melon fields during spring and summer. Samples were stored few days to several weeks before DNA extraction.

2.2. DNA extraction

Ethanol was removed and 50 µl of 5% chelex 100 (Chelex 100, Bio-rad) were added to each tube. Aphids were coarse grinded and submitted to a thermal shock at 56° for 30 min and 95–100° for 5 min. After a short centrifugation (2000 × rpm for 30 s) to pellet debris and Chelex beads, DNA remains in the supernatant which can be stored a few days at −20°C before PCR amplification.

2.3. Microsatellite amplification

The primer sequences, amplifying eight microsatellite loci specific of the *A. gossypii* genome [5], are given in the supplemental 1. The forward primer of each microsatellite locus was labelled with a fluorescent dye (FAM, NED, PET, VIC) chosen to analyze simultaneously the eight microsatellite loci (Ago24-FAM, Ago53-VIC, Ago59-NED, Ago66-VIC, Ago69-NED, Ago84-PET, Ago89-PET and Ago126-FAM). The primers labeled by NED, PET and VIC fluorochromes were supplied by Applied Biosystems™ (https://www.fishersci.fr), all other primers were supplied by Eurofins Genomics (https://eurofinsgenomics.eu/). DNA amplifications were performed in two polymerase chain reactions (PCR) in a final volume of 5 µL in a thermocycler (Mastercycler, Eppendorf). The first PCR is a multiplex of Ago53, Ago59, Ago66, Ago69, Ago84, Ago89 and Ago126, containing 2.5 µL of QIAGEN Multiplex PCR Master dNTPmix, 5 units/µl of HotStart-Taq DNA polymerase and 3 mM $MgCl_2$, 0.2 µm of each primer, 1 µL of Chelex DNA extraction diluted at 1/10 and RNAase free water to supplement at 5 µL. Amplifications were performed according to a programming with 15 min at 95°C followed by 25 cycles of 30 s at 95°C, 90 s at 56°C and 30 s at 72°C and a final extension during 30 min at 60°C. The second PCR, using the primer specific of the locus Ago24, was performed in the same conditions except for the primer concentration (0.1 µm) and the thermocycler programming: 5 min at 95°C, 35 cycles of 30 s at 95°C, 45 s at 62°C and 30 s at 72°C, and a final elongation of 7 min at 72°C.

2.4. Microsatellite analyses

A mix containing 10 µL of Hi-Di Formamide (Applied Biosystems) and 0.15 µL of GeneScan-500LIZ Size Standard was deposited in each well of a specific plate for automatic sequencer (ABI 3100 Genetic Analyser, 3730XL), and then 1 µL of each of the two PCRs was added in the well and submitted to denaturation in a thermocycler (95°C 3 min – 4°C 10 min). Separation and detection of PCR products were carried out by a capillary electrophoresis with an automatic sequencer (ABI 3100 Genetic Analyser, 3730XL). We determined the size of the allele at each locus by comparison with GeneScan-500LIZ Size Standard with GeneMapper v3.7 software (Applied Biosystems, Foster City, California, USA). When only one peak was observed for a locus, the locus was considered homozygous. Moreover, aphids with a known combination of alleles (collected in rearings available in the lab), were used too, helping in the calibration of the reading on Genemapper. These controls were reinforced when we changed the device, its capillars or polymers for migration. Cross reading was done anytime a changing in the reader occurred overtime: current reader shared expertise for allele size determination with the new reader.

All individuals collected on *Frangula alnus* did not amplified the microsatellites Ago24 and Ago84 while they were expected belonging to *Aphis frangulae* species. Because the set of microsatellites were defined for *Aphis gossypii* species, we assumed that individuals collected on *Frangula alnus* carry two null alleles for both microsatellites Ago24 and Ago84; these alleles were coded 0.

Only individuals for which at least six out of the eight microsatellites were amplified were kept in the data set. To minimize the risk of miss-reading in Genemapper, we checked samples for which combination of alleles was observed only once. For those combinations, we assumed a miss-reading when the combination differs from any other one in size alleles for only one DNA base at one of the 16 alleles. As far as possible we checked this assumption by a second reading in Genemapper. Then, we corrected the allele size and we assigned a MultiLocus Genotype name (MLG) to each unique combination of alleles.

All the process was checked by the R script given in the supplemental 2.

2.5. Population structure analysis

The 2358 different MLGs were subjected to a Bayesian clustering [6], using an admixture model with a burn-in of 250,000 iterations and a subsequent Markov Chain of 500,000 iterations. For each putative number of clusters (*K*, ranging from 1 to 10), we compared 10 replicate runs, to assess the consistency of the estimated values. We used the Evanno method to determine the most likely number of genetic clusters [7]. For each K, the $L(K)_i$, i.e. estimated Ln Prob of Data in the results file given by the structure software for the run i, was collected, and the $L"(K)_i = L(K)_i - 2L(K\text{-}1)_i + L(K\text{-}2)_i$ was calculated. The mean of L(K) and L(K)" were plotted (see supplemental 3). The likeliest number of K (for K>2) is given by the peak of L(K)".

For the present data set, the likeliest number of K was equal to nine, we then performed one run of the admixture model with a burn-in of 500,000 iterations and a Markov Chain of 1000,000 iterations. Probabilities of assignation of each MLG to the nine clusters were obtained and plotted in Fig. 3.

Ethics Statement

No concern.

CRediT Author Statement

Pascale Mistral: has organized and participated to all sampling since 2005. She organized the dataset for sampling characteristics. She produced most of the genetic data and their curation; **Flavie Vanlerberghe-Masutti:** coordinated or participated to all successive projects for which aphids were sampled, then she was highly involved in funding acquisition. She reviewed and edited the data paper; **Sonia Elbelt:** worked on curation/validation of the dataset. She built the figures and tables, annotated the R-script and wrote the original draft; **Nathalie Boissot:** coordinated or participated to successive projects for which aphids were sampled since 2004, then she was highly involved in funding acquisition. She supervised the data curation/validation and carried on the Bayesian analysis. She reviewed and edited the data paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

We thank Joel Chadoeuf (INRAE-GAFL) for helping in the R script writing. Aphids sampling and genotyping data presented here were part of studies mainly supported by the Departments BAP and SPE of INRAE, by the French Ministère des Affaires Etrangères, the French Ministère de la Recherche and the French Ministère of Agriculture.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.106967.

References

[1] T.A. Ebert, B. Cartwright, Biology and ecology of *Aphis gossypii* Glover (Homoptera: aphididae), Southwest. Entomol. 22 (1997) 116–153.
[2] H.F. van Emden, R. Harrigton, Aphids as Crop Pests, CABI, Oxfordshire, UK, 2007.
[3] S. Thomas, N. Boissot, F. Vanlerberghe-Masutti, What do spring migrants reveal about sex and host selection in the melon aphid? BMC Evol. Biol. 12 (2012), doi:10.1186/1471-2148-12-47.
[4] R.L. Blackman, V.F. Eastop, Taxonomic issues, in: H.F. van Emden, R. Harrigton (Eds.), Aphids as Crop Pests, CABI, Oxfordshire, UK, 2007, pp. 1–29.
[5] F. Vanlerberghe-Masutti, P. Chavigny, S.J. Fuller, Characterization of microsatellite loci in the aphid species *Aphis gossypii* Glover, Mol. Ecol. 8 (1999) 693–695, doi:10.1046/j.1365-294x.1999.00876.x.
[6] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[7] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (2005) 2611–2620.
[8] A. Schoeny, P. Gognalons, Data on winged insect dynamics in melon crops in southeastern France, Data Br. 29 (2020), doi:10.1016/j.dib.2020.105132.