(Article begins on next page)

06/05/2024 21:13

Journal Pre-proof

Soybean Aphid Biotype 1 Genome: Insights into the invasive biology and adaptive evolution of a major agricultural pest.

Rosanna Giordano, Ravi Kiran Donthu, Aleksey Zimin, Irene Consuelo Julca Chavez, Toni Gabaldon, Manuella van Munster, Lawrence Hon, Richard Hall, Jonathan Badger, Minh Nguyen, Alejandra Flores, Bruce Potter, Tugrul Giray, Felipe N. Soto-Adames, Everett Weber, Jose A.P. Marcelino, Christopher J. Fields, David J. Voegtlin, Curt B. Hill, Glen L. Hartman, Soybean aphid research community

PII: S0965-1748(20)30023-0

DOI: https://doi.org/10.1016/j.ibmb.2020.103334

Reference: IB 103334

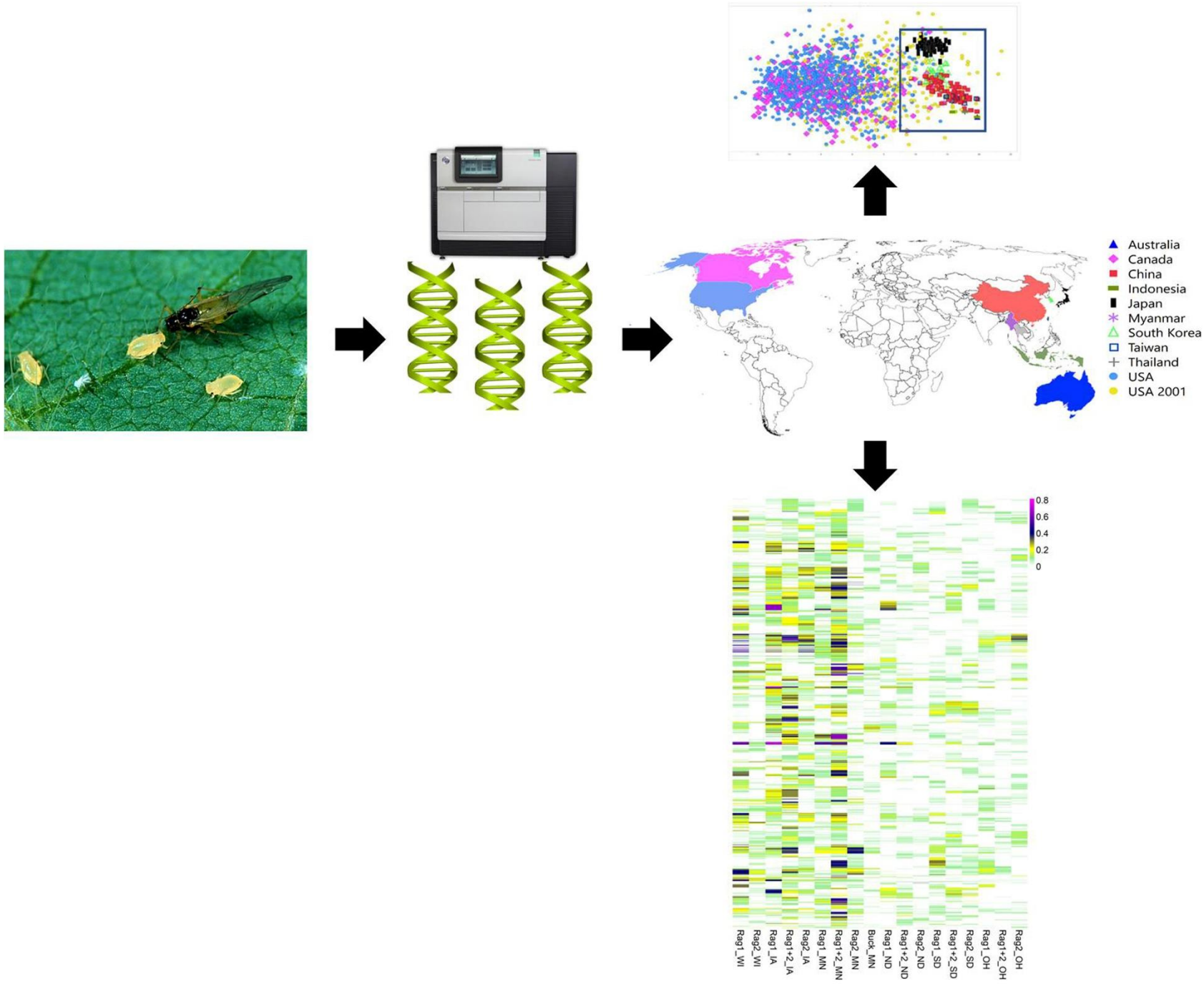
To appear in: *Insect Biochemistry and Molecular Biology*

Received Date: 22 November 2019

Revised Date: 7 January 2020

Accepted Date: 10 February 2020

Please cite this article as: Giordano, R., Donthu, R.K., Zimin, A., Julca Chavez, I.C., Gabaldon, T., van Munster, M., Hon, L., Hall, R., Badger, J., Nguyen, M., Flores, A., Potter, B., Giray, T., Soto-Adames, F.N., Weber, E., Marcelino, J.A.P., Fields, C.J., Voegtlin, D.J., Hill, C.B., Hartman, G.L., Soybean aphid research community, Soybean Aphid Biotype 1 Genome: Insights into the invasive biology and adaptive evolution of a major agricultural pest., *Insect Biochemistry and Molecular Biology*, https://doi.org/10.1016/j.ibmb.2020.103334.


Australia
Canada
China
Indonesia
Japan
Myanmar
South Korea
Taiwan
Thailand
USA
USA 2001
0.8
0.6
0.4
0.2
0
Rag1_WI
Rag2_WI
Rag1_IA
Rag1+2_IA
Rag2_IA
Rag1_MN
Rag1+2_MN
Rag2_MN
Buck_MN
Rag1_ND
Rag1+2_ND
Rag2_ND
Rag1_SD
Rag1+2_SD
Rag2_SD
Rag1_OH
Rag1+2_OH
Rag2_OH

Title

Soybean Aphid Biotype 1 Genome: Insights into the invasive biology and adaptive evolution of a major agricultural pest.

Rosanna Giordano[1,2,Δ], Ravi Kiran Donthu[1,2,Δ], Aleksey Zimin[3], Irene Consuelo Julca Chavez[4,5,6], Toni Gabaldon[4,5,6,7], Manuella van Munster[8], Lawrence Hon[9], Richard Hall[10], Jonathan Badger[11], Minh Nguyen[12], Alejandra Flores[13], Bruce Potter[14], Tugrul Giray[15], Felipe N. Soto-Adames[16], Everett Weber[2], Jose A.P. Marcelino[1,2,17], Christopher J. Fields[18], David J. Voegtlin[19], Curt B. Hill[20], Glen L. Hartman[21], Soybean aphid research community*

[1] Puerto Rico Science, Technology and Research Trust, San Juan, PR

[2] Know Your Bee, Inc. San Juan, PR

[3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

[4] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[5] Barcelona Supercomputing Centre (BSC-CNS), Barcelona, Spain

[6] Institute for Research in Biomedicine, Barcelona, Spain

[7] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[8] CIRAD-INRA-Montpellier SupAgro, TA A54/K, Campus International de Baillarguet, Montpellier, France

[9] Color Genomics, Burlingame, CA, USA

[10] Pacific Biosciences, Menlo Park, CA, USA

[11] Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institute of Health, DHHS, Bethesda, MD, USA

[12] Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA

[13] College of Liberal Arts and Sciences, School of Molecular and Cellular Biology, University of Illinois, Urbana, IL, USA

[14]University of Minnesota, Southwest Research and Outreach Center, Lamberton, MN, USA

[15] Department of Biology, University of Puerto Rico, San Juan, PR, USA

[16] Florida Department of Agriculture and Consumer Services, Division of Plant Industry, Entomology, Gainesville, FL, USA

[17] Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA

[18] HPCBio, Roy J. Carver Biotechnology Center, University of Illinois, Urbana, IL, USA

[19] Illinois Natural History Survey, University of Illinois, Urbana, IL, USA

[20]Agricen Sciences, Pilot Point, Texas, USA

[21] USDA-ARS and Department of Crop Sciences, University of Illinois, Urbana, IL, USA

[Δ]These authors contributed equally to this work.

*Names and affiliations are listed in the appendix.

Corresponding authors: Rosanna Giordano (rgiordano@prsciencetrust.org; rgiordano500@gmail.com) and Ravi Kiran Donthu (rkiran@prsciencetrust.org; donthuanalyst@gmail.com), Puerto Rico Science, Technology; and Know Your Bee, Inc., San Juan, PR.

Abstract

The soybean aphid, *Aphis glycines* Matsumura (Hemiptera: Aphididae) is a serious pest of the soybean plant, *Glycine max*, a major world-wide agricultural crop. We assembled a *de novo* genome sequence of *Ap. glycines* Biotype 1, from a culture established shortly after this species invaded North America. 20.4% of the *Ap. glycines* proteome is duplicated. These in-paralogs are enriched with Gene Ontology (GO) categories mostly related to apoptosis, a possible adaptation to plant chemistry and other environmental stressors. Approximately one-third of these genes show parallel duplication in other aphids. But *Ap. gossypii*, its closest related species, has the lowest number of these duplicated genes. An Illumina GoldenGate assay of 2,380 SNPs was used to determine the world-wide population structure of *Ap. Glycines*. China and South Korean aphids are the closest to those in North America. China is the likely origin of other Asian aphid populations. The most distantly related aphids to those in North America are from Australia. The diversity of *Ap. glycines* in North America has decreased over time since its arrival. The genetic diversity of *Ap. glycines* North American population sampled

shortly after its first detection in 2001 up to 2012 does not appear to correlate with geography. However, aphids collected on soybean *Rag* experimental varieties in Minnesota (MN), Iowa (IA), and Wisconsin (WI), closer to high density *Rhamnus cathartica* stands, appear to have higher capacity to colonize resistant soybean plants than aphids sampled in Ohio (OH), North Dakota (ND), and South Dakota (SD). Samples from the former states have SNP alleles with high F_{ST} values and frequencies, that overlap with genes involved in iron metabolism, a crucial metabolic pathway that may be affected by the *Rag*-associated soybean plant response. The *Ap. glycines* Biotype 1 genome will provide needed information for future analyses of mechanisms of aphid virulence and pesticide resistance as well as facilitate comparative analyses between aphids with differing natural history and host plant range.

1. Introduction

Native to Asia, the soybean plant, *Glycines max* (L.) has been grown in China for 4000-5000 years (Ma, 1984) and its cultivation spread to other Asian countries approximately 2,500 years ago (Wu et al., 2004). The soybean aphid, *Aphis glycines*, native to the same region, is a highly successful organism with a wide geographic distribution. In Asia it can be found over a range that spans from northern China, eastern Russia, Japan, Korea, to the more southern areas of Thailand, Malaysia, Indonesia, the Philippines, Vietnam and Myanmar (Wu et al., 2004; Ragsdale et al., 2004; Krupke et al., 2005). More recently, facilitated by commerce and human movement, it has invaded Australia (Fletcher and Desborough, 2000), the United States and Canada (Venette, 2004; Ragsdale et al., 2004).

Like most aphids, *Ap. glycines* has a life cycle during which both sexual and asexual morphs are produced (holocyclic) on alternating plant hosts (heteroecious). *Rhamnus* sp. constitute the primary host, which the aphid uses to overwinter and reproduce sexually (Blackman and Eastop, 1984). The cultivated soybean plant is used during the summer months, when the parthenogenetic form can reach extremely high population densities. However, other plant species such as *G. soja* Sieb. & Zucc., and other species (Wang et al., 1962; Ragsdale et al., 2004; Hill et al., 2004b) have been reported as summer hosts. During the summer, winged morphs (alates) can develop in response to low host quality, crowding or other stressors. These alates disperse to new host plants locally and in some cases wind aids in long-distance dispersal. Fall, temperatures, photoperiod and changes in soybean host quality trigger the production of winged females that viviparously produces the sexual generation (gynoparae). The gynoparae fly to *Rhamnus* where they feed and give birth to nymphs (oviparae) destined to bear the overwintering eggs. Alate males, produced on senescing soybean, seek the oviparae on *Rhamnus* and mate. Mated oviparae lay fertilized eggs in the folds of *Rhamnus* buds (Ragsdale et al., 2004) (Fig. 1). In Asia, *Rhamnus davurica* Pallus and *R. japonica* Maxim. are most commonly used as overwintering hosts (Takahashi et al., 1993; Kim et al., 2010), while in North America *R. cathartica*, also an invasive species widely diffused in the north–central region of the U.S., is utilized as the overwintering plant host (Voegtlin et al., 2004; Ragsdale et al., 2004).

Similar to many other insects, the most widely used control method for soybean aphid has been the application of chemical pesticides (Hodgson et al., 2012; Ragsdale et al., 2011; Hesler et al., 2013). However, insects have commonly met this challenge by developing resistance to highly used modes of action of insecticidal compounds (Pedigo and Rice, 2009; Mahmood et al., 2014). The soybean aphid is no exception and resistance to organophosphates and pyrethroids has been observed in Asia (Wang et al., 2011a,b; Xi et al., 2015) and North America (Hanson et al., 2017).

The production of soybean in China is mainly located in the north and northeast region and the soybean aphid is the most serious pest threat to productivity (Wu et al., 2004) (A compendium of translated papers regarding past research conducted in China on the soybean aphid is available at http://www.ksu.edu/issa/aphids/reporthtml/citations.html (Wu et al., 2004). In Asia, the soybean aphid, where it has co-existed with the cultivated soybean for several thousand years, has a large number of natural enemies that serve to moderate its populations. These include 15 species of aphelinids and braconids parasitoids, 9 species of hyperparasitoids as well as multiple predators such as anthocorids, chamaemyiids, chrysopids, coccinellids, linyphiids, lygaeids, mirids, nabids, and syrphids (Wu et al., 2004). Within Asia, the soybean aphid inhabits a geographic landscape with highly varied topography including mountains and large bodies of water that could serve as barriers, however, its dispersal was facilitated by human activity and the concomitant dissemination of the soybean plant, an easy to grow source of protein and oil and is now present in much of Asia (Wu et al., 2004).

The recent increase in world-wide commerce and human mobility has facilitated the movement of the soybean aphid beyond the Asian continent, making it one of the most important invasive agricultural insect pests in North America. First observed in July of 2000 on soybean fields in Wisconsin, Illinois and Minnesota (Hartman et al., 2001; Alleman et al., 2002; Venette and Ragsdale, 2004), it rapidly spread to 22 states and three Canadian provinces in 4 years. It has been proposed that it was likely present in the U.S. for several years prior to 2000, but in low numbers that escaped detection and or confirmation (Hunt et al., 2003; Venette and Ragsdale, 2004; Ragsdale et al., 2004). *Ap. glycines* is now established in most of the soybean growing areas of North America and its economic impact in terms of crop loss is significant. In 2001, yield losses greater than 50% were reported in Minnesota. Ragsdale et al. (2007) reported yield losses of 40%, and in 2003 losses were estimated at $80 million in Minnesota and $45 million in Illinois. In 2003 the state of Illinois spent an estimated $9 to $12 million in insecticides to control the soybean aphid. Damage estimates from the soybean aphid, if left untreated, are estimated at $2.4 billion annually (Song et al., 2006). Large aphid populations reduce soybean production directly by causing severe plant damage during feeding, resulting in leaf distortion, stunting, and desiccation. Feeding by a relatively small number of aphids can affect photosynthesis (Macedo et al., 2003). However, soybean aphids also indirectly affect soybean plants by facilitating the growth of black sooty mold fungus that grows on aphid honeydew and inhibits photosynthesis (Malumphy, 1997; Hartman et al., 2001). In addition to direct feeding damage, the soybean aphid transmits several plant viruses such as *Soybean mosaic virus* (SMV), *Soybean dwarf virus* (SbDV), as well as viruses of other

crops such as *Cucumber mosaic virus* (CMV) and *Potato virus Y* (PVY) (Sama et al., 1974; Iwaki et al., 1980; Hartman et al., 2001; Hill et al., 2001; Clark and Perry, 2002; Domier et al., 2003; Davis et al., 2005; Sass et al., 2004). Probe feeding by migrating soybean aphids can transmit viruses to non-hosts such as potato, *Solanum tuberosum* L., (Davis and Radcliffe, 2008) and bean, *Phaseolus spp*., (Mueller et al., 2010).

While there have been efforts to establish environmentally sound biological controls methods (Chacón et al., 2008; Heimpel et al., 2004; Nielsen and Hajek, 2005; Rutledge and O'Neil, 2005; Wu et al., 2004; Wyckhuys et al., 2007) the application of insecticides to reduce soybean aphid populations is the most common management method (Hodgson et al., 2012; Magalhaes, 2008; Myers et al., 2005). For some U.S. states, as much as 57% of soybean acres have been reported as treated with insecticide during outbreak years (Ragsdale et al., 2007). Scouting and insecticide treatments based on economic threshold have been shown to be an economical way to manage soybean aphids with insecticide (Ragsdale et al., 2007; Hodgson et al., 2012; Koch et al., 2016; Ragsdale et al., 2011).

Most aphid species are specialized to feed on a particular plant family or a few plant species within a family (Blackman and Eastop, 2000; Powell et al., 2006). *Ap. glycines* is highly specialized towards soybean and its closest relatives, likely the result of a long period of co-evolution between ancestors of *Ap. glycines* and *Glycine* plant species in their center of origin, probably in present day northwest China (Wu et al., 2004).

The basics of the life cycle of *Ap. glycines*, were constant through the first few years of its establishment in North America (Fig. 1). Soybean was utilized as the summer host and *R. cathartica, R. lanceolata* and *R. alnifolia* as winter host plants (Voegtlin et al., 2004). The latter two species are uncommon natives and not of significance in the year-to-year survival of the soybean aphid in North America (Fig. 1). In 2006 two biological changes were observed in the soybean aphid: the detection of virulent biotypes and the colonization of a new genus of overwintering host plant.

As part of the research effort to limit the impact of *Ap. glycines* on soybean production, a portion of the USDA soybean germplasm collection, housed at the University of Illinois, was tested and several ancestral lines were discovered with host resistance against the soybean aphid (Hill et al., 2004a). From this initial screening, two ancestral soybean lines found to have host resistance genes against the soybean aphid were identified. The resistance in these lines was characterized for mode of action and inheritance. It was found that each line had single, dominant acting genes, *Rag1* (Hill et al., 2006a) and *Rag (Jackson)* (Hill et al., 2006b; Li et al., 2007) that conditioned antibiosis-type resistance against the aphid pest. These genes were subsequently transferred through conventional backcross breeding into elite pre-commercial lines. In 2006, experimental soybean plots of soybean breeding lines with the *Rag1* gene, planted in the field in Ohio, were unexpectedly found to be colonized by soybean aphids. A clonal colony of these aphids was established in the laboratory and tested in a greenhouse on aphid host resistant plant lines, and compared to aphids from a soybean aphid colony established in 2001 from samples collected in Illinois shortly after the soybean aphid was detected in the U.S. The latter were unable to colonize any of the plants with host

resistance, while the Ohio-derived culture showed virulence on the resistant soybean genotypes Dowling (*Rag1*), LD05-16611 (*Rag1*), and Jackson (*Rag(Jackson)).* The ability of this new soybean aphid isolate, to colonize plants with *Rag1* or *Rag(Jackson),* which likely are allelic host resistance genes (Hill et al., 2012), demonstrated that the Ohio isolate was a representative of a new, previously unknown *Ap. glycines* Biotype 2 (B2) that could overcome *Rag1*-conditioned resistance and had a different virulence spectrum compared to the original avirulent isolate collected in Illinois, now called Biotype 1 (B1) (Kim et al., 2008; Alt and Ryan-Mahmutagic, 2013), and whose genome is described herein.

A second significant biological change was observed during the fall of 2006 when soybean aphid colonies and eggs were observed on *Frangula alnus* (glossy leaved buckthorn) at three widely separate locations in Northern Indiana. For aphids the switch to a different woody plant species that serves as the overwintering primary host, is uncommon due to the specialization of the fundatrix morph on the primary host plant (Moran, 1988). In the spring of 2007, colonies of *Ap. glycines* were again observed on *F. alnus* at two locations, demonstrating that the aphid had successfully overwintered on this new host plant (O'Neil, R. and Voegtlin, D.J., Personal communication). Previous observations and laboratory tests had shown that the *Ap. glycines* gynoparae (Fig. 1) would accept *F. alnus* in the fall, feed and produce nymphs, but these would not mature into oviparae and thus not deposit overwintering eggs (Voegtlin et al., 2004). Aphids from Indiana found to have survived over winter on *F. alnus* were taken into culture and tested on a panel of aphid-resistant soybean lines to determine their virulence spectra (Hill et al., 2010). From the results of the tests, an aphid clone, established from viviparous aphids collected on *F. alnus*, behaved as a new biotype (Biotype 3; B3), which was able to colonize soybean genotypes with the *Rag2* gene (Hill et al., 2009).

These findings showed that the soybean aphid possessed potentially significant genetic variability that resulted in virulence, posing a threat to the durability of plant host resistance used to manage this pest. This knowledge prompted soybean breeders to expand their search for new host resistance sources (Hill et al., 2017) and develop genetic strategies to improve the durability of host resistance genes, such as pyramiding multiple resistance genes together within soybean cultivars (McCarville, et al., 2014; Ajayi-Oyetunde et al., 2016**)**, to retard the adaptation to host resistance and slow the erosion of resistance efficacy. Multiple *Rag* genes have been mapped in soybean and several commercial varieties with *Rag1*, *Rag2* and *Rag1+2* are commercially available (McCarville et al., 2014; Hesler et al., 2013). However, several virulent *Ap. glycines* biotypes have been documented: B2, virulent on *Rag1*; B3, virulent on *Rag2*; B4, virulent on *Rag1*, *Rag2*, and *Rag1+2* (Kim et al., 2008; Hill et al., 2010; Alt and Ryan-Mahmutagic, 2013). The facility with which the *Ap. glycines* North American population has developed virulent biotype to resistant plant varieties has prompted the question of whether aphids in North America hybridized with a resident species and whether this "hybrid vigour" contributed to its success.

Two possible candidate species that also utilize *Rhamnus* as an overwintering host are *Ap. gossypii* and *Ap. nasturtii* (Lagos, 2014). Hybridization between different species

of aphids has been documented (Mueller, 1985) as well as the hybridization producing fertile offspring in the laboratory between *Ap. grossulariae* and *Ap. triglochinis* where the morphology and host preference of the former usually dominated in the hybrid clones (Rakauskas, 2000). Hybridization has also been demonstrated between *Ap. glycines* and *Ap. gossypii*. While *Ap. gossypii* does not share soy as a summer host it does share *Rhamnus* as the overwintering host plant. In China where the two species share *R. purshiana* (Cascara buckthorn or Cascara sagrada), Zhang and Zhong (1982) observed natural crossbreeding between the cotton and soybean aphid in Jilin Province, China and conducted laboratory hybridization experiments that demonstrated that mating between the species occurred. A greater number of viable eggs occurred in the cross *Ap. glycines* female x *Ap. gossypii* males than its reciprocal and offspring of both crosses could only live on the corresponding host of the female parent.

Efforts have been made to compare the population genetic structure of the ancestral Asian and invasive U.S. populations (Michel et al., 2009; Jun et al., 2013). Using populations from Ontario, Canada, nine different U.S. midwestern states and seven microsatellites, previously designed for *Ap. fabae* and *Ap. gossypii*, found significant genetic differentiation between South Korean and North American populations. However, for the latter, genetic diversity was associated with time of collection, June to September 2008, rather than geographic location, leading to the conclusion that this observed pattern was the result of successful asexual clonal populations expanding and colonizing other localities during a growing season (Michel et al., 2009). Eighteen simple sequence repeats (SSRs) used to examine the population structure of the soybean aphids collected from two localities in the U.S., two in South Korea and one in Japan had resolution to discern differences in the aphids originating from the different countries but not between the two samples within the U.S. and South Korea (Jun et al., 2013).

Genomic resources for agricultural crops and insects that affect them are increasing. Currently there are 12 publicly available genomes of agricultural aphid pests which differ in genome size, life history patterns, geographic distribution and impact as pests: *Ap. gossypii* (Quan et al., 2019), *Myzus persicae* (Mathers et al., 2017), *M. cerasi* (AphidBase; https://bipaa.genouest.org/is/aphidbase/), *Acyrthosiphon pisum* (The International Aphid Genomics Consortium, 2010), *Diuraphis noxia* (Nicholson et al., 2015), *Melanaphis sacchari* (NCBI; PRJNA413550), *Rhopalosiphum maidis* (NCBI; PRJNA480062), *R. padi* (AphidBase; https://bipaa.genouest.org/is/aphidbase/), *Schizaphis graminum*, and *Sipha slava* (NCBI; PRJNA472250), including the genome of *Ap. glycines* obtained by sequencing specimens from laboratory colonies and field specimens from six geographic localities in the Midwest U.S. (Wenger et al., 2017) and the genome of the strain of *Ap. glycines* (B1) presented herein (Table 1). In addition to the recently-obtained genomes of the cedar aphid *Cinara cedri* (Julca et al., in press) and of the phylloxeran *Daktulosphaira vitifoliae* (Rispe et al., 2019, in press) were kindly provided prior to publication for comparative analysis.

This paper provides a high-quality genome and annotation of *Ap. glycines* B1. A laboratory culture established from specimens collected in the field in Illinois in 2001. We include an analysis of the soybean aphid B1 genome with respect to the currently

available aphid genomes mentioned above including its sister species, the cotton aphid, closely related but with widely different host ranges. *Ap. glycines* uses the soybean plant as a summer host and a few species in the genus *Rhamnus* as the overwintering host, while *Ap. gossypii* utilizes over 900 species of plants (Blackman and Eastop, 1984; Carletto et al., 2009; Wang et al., 2016). Despite its widespread distribution and highly polyphagous nature the cotton aphid has the smallest genome of the currently available aphid genome assemblies and was found to have the lowest number of private genes (Quan et al., 2019). A superficial look at genome size differences does not hold the answer to the differences in the natural history of aphids. Rather, answers are likely to lie in the manner in which gene expression is regulated. Mathers et al. (2017) showed that identical clones of the polyphagous *M. persicae* can colonize different distantly related host plants via the differential regulation of expanded gene families which collectively upregulate within days of experiencing a change in host plant.

We present a phylome report, the complete collection of phylogenetic trees of genes encoded in the soybean aphid genome and the currently available aphid genomes to elucidate the evolutionary history of this pest. In addition, because structural cuticular proteins (CPs) are the major constituents of arthropod exoskeleton and also candidates for host receptors of plant viruses we have investigated the full set of structural CPs present in this aphid species (Webster, 2018; Kamanga, 2019). In this study we describe the different CPs subfamilies detected in the *Ap. glycines* genome after extensive manual curation that led to the annotation of the full set of this group of proteins. Phylogenetic analyses were done on two specific subfamilies of CPs, the RR-1 and RR-2 proteins, that contain a central chitin-binding domain (Andersen et al., 1995; Rebers and Willis, 2001; Willis, 2010) such as the conserved 64- amino- acids R&R domain (Cornman and Willis, 2008).

Furthermore, we also include an analysis of the soybean aphid world-wide population structure and its invasion of the North American continent using single nucleotide polymorphisms (SNPs) and specimens collected from across its world geographic distribution between 2001 and 2013. We trace the genetic changes of this population during its early period of colonization of the U.S. and Canada, with the aim to determine the adaptive process and genes that underwent selection as it adapted to the North American landscape. We also examine the influence of resistant soybean cultivars on the genetic diversity of aphids that colonize them and the genes associated with this selection process (See Fig. S1 for work flow diagram).

North America presented the soybean aphid an environment with drastically different topography, resources, predators and insect population control methods than it experienced in its original Asian environment. Uncovering how the genome of this species has and continues to navigate the opportunities and challenges that present themselves will inform as to the best manner to control it and other agricultural pests.

2. Materials and Methods

2.1 Laboratory aphid rearing and field collections of samples

DNA for the sequencing of the genome of *Ap. glycines* was obtained from a laboratory culture of B1, established from specimens collected in Urbana, Illinois in 2001 and kept in the laboratory from that time onwards. *Ap. glycines* specimens were reared on individual plant leaves of *Glycines max*, variety Williams 82 (W82), placed in petri dishes (100 x 20 mm) with a moistened cotton disk. Aphids were maintained in Percival incubators at 25°C with a light regimen of 16L/8D. Aphids were collected with a paintbrush and immediately placed in a tube on dry ice. Parthenogenetic soybean aphids were collected in the field for the SNP based population analysis, preserved in 95% ethanol and stored at -20°C prior to being processed.

2.2 Extraction of DNA used for Illumina, 454 and PacBio

DNA was extracted using a phenol/chloroform method. A starting material of ~100ul of aphids was used for the extraction. 1) Aphids were ground in Drosophila homogenization buffer: DHB - 0.1 M NaCl, 0.2 M sucrose, 0.01 M EDTA (pH8) and 0.03 M Tris (pH8), the solution was sterile and stored at 4°C (Teknova) and phage lysis buffer: PLB--0.25M EDTA, 0.5M Tris (pH9.2) and 2.5% SDS, this solution was sterile and stored at room temperature (RT) (Teknova). Tubes incubated at 65°C for 30 min after which they were spun briefly at low speed and set to incubate overnight at 37°C with 5µl of 20mg/ml of Proteinase K (-20°C). 2). 30µl of 3M KAc was added to the tubes, mixed gently, and placed on ice for 30 minutes. Tubes were centrifuged in a refrigerated microfuge for 10 minutes after which the supernatant was removed. 3) An equal volume (500µl) of Tris equilibrated phenol (ChCl3:Phenol) was added and the tubes mixed by hand. Tubes were then spun for 5 minutes at room temperature. The upper aqueous phase (475µl) was removed to fresh tubes while avoiding the interphase material. 4) An equal amount of ChCl3 was added. The tubes were shaken by hand, spun for 5 minutes at RT, the aqueous phase retrieved and placed into new tubes. 5) 1µl of 32mg/ml of RNaseA (-20C) (Sigma R4642) was added to tubes, which were mixed and incubated at 37°C for 15min. 6) 100-95% ethanol, in a volume of two times the amount of supernatant, (700-800µl) was added to tubes and left overnight at -20°C. 7) Tubes were spun in refrigerated centrifuge for 30 min. The supernatant was removed while being careful not to disturb the pellet, which was washed with 1ml of ethanol and stored at -20°C. 8) Tubes were spun in refrigerated centrifuge for 5 minutes then dried in an incubator at 39°C while not allowing the DNA to get overly dry to facilitate re-suspension. 9) 20µl of TE was added to tubes to resuspend DNA at 37°C overnight. 10) DNA from separate tubes was pooled into a single tube with a concentration of ~1180 ng/µl.

2.3 Extraction of RNA, library construction and sequencing

For 454 data, total RNA was extracted from 3 groups of aphids: B1, B2 and B3 using Trizol. mRNA was isolated from 20µg of total RNA using Oligotex (Qiagen, CA). cDNA was synthesized using random hexamers with the Superscript Double-Stranded cDNA synthesis kit (Invitrogen, CA). cDNA was then nebulized to a size of 400-1000 bp and blunt-ended. 454 adaptors were obligated to both ends; adaptors with unique sequence identifiers (barcodes) were used for the different samples to enable sample

identification upon sequencing. The adaptored cDNA was amplified for 10 cycles and normalized with the Trimmer Direct kit (Evrogen, Russia). The three barcoded normalized cDNA libraries were pooled and sequenced on two 1/16th regions of a 454-Titanium plate (titration). The titration yielded 79,326 reads with an average length of 385bp.

For Illumina data, RNA was extracted with Trizol (Thermo Fisher, MA) as per the manufacturer's protocol with one modification: RNA was treated with DNAse (Qiagen, CA) before precipitation. RNA was eluted in RNAse-free water (Thermo Fisher), quantitated with Qubit (Thermo Fisher) and the integrity of the RNA rRNA bands and absence of DNA were evaluated in a 1% Ex-Gel next to a 1kb DNA ladder (Thermo Fisher).

RNAseq libraries were constructed using the TruSeq RNA Sample Preparation Kit (Illumina, CA). Briefly, messenger RNA was selected from one microgram of high quality total RNA. First-strand synthesis was synthesized with a random hexamer and SuperScript II (Thermo Fisher, MA). Doble stranded DNA was blunt-ended, 3'-end A-tailed and ligated to indexed adaptors. The adaptor-ligated double-stranded cDNA was amplified by PCR for 10 cycles. The final libraries were quantitated Qubit (Thermo Fisher) and the average size was determined on an Agilent bioanalyzer DNA7500 DNA chip (Agilent Technologies, DE) and diluted to 10nM. The individually barcoded libraries were pooled in equimolar concentration. The pooled libraries were further quantitated by qPCR on an ABI 7900.

The multiplexed libraries were loaded onto three lanes of an 8-lane flowcell for cluster formation and sequenced on an Illumina Genome Analyzer IIx. The libraries were sequenced from one end of the molecules to a total read length of 100nt. The raw .bcl files were converted into demultiplexed fastq files with the software Cassava 1.6 (Illumina, CA).

2.4 Extraction of DNA for SNP analysis

DNA was extracted using the Qiagen DNeasy Blood & Tissue kit (Cat No./ID: 69504) according to the manufacturer's instructions with some minor modifications. Using a fine sable paintbrush and with the aid of a microscope, individual aphids preserved in 95% ethanol and stored at -20°C, were placed on clean kimwipes to absorb ethanol and dry out and then transferred, with a fine sable paintbrush, to an eppendorf tube with 180ul of lysis solution and 5µl of Proteinase K.

While visualizing the aphid under the scope, the specimen was macerated against the side of the walls of the tube with a pestle (Polypropylene, Bel-Art Products, Cat # 19923001). Tubes were briefly pulse-vortexed to mix then were placed in a heat block to incubate overnight at 50°C. Tubes were spun down for 30 seconds at low speed in a small bench top spinner to bring down any condensation on the inside of the caps. Extraction was treated with the addition of 1µl of RNAase (R4642 Sigma-Aldrich) ~24 mg/ml. Tubes were briefly vortexed and incubated at room temperature (25°C) for 10

457 min. Tubes were centrifuged for 30 seconds. 200µl of buffer AL was added and tubes
458 mixed briefly by pulse-vortex. Tubes were incubated at 70°C for 5 min to dissolve
459 precipitate, vortexed briefly at low speed, and incubated for an additional 3 min or until
460 all precipitate was dissolved. Tubes were spun briefly at low speed to bring down
461 condensation on the inside of the caps, and cooled for 5 to ten minutes. 200µl of cold (-
462 20°C) ethanol (96-100%) was added and tubes vortexed briefly after which they were
463 placed at 4°C overnight to allow DNA to precipitate. Tubes were briefly centrifuged and
464 the entire lysate transferred to Promega columns (Wizard SV Minicolumns Part #
465 A129B) without wetting the rim, and centrifuged at 8000 rpm for 1 min. The flow
466 through was discarded and the column membrane washed with 500µl Buffer AW1,
467 centrifuged at 8,000 rpm for 1 min and rewashed again with 500µl Buffer AW2 and
468 centrifuged at 8,000 rpm for 1 min. A final centrifuge step at 12,000 rpm for 3 min was
469 used to dry the membrane completely. The column was then placed in a clean, labeled,
470 1.5 ml Eppendorf tube and 50µl of Sigma tissue culture water was added to the center of
471 the membrane and allowed to saturate the membrane for 3 minutes. Membrane was
472 centrifuged at 12,000 rpm for 3 min to elute the DNA. Tubes with eluted DNA were
473 incubated in a heat block at 60°C for ~1/2 hr, to insure that all residual ethanol from the
474 wash buffers evaporated which reduced the volume in tube to 30µl +/-3µl. Tubes were
475 vortexed gently and spun down briefly. DNA was measured using a Qubit Fluorometer
476 (Thermo Fisher Scientific, U.S.). As aphids used differed in size the DNA obtained with
477 the above protocol ranged from ~230 to 650ng of total DNA from a single aphid. Aphid
478 specimens resulting in a concentration of 300 to 400ng in a 7µl volume were chosen for
479 the downstream steps. DNA resulting in a concentration of 300 to 400ng (~395ng) in a 7-
480 30 µl volume, was placed in individual wells of a 96 well plate. The plates were sealed
481 and run in a SpinVac to dry without heat for 1 hr. Plates were checked to confirm if dry,
482 if not, the procedure was repeated for another 15 minutes. 7µl of water was added to
483 wells in plated, covered with film and the DNA allowed to re-suspend overnight at 4°C.
484 If the plate was not run right away it was stored a -20°C.
485
486 *2.5 Sequencing of genome*
487
488 An Illumina HiSeq 2000 and 454 Titanium system was used to generate Illumina
489 and 454 sequences (NCBI SRA accessions: PRJNA551277). Two types of libraries were
490 prepared and sequenced with 454 Titanium platform: 1) random shotgun, in which
491 genomic DNA was randomly sheared to a size of 600nt to 1.2kb and 2) paired-end, in
492 which DNA was sheared to a size of 8kb and 20kb fragments. On Illumina HiSeq 2000
493 system, the shotgun libraries, with a fragment size of 200bp, were sequenced from both
494 ends (paired-end sequencing), each read being 100nt in length. Mate-pair libraries with a
495 jump size of 3kb and 8kb were sequenced at 35nt from each end of the fragments. Using
496 Pacific Biosciences (PacBio) RSII sequencing platform with C2 chemistry, we sequenced
497 a 10K library on 8 SMRT cells which yielded a total of 193,586 sequences that passed
498 quality filters (NCBI SRA accessions: PRJNA551277). Mean length of these sequences
499 was 4,274 bases. Total number of bases in all the 193,586 sequences was 1,299,749,757.
500
501 *2.6 Genome sequence assembly*
502

We used sequencing reads from Illumina HiSeq 2500, Pacific Biosciences (PacBio) RSII and 454 FLX Titanium sequencers. Illumina sequencing data contained both paired-end reads and mate pairs with 3kb and 10kb target insert sizes. The 454 sequencing data contained mate pairs with target insert size of 8Kb. The PacBio reads were produced on the RSII sequencer with P6-C4 chemistry. MaSuRCA assembler version 3.2.2 (Zimin et al., 2017) was used to assemble sequencing reads from the three different sequencing platforms. At initial step, MaSuRCA error-corrects Illumina reads, followed by filtering of the Illumina paired reads by removing PCR duplicates and short non-junction pairs. It then transforms Illumina paired-end reads into super-reads (Zimin et al., 2013). The super-reads were assembled into mega reads using PacBio reads as templates. MaSuRCA then assembled the mega-reads along with error corrected and filtered Illumina and 454 paired reads with CABOG assembler version 8.2.

2.7 Optical BioNano Genome (BNG) map construction and assembly

Aphids were harvested from leaves, immediately frozen on dry ice and shipped to MOgene LC (St. Louis, MO) for optical map construction. High-molecular-mass DNA was extracted using the Bionano IrysPrep Animal Tissue DNA Isolation Fibrous Tissue User Guide" (Document # 30071, v.A, 2016). In brief, tissue was briefly fixed in formaldehyde to protect DNA from mechanical shearing. This was followed by homogenization using a rotor stator. Subsequently the crude homogenate of the extracted DNA was embedded in agarose plugs to undergo purification. The process yielded 300ng of high molecular weight DNA (HMW).

Using the Knickers software (v1.5.5), we determined that the best nicking enzyme for this genome was BssSI (New England BioLabs), with a labelling density of approximately 16 nicks per 100kb (http://www.bnxinstall.com/knickers/Knickers.htm). To obtain Nicked, Labeled, Repaired and Stained (NLRS) NLRS-gDNA 300 ng of g DNA was used using the protocol in the IrysPrep Labeling-NLRS User Guide (Document #30024, v.G, 2016). In brief, extracted genomic DNA was placed in a Nicking master mix and allowed to incubate for 2 hrs at 37°C. This was subsequently combined with the labeling master mix and incubated for 1hr at 72°C. A repair master mix was then added for 0.5 hrs at 37°C for the purpose of repairing the nicks. Lastly the mixture was stained and incubated overnight at 4°C. At the end of the NLRS procedure the labeled sample was quantified using the Bionano Irys System. NLRS-gDNA was loaded onto IrysChip (part # 20249, v2; SN: 850024985) and the IrysChip was scanned using the protocol given in the Irys User Guide (Document # 30047, v.B, 2016). The raw data output of 221.9 GB obtained from these scans was analyzed using IrysView software (v2.5.1) and the protocol given in "IrysView v2.5.1 Software Training Guide" (Document # 30035, v.G, 2016). The filtered data output consisted of 102.6 Gb.

Using the BioNano Genomics assembly pipeline, genomic maps of DNA molecules in bnx format were aligned against each other and assembled into BioNano Genome map contigs. There were 665 BioNano Genome Map (BNG) contigs that covered 358 Mb of the *Ap. glycines* genome. MaSuRCA was used to generate scaffolds that were further extended as well as joined with other scaffolds utilizing BNG contigs.

Using BioNano Genomics software Refaligner, sequence assembly scaffolds were aligned against BNG contigs. These alignments were processed with the BioNano Genomics pipeline and a total of 85 hybrid scaffolds that spanned 303 Mb were generated. There were 198 sequence assembly scaffolds integrated into the hybrid scaffolds and these covered approximately 280 Mbp of the genome. The utilization of BNG contigs resulted in the reduction of the number of scaffolds in the *Ap. glycines* genome assembly from 3,261 to 3,254 scaffolds. The N50 scaffold length increased from 2,957,263 bp to 5,358,903. The increase in the N50 scaffold length is due to merging the largest scaffolds of the sequence assembly using BNG contigs as the template.

2.8 Filtering of assembly scaffolds

Genome assembly scaffolds were aligned against NCBI non-redundant (NR) protein database (version from 2017-11) using BLASTX command of diamond aligner (version 0.9.10). All the Illumina and 454 reads used to assemble the genome were aligned against the assembled scaffolds using BWA-mem (version 0.7.15). These two alignments were given as input to Blobtools (version 0.9.19.6) (Laetsch et al., 2017) to identify scaffolds that belonged to proteobacteria and these were subsequently removed from the downstream analysis. The supplementary file 1 contains the parameters used to create BlobDB database using the diamond BLASTX results and parameters to create and view the blobplot.

2.9 Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis

To evaluate the relative completeness of the assembly, BUSCO (Simão et al., 2015) version 3.0.1 was run on the final version of assembled scaffolds with the insect single copy ortholog database version 9.

2.10 Assembly of transcriptome reads

To assist in the annotation of the soybean aphid genome two transcriptome assemblies were generated using 43,138,024 single end Illumina RNA Seq reads and a second using 4,403,008 454 sequences. Illumina RNA Seq reads were first preprocessed with Trimmomatic (Bolger et al., 2014) software to trim adapter bases using parameter ILLUMINACLIP and all reads shorter than 25 bases were removed using parameter MINLEN. To improve the efficiency of assembling the data, *in silico* read normalization was performed on trimmed reads using Trinity's script (Grabherr et al., 2011) with parameters --JM 500G --max_cov 30 --pairs_together and --PARALLEL_STATS. Illumina reads thus normalized were assembled using Trinity version 2.1.1 (Grabherr et al., 2011) in genome guided mode with parameters --genome_guided_bam --genome_guided_max_intron 10000 --max_memory 50G. To assemble 454 transcriptome sequences, newbler (Margulies et al., 2005) was run with all default parameters.

2.11 Alignments of RNA Seq reads against genome sequence

RNA Seq reads of previously published soybean aphid were downloaded from the NCBI short read archive database with accession numbers: SRP031835, SRP033884, SRP050997, SRP062763. Raw reads were preprocessed using Trimmomatic (Bolger et al., 2014) to trim low quality bases and adapter sequences using parameters LEADING:28 TRAILING:28 SLIDINGWINDOW:4:20 MINLEN:30 ILLUMINACLIP:2:15:10 and subsequently were aligned against the assembled scaffolds using STAR aligner (version 2.5.3a) (Dobin et al., 2013) using all default parameters. Similarly, RNA Seq reads used in creating the transcriptome assemblies were also aligned against the assembled scaffolds using STAR aligner.

2.12 Annotation of soybean aphid genome

To annotate the genome sequence of soybean aphid, MAKER annotation pipeline version 3.01.1 (Cantarel et al., 2008) was used. The first round of MAKER was run by giving as input a transcriptome assembly generated using 454 sequences, another transcriptome assembly generated using Illumina paired end reads, protein sequences from closely related species such as cotton aphid (Quan et al., 2019), *Drosophila melanogaster* (downloaded from flybase version FB2016_02), *Diuraphis noxia* (Nicholson et al., 2015), and *Myzus persicae* (clone G006 and clone O downloaded from AphidBase), all the protein sequences from swissprot database (version 2016-05) and alignments of RNA Seq reads against the genome sequence.

By running command “maker -CTL” four parameter files were created. Of all the files thus generated maker_opts.ctl file was modified to include the full path to all the above data. Full path to the genome sequence was given using the parameter “genome”, full path to transcriptome assemblies was given using the parameter “est”, full path to the RNA-Seq read alignments was given using the parameter “est_gff”, protein sequences of closely related species was given using parameter “protein”. To infer gene predictions using transcriptome assemblies and closely related species’ proteins, est2genome and protein2genome were set to 1. MAKER accepts read alignments in GFF format. To convert read alignments in BAM format to GFF format, they were first converted to bed format using bedtools bamtobed tool and then converted to BAM format using genometools bed_to_gff3 tool.

After the completion of the first round of MAKER run, fasta_merge and gff3_merge was run to generate FASTA file of protein and transcript sequences and the genome annotation in GFF3 format. Using the gene models created in the first round of MAKER, sequences for training Augustus (Stanke et al., 2006) were extracted. This is achieved by extracting the genomic regions that contain mRNA annotations along with 1000 bases up and downstream of the mRNA annotations using bedtools getfasta (Quinlan and Hall, 2010) tool. These sequences were given as input to BUSCO and BUSCO was run using parameters -m genome, -long, -sp pea_aphid -l insect_odb9. After the BUSCO run was completed, the new config files that were generated by BUSCO were renamed and copied to the species config folder of Augustus.

To train SNAP (Korf, 2004) using the best models created from the first-round MAKER, gene models with AED score of 0.25 or better and a sequence of 50 bases long were extracted using maker2zff script using parameters -x 0.25 and -l 50. Training

parameters were created by running forge command on the annotations and sequences obtained after running the maker2zff script. Hmm-assembler.pl script was run to generate HMM models. The file with HMM models was given as input to MAKER.

For the second round of MAKER in the maker_opts.ctl file, est2genome and protein2genome was set to 0. “snaphmm” was assigned the full path to the HMM file that was created subsequent to the training of SNAP as mentioned above. “augustus_species” was set to the new species folder that was created in the Augustus config folder and it contains the parameters generated by BUSCO after training Augustus. After the completion of the second round of MAKER fasta_merge and gff3_merge was run to extract genome annotation in GFF3 format and transcript sequences in FASTA format. Annotation file thus obtained was examined using jbrowse (Buels et al., 2016) to check the integrity of annotation.

To obtain the functional annotation of the *Ap. glycines* genes, protein sequences in FASTA format were aligned against UniProt database sequences and the first 20 best alignments for each query *Ap. glycines* protein sequence were extracted. Using the "Retrieve ID/mapping" (https://www.uniprot.org/uploadlists/) tool of UniProt database, we extracted protein names based on the UniProt gene IDs from the 20 best alignments. All entries with protein name “Uncharacterized protein” were excluded. From the remaining entries the protein name of the first entry is assigned to the *Ap. glycines* query protein. Using the same approach, we extracted GO annotations and protein names from the UniProt database based on the 20 best alignments for each query *Ap. glycines* protein sequence (Table S1 and S2). In addition, we ran the AphidBase pipeline to align gene sequences against the NCBI non-redundant protein database followed by uploading of the BLAST results in XML format to BLAST2GO program (Conesa et al., 2005). Subsequently the BLAST2GO program assigned GO terms to each gene by querying the GO database using the protein id from the BLAST results. GO annotations obtained from UniProt and NCBI were consolidated and from these a final file was generated (Table S1).

2.13 Retrieval of the full set of cuticular proteins in Ap. glycines genome

To retrieve the full set of genes coding for CPs (including CPs with the R&R motif defined as CPR proteins; Rebers and Riddiford, 1988) in the *Ap. glycines* B1 genome, CutProtFam annotation site (http://aias.biol.uoa.gr/CutProtFam-Pred/) was used, with standard settings (Ioannidou et al., 2014). Annotated genes were then fully curated on AphidBase through web-Apollo.

2.14 Aphis glycines phylome reconstruction

The *Ap. glycines* phylome was reconstructed using the PhylomeDB pipeline (Huerta-Cepas et al., 2011). In brief, for each protein-coding gene in the soybean aphid genome we searched for homologs (Smith-Waterman Blast search, e-value cutoff < 1e-05, minimum contiguous overlap over the query sequence cutoff 50%) in a protein database containing the proteomes of the 16 species considered (Table S3). The most

similar 150 homologues were aligned using three different programs (MUSCLE (Edgar, 2004), MAFFT (Katoh et al., 2005) and KALIGN (Lassmann and Sonnhammer, 2005) in forward and reverse direction. These six alignments were combined using M-COFFEE (Wallace et al., 2006), and trimmed with trimAl v.1.3 (Capella-Gutiérrez et al., 2009) using a consistency cut-off of 0.16667 and a gap threshold of 0.1). Phylogenetic trees were built using Maximum Likelihood approach as implemented in PhyML v3.0 (Guindon and Gascuel, 2003) using the best fitting model among seven different ones (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff). The two models best fitting the data were determined based on likelihoods of an initial Neighbor Joining tree topology and using the AIC criterion. We used four rate categories and inferred fraction of invariant positions and rate parameters from the data. All alignments and trees are available for browsing or download at PhylomeDB with the PhylomeID 709 (Huerta-Cepas et al., 2014).

2.15 Alignment and phylogenetic reconstruction of cuticular proteins RR-1 and RR-2 sub-groups

Phylogenetic analyses were performed using the corresponding protein sequences sets of updated RR-1 or RR-2 genes retrieved from five aphid genomes: *Ap. glycines* B1, *M. persicae* (Mathers et al., 2017), *A. pisum* (Gallot et al., 2010), *D. noxia* (Nicholson et al., 2015), *R. padi* and the close-related aphid species, *Daktulosphaira vitifoliae*. RR-1 and RR-2 sub-groups were treated separately. After removal of predicted signal peptides using SignalP-5.0 Server (Almagro Armenteros et al., 2019), RR-1 mature protein sequences were used in phylogenetic analyses. For RR-2 proteins, only the extended 69 amino acids RR domain (pfam00379) was used for phylogenetic analyses, because they tend to be highly divergent and difficult to align along their full length. RR-2 proteins from *Ap. glycines*, *M. persicae*, *A. pisum*, *D. noxia*, *R. padi* and *D. vitifolia*, were aligned using Clustal Omega (Sievers et al., 2011) and the aligned extended domain of each RR-2 protein was extracted for further phylogenetic analyses.

Phylogenetic analyses of the RR-1 and RR-2 proteins were then assessed using the Seaview software (Gouy et al., 2009). To generate alignments, MUSCLE software (Edgar, 2004), a part of the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) sequence analyses tool kit, was used (Madeira et al., 2019). Ambiguous regions after alignment (i.e. containing gaps and / or poorly aligned) were removed with Gblocks (v0.91b) using the following parameters: minimum length of a block after gap cleaning: 10, no gap positions were allowed in the final alignment and all segments with contiguous non conserved positions bigger than 8 were rejected, minimum number of sequences for a flanking position: 85%.

Phylogenetic trees were reconstructed using the maximum likelihood method implemented in the PhyML program (v3.1/3.0 aLRT, and SeaView v 4.6.2). The WAG amino-acid substitution model was selected, assuming an estimated proportion of invariant sites, and 4-categories gamma-distributed rate to account for rate heterogeneity across sites. The gamma shape parameter was estimated directly from the data (gamma=3.517) and reliability for internal branch was assessed using the aLRT test (SH-

Like).

2.16 Prediction of gene duplications, and orthology and paralogy relationships

Orthology and paralogy relationships were predicted based on phylogenetic evidence from the soybean aphid phylome. We used ETE v3 (Huerta-Cepas et al., 2010a) to infer duplication and speciation relationships using a species overlap approach. The relative age of detected duplications was estimated using a phylostratigraphic approach that uses the information on which species diverged prior and after the duplication node (Huerta-Cepas and Gabaldón, 2011). Duplication frequencies at each node in the species tree were calculated by dividing the number of duplications mapped to a given node in the species tree by all the gene trees that contain that node. For this analysis we excluded gene trees that contained large species-specific expansions (expansions that contained more than five members). All orthology and paralogy relationships are available through PhylomeDB (Huerta-Cepas et al., 2014).

2.17 Gene ontology term enrichment for phylome analysis

Gene Ontology (GO) terms enrichment analysis was performed using FatiGO (Al-Shahrour et al., 2007). We compared two lists of proteins (*Ap. glycines* specific duplications and duplications at the ancestral node of all aphids) against all the other proteins encoded in the genome.

2.18 Species tree reconstruction

The trimmed alignments of 67 larger genes (>10 Kb) that had single orthologs in the 16 species considered were selected and concatenated. The final alignment containing 109,282 amino acid positions was used to reconstruct the maximum likelihood species tree with RAxML v8.1.17 (Stamatakis, 2014) using the LG amino acid substitution model, and 100 bootstrap replicates.

2.19 SNP Discovery and genotyping using Illumina Golden Gate Assay

RNA-Seq reads from *Ap. glycines* B1, B2 and B3 reared on susceptible plants (Dowling) were trimmed using the FASTX toolkit (Gordon and Hannon, 2010). Bases with quality score less than 20 were trimmed from 3' end and reads that were less than 50 nucleotides in length were discarded. A total of 10,089,179 reads from B1, 8,081,931 reads from B2 and 12,458,830 reads from B3 were used for *in silico* SNP discovery. Reads from each individual biotype were aligned against the preliminary set of contigs assembled using Illumina and 454 sequences by running tophat v1.3.1 (Trapnell et al., 2009) with parameters –solexa1.3-quals and –g 1. Only the single best alignments were used for the downstream SNP discovery pipeline. For query reads with more than one best alignment, tophat chose at random only one of the best alignments. Alignment output files in BAM format were sorted using samtools (Li et al., 2009) based on

alignment coordinates on the contigs. Sorted BAM files were processed using samtools mpileup and bcftools with default parameters to identify potential SNPs. The maximum coverage used to allow the detection of a SNP/indel was 100, this was achieved by setting parameter varFilter to -D100. SNPs identified using reads from each individual biotype were combined into a single VCF file. There was a total of 45,071 SNPs identified using reads from all three biotypes. Of all the SNPs, 30,509 SNPs had one hundred bases flanking on either side of each SNP on the assembled contigs. This set of SNPs was sent to Illumina to generate genotype designability scores.

A GoldenGate Universal-32, which contained 3072 plex Assay Kit with UDG and custom designed Soybean Aphid Custom Oligo Assay Pools was generated by Illumina (San Diego, CA). Briefly the manufacturing steps included the following: the assay design tool was used to identify 50 base upstream or down-stream of the identified SNP and associated flanking regions to determine which strand would function best as a probe. Probes were synthesized to the flanking region of interest and these included a universal forward or reverse primer, with the latter containing the locus specific region, the Illumicode Sequence tag and the Universal reverse sequence primer. DNA oligos complementary to the allele specific sequence are synthesized and attached to a bead. These are pooled and applied to a bead chip where multiples of each bead type localize in each of the 32 sample areas on the chip. The Illumina manufacturing QC uses a decode process that sequences each unique Illumina code sequence tag to check its location (X, Y coordinate on the chip) and that each bead type is represented (Gunderson et al., 2004). The SNP specific bead chip as well as the SNP specific primer pool is the product of this process. Probes are then pooled and stored at -20°C and used in the golden Gate genotyping assay. The custom GoldenGate chip outlined above was used to process 250ng, according to the manufacturer's instruction, for each of all samples used in the population analysis. Slides were scanned using an Illumina iScan beadscanner and image processing and QC analysis was carried out using GenomeStudio software.

A total of 3,072 SNPs with best designability scores were selected for genotyping a total of 4,421 samples collected from Australia, Canada, China, Indonesia, Japan, Myanmar, South Korea, Taiwan, Thailand and USA. Using Illumina genome studio, genotype clusters for all 3,072 SNPs were manually examined and edited. Of 4,421 samples, 212 were excluded because the call rate was less than 95% and 418 were excluded because they were lab culture samples. Of 3,072 SNP clusters, 637 SNP genotype clusters were manually flagged as being poor quality and removed from the analysis. Of the remaining 2,435 SNPs, 55 had no genotypes in more than 100 samples and were subsequently discarded from the downstream analysis. This resulted in the final set of 2,380 SNP genotypes in 3,791 samples (Table 2) that was used in the downstream analysis.

2.20 Annotation of genes overlapping SNPs

There were 1,700 genes that overlapped with 2,380 SNPs. Of the genes found to overlap, GO terms were obtained for 1,185 genes, Eukaryotic Orthologous Groups (KOG) categories were obtained for 1,025 genes, Kyoto Encyclopedia of Genes and

Genomes (KEGG) pathway names were identified for 641 genes. To obtain KOG categories, RPS BLAST of gene sequences was run against the KOG database with -max_target_seqs 1 -evalue 1e-10 as parameters. GO annotations were downloaded from AphidBase. In turn, to obtain KEGG K numbers for each gene, protein sequences in FASTA format were submitted to the KEGG's GhostKOALA server (https://www.kegg.jp/ghostkoala/).

Databases used for gene annotation, while having data on multiple organisms, vertebrate and invertebrates, have the greatest amount of information for model organisms that have been well studied. If we restrict our analysis to insects it would not be possible to identify pathway information for many genes in our study. Moreover, much of the existing insect annotation is derived from the well-studied model species such as human, rat, mouse. Hence, some of the genes and pathway names listed have human specific nomenclature.

2.21 Assessment and management of ascertainment bias.

Our SNP discovery process is based on the alignments of sequence reads from U.S. samples against the reference genome of U.S. *Ap. glycines*. There is an ascertainment bias 1) when SNPs ascertained in one population are used to genotype other populations 2) when SNPs ascertained using a small set of samples are used to genotype larger set of samples of the same population (Nielsen et al., 2004; Lachance and Tishkoff 2013). As a result of ascertainment bias, very few SNPs with allele frequencies close to 0 or 1 are found in the populations used for SNP discovery while SNPs with these frequencies are more frequent in the populations not used for the SNP ascertainment (Albrechtsen et al., 2010). We detected this pattern in the allele frequency spectrum generated for U.S./Canada and Asia/Australia populations (Fig. S2).

The allele frequencies for the U.S./Canada population show a bell-shaped distribution, with values ranging from 0.3 to 0.7, while those of the Asian/Australian population combined have a bimodal curve with frequencies ranging from 0 to 1 (Fig. S2). The difference in the allele frequency distribution is a reflection of the manner in which SNPs were identified. Namely, highly polymorphic loci determined from sequencing reads of U.S. samples were chosen as SNP candidates. With this approach, and by not having sequence reads from the Asian/Australian population, our assay resulted in containing a high number of SNPs with frequencies closer to 0 and 1 in the Asian/Australian population.

Unless one obtains whole genome sequence of every individual in the population, it is not possible to remove SNP ascertainment bias completely. It has been proposed that sequencing data from samples of all populations being compared can help to address this problem, however, this is also prone to bias as not every individual in the population would be considered (Lachance and Tishkoff 2013). As a means to compensate for the ascertainment bias, when comparing U.S./Canada and Asian/Australian populations, we resolved to restrict our analysis to the use of 926 SNPs that fall within the allele frequency range of 0.3 and 0.7 in the combined Asia/Australia population, as these are

present in all populations being evaluated (Fig. S2). While we run the risk of eliminating informative SNPs in the Asian/Australian populations, this more conservative approach limits the use of SNPs that are fixed in these populations. This selection did not eliminate all SNPs with frequencies of 0 and 1, rather it chose SNPs for each population, with allele frequencies that formed a bell-shaped distribution, as can be seen in Fig. S2. We analyzed and compared U.S./Canada and Asia/Australia populations using both the original complete 2,380 SNPs as well as the reduced 926 SNPs obtained via the method outlined above.

2.22 Principle component analysis

Principle component analysis (PCA) was conducted using JMP version 13. A VCF file with SNP genotype data was converted into a tab delimited file with genotypes coded as "0" for the homozygous reference allele, "1" for the heterozygote and "2" for the homozygous alternate allele. After importing the tab delimited text file into JMP, missing genotypes were imputed using "Multivariate Normal Imputation" function in JMP. "Principle components" function under "Multi variate methods" was used to run the principle component analysis on the imputed genotypes. The graph builder function of JMP was used to generate a PCA plot with the first two principle components.

2.23 Identification of clonal copies

To identify clonal copies among samples, principle components were obtained for all samples. The first three principle component values for each sample were rounded to non-decimal values. All samples with the same principle component values were grouped into clusters of clones.

2.24 Calculation of F_{ST} values

VCF tools version 0.1.15 (Danecek et al., 2011) was used to calculate F_{ST} values according to the method described in Weir and Cockerham 1984. VCF file with 3,791 samples and 2,380 SNPs was given as input to the VCFtools using --weir-fst-pop option for each population in the pairwise comparison. F_{ST} values were calculated for all pairwise comparisons between all populations sampled: Australia, China, Japan, South Korea, Indonesia, Taiwan, Thailand, Myanmar, Canada and U.S. F_{ST} values were also calculated using the same set of SNPs to compare U.S. samples collected in 2001 and those sampled in 2005, 2006, 2008, 2009, 2010, 2011, 2012. In addition, F_{ST} values were calculated in comparisons between aphids from susceptible soybean plants and *Rag* varieties: *Rag1*, *Rag2* and *Rag1+2*.

2.25 Manhattan plots

Tab delimited files with F_{ST} values for all markers in pairwise comparisons were imported into JMP version 13. The graph builder function of JMP was used to generate Manhattan plots by assigning SNP chromosome coordinates to the x-axis and F_{ST} value to the y-axis.

915
916 *2.26 Heat maps*
917
918 Comma delimited files with F_{ST} values were imported into R using the read_csv
919 function. Heat maps were generated on the imported F_{ST} values using pheatmap function
920 of R package pheatmap (Kolde, 2015).
921
922 *2.27 Over representation analysis*
923
924 Over representation analysis was performed for multiple sets of genes that overlap
925 with SNPs with F_{ST} values 1) >0.14 in a comparison between U.S. samples collected in
926 2001 and 2005 2) >0.1 in a comparison between U.S. samples collected in 2001 and
927 2009, 2010, 2011, and 2012 3) >0.2 in comparison between *Rag* (*Rag1*; *Rag1+2*; *Rag2*)
928 and susceptible aphid samples. To identify the GO terms or KEGG pathways
929 overrepresented among these two sets of genes, hypergeometric analysis was performed
930 using the GOStats package (Falcon and Gentleman 2007). The genes that overlapped
931 with the 2,380 SNPs used in this study were considered as "universe". The read.table
932 function was used to import input files into R. For the GO terms over representation
933 analysis, GOALLFrame and GeneSetCollection data objects were created using
934 GOAllFrame and GeneSetCollection functions of GSEABase package (Morgan et al.,
935 2019). The GSEAGOHyperGParams and hyperGTest functions were used to perform
936 hypergeometric test on GO terms, while GSEAKEGGHyperGParams and hyperGTest
937 functions were used to perform hypergeometric test on KEGG pathway terms.
938
939 *2.28 Identification of non-synonymous SNPs*
940
941 To identify the non-synonymous SNPs among the 2,380 SNPs, Ensembl Variant
942 Effect Predictor (McLaren et al., 2016) was run on an input file with 2,380 SNPs in VCF
943 format using paramenter "-i" along with the gene annotation file in GFF format with
944 parameter "-gff" and the genome sequence in FASTA format using parameter "-fasta".
945
946 *2.29 Data availability*
947
948 The genome sequence assembly scaffolds, gene annotation and functional
949 annotation files are available at AphidBase (https://bipaa.genouest.org/is/aphidbase/). The
950 genome sequence assembly and gene annotation was also deposited at NCBI GenBank
951 under the accession VYZN01000000; GenBank assembly accession GCA_009761285.1;
952 BioProject PRJNA551277; BioSample SAMN12143004. The raw sequence data was
953 deposited at NCBI SRA database under accession PRJNA551277. The SNP genotype
954 data was deposited at the European Variation Archive under project PRJEB35243 and
955 analyses ERZ1108186 (https://www.ebi.ac.uk/
956 ena/data/view/PRJEB35243).
957

958 **3. Results and discussion**

3.1. Genome assembly and evaluation

Of the currently available aphid genome sequence assemblies the soybean aphid is amongst one of the three smallest. The assembly of *Ap. glycines* B1 has an estimated size of 308 Mbp, 3,224 scaffolds and an N50 value of 6 Mbp making it next best assembly after *R. maidis* (Table 1). The smallest aphid assembly is *Ap. glycines* sister species *Ap. gossypii* followed by *M. sacchari*. The most recently sequenced genomes, obtained with technologies that produce longer reads and the use of new mapping tools, have the smallest number of scaffolds: *R. maidis*, and *M. sacchari* followed by the *Ap. glycines* B1 assembly included herein. Of all the single copy orthologs tested by BUSCO, 92.2% were identified to full length in the assembly and 88.9% were found as single copy. Only 1.2% of BUSCOs were fragmented and 6.6% were missing.

Aphids listed in Table 1 differ in their life histories and plant host range. Some are specialist and use a limited number of host plants, such as *Ap. glycines*, whose host plant range was mentioned in the introduction. *M. cerasi* utilizes several species in the genus *Prunus* and a limited number of secondary hosts in the families Asteraceae, Brassicaceae Rubiaceae and Scrophulariaceae. Most of the aphids listed, *D. noxia*, *M. sacchari*, *R. maidis*, *R. padi*, *S. flava*, *S. graminum* and *M. sacchari* have a middle level plant host range and utilize various number and species of grasses (Kindler and Springer 1989; Mezey and Szalay-Marzsó, 2001; Blackman and Eastop 1984). The remaining species range from the polyphagous species of *M. persicae* and *A. pisum* to the highly polyphagous *Ap. gossypii*. This latter species, unlike other members of the *Aphis frangulae* group, can overwinter on several other plant genera besides Rhamnaceae. However, the full range of the cotton aphid's capacity to exploit different species of plants and their respective chemistries is best seen in the number of summer host that it can utilize that span over 92 species of plant families (van Emden and Harrington, 2007; Blackman and Eastopm, 1984). The current limited sample size of complete genome assemblies, from various and mostly distantly related aphid genera, does not permit a ready examination of the possible links between genome size and life history.

3.2 Phylome analysis

To elucidate the evolutionary history of *Ap. glycines*, we reconstructed the phylome in the context of sixteen other insect genomes (Table S3). This phylome was analyzed to infer duplication and speciation events, and derive paralogy and orthology relationships (Gabaldón, 2008). The soybean aphid phylome, including the alignments, phylogenetic trees and orthology and paralogy relationships, is available for browsing and downloading in PhylomeDB (phylomeID: 709, http://www.phylomedb.org) (Huerta-Cepas et al., 2014).

The phylome of *Ap. glycines* includes 14,914 gene trees, which cover 76.7% of the proteome. Genes with less than two homologs do not have sufficient information to generate a tree and therefore were not included when gene trees were generated. A total of 13,845 proteins (71.2%) have an ortholog in at least one of the other species that were analyzed.

When considering orthologs present in all sixteen species, we determined that on average 1,848 are present in each species. Of these only 811 have single-copy orthologs present in all species (Fig. 2, Table S4). When Hemipteran species were considered separately, we found an average of 288 orthologs and of these 130 were single-copy. Whereas for aphid species, we found 141 orthologs of which 81 were single-copy.

We reconstructed the evolutionary relationships of all 16 species included in the analysis by using the alignment of 67 single-copy orthologs longer than 10 Kb. The resulting species tree (Fig. 2) was congruent with previous analyses (Nováková et al., 2013).

An analysis of *Ap. glycines* gene duplications, including large gene family expansions, showed that there is a total of 3,972 soybean aphid proteins (20.4% of the proteome) that have paralogs. These genes considered as in-paralogs can be assigned to 1,028 specific gene expansions (Table S5). Most expansions (785, 76%) have small to moderate number of copies (2-5), and a few (133, 13%), have larger expansions corresponding to >10 copies (Fig. S3). As previously reported for other aphid genomes, *Ap. glycines* also has a number of genes that have very large expansions of up to 483 in-paralogs (The International Aphid Genomics Consortium, 2010; Mathers et al., 2017; Huerta-Cepas et al., 2010b).

A functional GO term enrichment analysis of *Ap. glycines* in-paralogs shows enrichment in large part for terms involved in apoptosis such as negative regulation of apoptotic process, homophilic cell adhesion via plasma membrane adhesion molecules, inhibition of cysteine-type endopeptidase activity involved in apoptotic process, negative regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis, JAK-STAT cascade, spermatid nucleus differentiation, protein monoubiquitination, protein desumoylation, and protein neddylation (Table S6). Similar enriched functions were found in other aphids-specific duplications (Mathers et al., 2017; Duncan et al., 2016; Huerta-Cepas et al., 2010b).

The proteins involved in the above listed functions affect processes of cell cycle, proliferation, contact inhibition and cell adhesion and death. Ubiquitination is a crucial process involved in apoptosis, autophagy, and the cell cycle. In humans, disturbance of these processes can lead to disease states such as cancer. While these processes are involved in cell death, they can function as protective mechanisms during exposure to stress and protect cells from apoptosis. Duplications of apoptotic related genes may facilitate *Ap. glycines*'s colonization of host plants with differing chemistry as well as permit a successful response to pesticide exposure.

We examined other aphid species in our analysis to determine whether they had gene duplications in parallel as those that occur in *Ap. glycines*. There are 1,621 (41%) *Ap. glycines* genes that are involved in 1,028 gene expansion events, of these 372 occur in at least one other aphid species (Table S5). Unexpectedly, *Ap. gossypii*, the most closely related species, in this comparison has the lowest number of parallel duplication events (Fig. S4). A functional analysis of the proteins of *Ap. glycines* that have parallel

duplications in other aphids examined in this study, indicate that most GO enrichment terms are related to apoptotic processes such as SUMO-protease specific activity, NEDD8 activity, apoptotic process, spermatid nucleus differentiation, sensory organ development, negative regulation of Wnt signalling pathway, regulation of JAK-STAT cascade, negative regulation of compound eye retinal cell death, antennal morphogenesis, defense response to Gram-negative bacterium (Table S6).

To identify genes under selection in *Ap. glycines* and its most closely related, species *Ap. gossypii*, we calculated the dN/dS ratios of 7,502 single-copy orthologs of *Ap. glycines* and *Ap. gossypii* using *M. persicae G006* as the outgroup. Of these orthologs, 3,825 passed the cut off filters (see Materials and Methods). Most of the genes (~98%) of both soybean and cotton aphid have dN/dS ratios lower than 1, suggesting the action of purifying selection, while the remaining fraction of genes (~2%) show dN/dS ratios higher than 1, indicative of positive selection (Table S7, Fig. 3).

Of the 3,825 single copy orthologs, six proteins were identified as under positive selection in both *Ap. glycines* and *Ap. gossypii* species, and only one, Groucho had known functional information. Groucho proteins are DNA-binding repressors that inhibit transcription by interacting with a repression domain (Paroush et al., 1994; Fisher et al., 1996; Aronson et al 1997; Dubnicoff et al., 1997; Jimenez et al., 1997)

There are 47 genes identified as under positive selection in *Ap. glycines*. Functional information is available for 31 of these genes. These encompass a range of metabolic functions from P450s involved in detoxification, to arrestin domain-containing protein that transports proteins between cells, to histone acetyltransferase that acetylates lysine on histone proteins (Table S8).

Of 42 genes determined to be under positive selection in *Ap. gossypii*, 24 have known functional annotations. Genes under this category also cover a wide variety of metabolic functions from Azurocidin, an anti-microbial protein, to optomotor-blind protein required for optic lobes and wing development, to the sodium channel protein Nach, involved in the clearance of tracheal liquid.

3.3 Cuticular proteins

The manual curation and annotation of *Ap. glycines* cuticular protein (CP) genes allowed the identification of 106 unique genes belonging to seven well-identified cuticular protein subfamilies present in Orthopteran insects (Willis, 2010) (Table S9). Similar representatives numbers in each CPs subfamilies are found in aphid genomes and in *D. vitifoliae*, the grape wine pest species belonging to Phylloxeroidea, a Superfamily considered to be the nearest sister taxon of the Aphidoidea. Of the genomes examined thus far, only *A. pisum* shows a major expansion of the RR-2 protein (Table S9). Such an increase of gene content in *A. pisum* has been discussed and appears to be a characteristic of this aphid species (Mathers et al., 2017). The authors explained this feature by an increase in lineage-specific genes and widespread duplication of genes from conserved families (Mathers et al., 2017). More specifically, in *Ap. glycines* the final CPs set

includes 13 and 71 unique genes harboring respectively the RR-1 and RR-2 motif (Table S9). As mentioned in the introduction section these subfamilies (named CPRs) are of major importance in insect physiology. They are by far the largest CPs subfamilies in every species of arthropod sequenced so far and appear to be restricted to this group of invertebrates (Willis, 2005). The R&R Consensus domain present in CPRs confer chitin-binding properties to these proteins and is involved in cuticle formation (Rebers and Riddiford, 1988). It seems that RR-1 proteins are preferentially present in soft (flexible) cuticle while RR-2 proteins are found in hard (rigid) cuticles (Willis, 2010). Interestingly these proteins are poor in cysteine residues. Andersen (2005) suggested that cystine could react with ortho-quinones and interfere with sclerotization of the cuticle.
Most RR-1 proteins from *Ap. glycines* seem to display 1-to-1 orthology relationships with other aphid species and this reduced complexity signals the absence of specific duplication trends for this protein subfamily (Fig. S5A). An ortholog of Stylin 01, originally identified in *A. pisum* and *M. persicae* was also found in *Ap. glycines* (AG6029153) (grey box, Fig. S5A). This RR-1 protein present at the tip of aphid stylets is believed to be a receptor of non-circulative viruses (i.e. viruses transmitted during short punctures without internalization of the viral particles) such as the *Cauliflower mosaic virus* (CaMV), or the CMV which is transmitted by *Ap. glycines* (Uzest, 2007; Webster 2018; Gildow et al., 2008). Indeed Stylin 01, named previously Mpcp4 in *M. persicae* (Dombrovsky, 2007), was shown to interact in yeast with the coat protein of the CMV. However, there is still no direct evidence of its role in CMV transmission (Liang and Gao, 2017).

Most CPR proteins harbor signal sequences, consistent with their extracellular/secretory localization, and most CPR genes display the canonical first intron in this signal peptide. Noteworthy, CPR gene subfamilies are located on different genome scaffolds (data not shown) showing a differentiated localization depending of the CPR nature (RR-1 or RR-2) as it was previously shown for *M. persicae* (Mathers et al., 2017). Moreover, some scaffolds harbor several RR-2 genes organized as tandem repeats. Within these tandem arrays some genes occur in pairs of almost identical adjacent sequences and were reported in other organisms such as *Aedes aegypti* (Cornman and Willis 2008). The presence of tandem repeats might reflect duplications events as suggested by phylogenetic analyses (Fig. S5B).

RR-2 proteins are also good candidates as plant virus receptors. CMV has been reported to interact with several RR-2 peptides detected in aphid stylets (Webster et al., 2017). However, it was not possible to precisely identify one specific candidate. Recently, Kamangar and colleagues (2019) reported the role of MPCP2, a RR-2 protein of *M. persicae*, in the transmission of PVY, another non-circulative virus. *Ap. glycines* ortholog (AG6024500) of MPCP2 (referred as Mp_000169000 in Fig. S5B) belongs to a well conserved cluster among different aphid species and *D. vitifoliae* (grey box, Fig. S5B). Since *Ap. glycines* transmit PVY (Davis et al., 2005) it would be useful to investigate the role of this RR2-protein in PVY transmission.

3.4 Origin and distribution of Ap. glycines populations

The 2,380 SNPs Illumina Golden Gate assay developed for this study was based on sequence data from *Ap. glycines* samples obtained in North America. When this assay is used to genotype populations not included in the SNP discovery process an ascertainment bias can result (Nielsen et al., 2005; Nielsen 2005; McTavish and Hills 2015). We chose to adjust for this bias when analyzing the world populations of *Ap. glycines* listed in Table 2, by using a subset of 926 SNPs (see Materials and Methods for specific details).

Using this set of 926 SNPs we conducted a PCA analysis using genotypes from individual soybean aphid specimens collected from 10 countries across *Ap. glycines's* world-wide distribution. Data from 2001 to 2013 (Table 2; Fig. 4). shows that the U.S./Canada and Asian/Australian populations are clustered in separate groups with U.S. samples collected in 2001 overlapping with Asian samples (Fig. 4 a, b). In U.S. the soybean aphid was first detected in 2000. Samples from 2001 are the closest approximation to the aphids that were introduced in North America. Their similarity to Asian samples is supported by the overlap seen in this analysis further confirming that *Ap. glycines* that invaded North America originated from Asia. Samples in the North American cluster display a more diffuse distribution than those in the Asian and Australian cluster.

While samples from each Asian country form their own cluster, there is considerable overlap between countries (Fig. 4). Samples from China overlap with South Korea, Taiwan, Indonesia, Thailand, and Myanmar but not Japan (Fig. 4 c, d) suggesting that the soybean aphid has dispersed from China to these countries. Populations of *Ap. glycines* from Japan do overlap with those from South Korea. This distribution is likely the result of the higher interactions that have taken place historically between South Korea and Japan. Due to the overlap between Indonesian and Australian samples it is likely that the former is the likely source of this relatively recent invasive population (Fig. 4 c and d).

The results and interpretations derived from the PCA analysis are in concordance with those derived from the pairwise F_{ST} values calculated for all countries (Fig. 4 e). The lowest F_{ST} values were observed between the U.S. and Canada and these form a cluster in the PCA plot (Fig. 4; a, b, e). Pairwise comparisons of the two North American populations against the Asian countries show that the lowest value is *vis a vis* South Korea, followed by China and Japan, indicating that the likely source of the North American population of *Ap. glycines* is South Korea and/or China. The highest F_{ST} value between the North American population and Asian countries is Myanmar. The population of *Ap. glycines* in Myanmar may be an isolated population that differentiated subsequent to its dispersal from China or conversely a local ancestral Asian population of *Ap. glycines.*

When Asian countries are compared to each other, China has the lowest F_{ST} value. This also supports that China was the source and point of dispersal of the current population of *Ap. glycines* to all other Asian countries. The lowest F_{ST} is seen between China and South Korea and the highest between China and Myanmar. The genotypic

composition of the current Asian population is likely a consequence of the recent human facilitated dispersal of *Ap. glycines* from China. However, when considering all the sampled populations the highest F_{ST} values are those observed between Myanmar and Australia followed by those between Australia and Thailand (Fig. 4 e). The highest F_{ST} value across all populations is between North America and Australia, likely because the latter, derived from Indonesia is a differentiated population, and like the U.S. population the result of a recent bottleneck. This relationship, and all the other pairwise F_{ST} comparisons are also illustrated in the Neighbor Joining tree (Fig. 4 f).

The same analysis was conducted with the full set of 2,380 SNPs (Fig. S6). The same relationship between populations from different countries were seen using Fst values even though the PCA plot reflects ascertainment bias in that the US/Canada and Asia/Australia form two separate distinct clusters (Fig. S6 a-f).

A comparison of PCA plots using the complete 2,380 (Fig. S7 A) and the reduced 926 (Fig. S7 B) SNP data sets, for U.S./Canada and Asian/Australian samples collected in different years: 2001; 2008; 2010-2013, for the U.S./Canada and Asia/Australia clusters, show separation of populations in the A series and their closeness in the B series.

The yearly analysis in Fig S6 B also shows that the 2001 U.S. samples overlap with Chinese samples from Hei Long Jiang and Jilin provinces, two of the major soybean growing areas of China, and not samples from Japan. From the available samples tested, the results indicate that the first introduction of *Ap. glycines* to the North American continent in 2001was likely from China. For subsequent years a direct overlap between U.S. and Asian samples is only seen in 2011 where U.S. aphids overlap with South Korean samples from the provinces of Cheonan and Suwon and Japanese samples from Tochigi prefecture. These results could be interpreted as a possible second introduction to the U.S. in 2011 from these localities or an overlap resulting from the high diversity of genotypes being generated in the U.S. invasive population as it adapted to the North American landscape.

3.5 Change in the U.S./Canada Ap. glycines population over time.

As the U.S./Canada population was the source for the SNP discovery process, the complete 2,380 SNP data set was utilized for subsequent analyses that pertained to this population. PCA plots generated using the total number of 2,380 SNPs for samples from the U.S. and Canada from 2001 to 2013 but divided in three time periods: 2001-2005; 2006-2009; and 2010-2013 show that the samples in the time period 2010-2013 are less diffused than the previous two periods, indicative of a decrease in genetic diversity with time (Fig. 5; A, B, C, D). These results lead to the conclusion that the U.S./Canada *Ap. glycines* population underwent directional selection as it adapted to the North American continent. These results are reflected in the F_{ST} values obtained when comparing the same time periods (Fig. 5). In contrast, PCA plots for Chinese and Japanese *Ap. glycines* populations for the time period from 2001 and 2011 do not show a decrease in diversity

over the same time periods (Fig. S8; A, B) and thus do not show the same directional pattern observed in the U.S./Canada population.

As indicated in previous work (Michel et al., 2009), our results indicated that overall, time was a better predictor of genetic differences in the U.S./Canada *Ap. glycines* population than geographic provenance. PCA analysis of samples from years that included collections from more than two states indicated no apparent structure to the *Ap. glycines* North American population with respect to geographic locality (Fig. S9). While apterous aphids move very short distances, historically it has been thought that aphid flight is common (Close and Tomlison, 1975; Llewellyn et al., 2003, Irwin et al., 2007; Shufran et al., 2009) and that most flights are migratory (Johnson, 1954). Recently it has been proposed that migration is a rarer event and that aphids tend to move shorter distances, with migration being an exception (Loxdale et al., 1993, 1999; Ward et al., 1998). Our data shows that there is overlap between all the states sampled. This could be interpreted that the aphids are involved in long range movement across the Midwest or that the degree of diversity generated in the *Ap. glycines* population within a state is greater than that between states and aphids may not be moving long distances. The *Ap. glycines* in the North American landscape can reach astronomically high population numbers, especially at the end of the summer when such population explosions can become airborne and a component of the "aerial plankton". The environmental parameters involved in the prediction of a given aphid species propensity to migrate short or long distance are highly complex it is likely that there is a continuum of migratory behavior that is species and environment dependent (Irwin et al., 2007; Parry, 2013).

We visualized the distribution of the 2,380 SNPs and their respective F_{ST} values across the genomic scaffolds for the years 2005 and 2009-2012. The Manhattan plots generated (Fig. S10) show that the SNPs with the highest F_{ST} values, and the corresponding genes that these overlap with, are concentrated in the first (1-5) and the last (14-79) scaffolds of the *Ap. glycines* B1 genome. The intervening scaffolds of 6-13 had SNPs with lower F_{ST} values. SNPs trailing behind those with high F_{ST} values are in close proximity on the scaffolds and are hitchhiked by the lead SNP. If the genes that overlap with high F_{ST} value SNPs are under positive selection then the hitchhiked genes could increase in frequency due to linkage with the selected genes as it has been proposed by the draft model (Nielsen 2005; Gillespie 2000, 2001).

The corresponding heat map for these samples (Fig. 6) shows that the F_{ST} values for most SNPs change through time. With the exception of the samples from 2005, those from other years show few SNPs at the highest F_{ST} values and these occur for usually one year and repeat for a maximum of three.

The higher the F_{ST} value the greater the difference in allele frequency of a SNP between the samples tested. A sample with a high number of clonal individuals would result in higher allele frequencies for the SNPs that they possessed which in turn increase its F_{ST} values. Most of the samples from the aphids collected at two localities in 2005 are clonal copies. The year 2005 when compared to the 2001 baseline has SNPs with

significantly higher F_{ST} values than the other years. *Ap. glycines* reproduces clonally in the summer months and all samples tested were apterous parthenogenetic individuals collected in the field. If a particular clone is successful it will have greater representation in a given sample. We examined the number of unique and clonal copies for each collection year (Fig. S11). For the year 2005 we had access to 41 individual samples from two localities, WI and IL, of these 32 were clonal and 9 unique. All the clonal individuals originated from the IL locality and represent a successful clonal lineage at this time and place.

We examined the GO terms (Table S10) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Table S11) for genes overlapping with SNPs having F_{ST} values greater than or equal to 0.14 for the comparison between 2001 and 2005 population. We visualized genes with high F_{ST} value SNPs that were assigned to enriched GO terms in comparison between samples collected in 2001 and 2005 to see their respective F_{ST} values in samples collected in subsequent years.

The GO ID's for genes overlapping with SNPs having high F_{ST} values for the 2005 year comparison (Table S10) such as programmed and regulation of cell death, regulation of apoptotic process, response to toxic substance, stress response to metal ion are indicative of exposure to stress. As indicated in the introduction, 2006 was the year when *Ap. glycines* were observed to colonize a new species of overwintering plant, *Frangula alnus*, and also when the first aphids were observed surviving on *Rag1* resistant cultivars in the field in Ohio. Furthermore, small experimental plots of *Rag* resistant cultivars had been planted in several localities in the Midwest such as IL and IA in the previous year. The stress response genes with high F_{ST} values may be indicative of the response of successful clones as they adapted to the new challenges of the North American landscape.

SNPs that had high F_{ST} value (>0.2) in 2005 fluctuated in subsequent years. With the exception of AG6029093 (Fig. S12), corresponding to the gene signal peptidase complex catalytic subunit SEC11 (EC 3.4.21.89) (Table S2), which contains a SNP with F_{ST} values of 0.23 and 0.19 for the years 2006 and 2009 respectively, all the other genes had F_{ST} values that were below 0.06.

We also examined the GO terms (Table S10) and KEGG pathways (Table S11) for genes containing high F_{ST} value SNPs for 2009, 2010, 2011 and 2012.

The GO terms repeated across the years (Table 3), that were associated with the category Biological Processes, correspond to cell signaling pathways localized in the plasma membrane and the myosin complex, as well as the molecular functions of hydrolase, phosphodiesterase activity, ribonucleotide and carbohydrate derivative binding. The GO term in the Biological Processes category of cellular response to chemical stimulus (2009, 2011 and 2012), 3',5'-cyclic-nucleotide phosphodiesterase activity (2009, 2010 and 2011) and myosin complex (2010, 2011 and 2012), were repeated for three years, with the latter two in consecutive years.

3.6 Response to Rag resistant varieties.

As part of the goal to examine the change in the structure of the *Ap. glycines* U.S./Canada population since the time of first colonization, and because the field deployment of *Rag* resistant varieties has been one of the significant environmental factors that has challenged the *Ap. glycines* population in North America, we conducted an analysis using aphids collected from *Rag* experimental plots from the states of Wisconsin, Minnesota, Iowa, North Dakota, South Dakota and Ohio.

Manhattan plots of *Ap. glycines* samples collected from *Rag* experiment plots and compared to samples collected on susceptible plants, for the years 2010 (WI) and 2013 (MN and IA) show overall higher F_{ST} values (Fig. S13) than the non-*Rag* plots *Ap. glycines* samples collected in the years 2003 to 2010 with the exception of 2005 (Fig. S10). In addition, SNPs with high F_{ST} values from the *Rag* experiment plots are not restricted to the first and latter numbered scaffolds of the *Ap. glycines* B1 genome assembly, as they were for samples collected on non-*Rag* field plants, but rather more uniformly distributed along the entire number of scaffolds. This is especially relevant for samples collected from the *Rag1* and *Rag1+2* soybean varieties. Previous laboratory tests have shown that these two resistant varieties present more challenging environments for the *Ap. glycines* to colonize and thrive on than *Rag2* (Ajayi-Oyetunde et al 2016; Hill et al 2017).

The distribution of SNPs and their respective F_{ST} values for all the localities from which *Rag* experimental samples were collected are shown in a heat map (Fig. 7). SNPs with the highest F_{ST} values are found in IA, WI and MN. In comparison, the remaining states, ND, SD, and OH, have few SNPs with similarly high F_{ST} values. An evaluation of the number of clonal and unique aphids from each sampling locality shows that aphid samples from IA, WI and MN, with SNPs with high F_{ST} values, have a higher number of clonal than unique individuals compared to those observed for ND, SD and OH (Fig. 8). We hypothesize that aphids collected in IA, WI and MN (Group 1) had the capacity to colonize the resistant soybean plants and reproduce clonally in higher numbers, while aphids collected in ND, SD, and OH (Group 2) colonized the resistant plants but were unable to reproduce clonally to the same degree, hence a greater number of unique individuals are detected at these latter locations.

The differences in the number of clonal individuals observed on resistant varieties between locations in Group 1 and Group 2 is reflected in the higher F_{ST} values seen for Group 1. These differences are likely the result of the former location proximity to areas with high density of *R. cathartica*, the over wintering primary host of *Ap. glycines* (Fig. 9). This is likely to influence the genetic makeup of summer *Ap. glycines* populations that colonize soybeans in multiple ways. One way is that there is a higher probability of *Rag* resistant aphid clones selected in one summer season to overwinter in near by *R. cathartica* stands and recolonize resistant soybean varieties planted the following year.

We determined the GO terms (Table S12) and KEGG pathways (Table S13) for genes overlapping with SNPs having F_{ST} values greater than or equal to 0.1 for the

comparison between *Rag* experimental and susceptible soybean varieties for both Group 1 and 2 localities. The highest number of genes were assigned to the Biological Processes and Molecular Function categories. These genes encompass a wide range of functions that include nervous system development, carbohydrate metabolism and mitochondrial function. We chose to focus on GO terms that were repeated for more than one year, location or treatment (Table 4).

All GO terms occur twice with the exception of oxidoreductase activity which occurs three times on IA *Rag1*, and *Rag1+2* as well as MN *Rag1+2*. Most of the GO terms listed in Table 4 are critical components of pathways involved in iron homeostasis and crucial to the function of fundamental processes such as respiration and nitrogen fixation (Rouault and Klausner, 1997; Nichol et al., 2002). Iron is commonly used by all organisms from bacteria to plants due to its abundance in the environment, versatility and reactivity, however, because of this flexibility it is necessary that it is tightly regulated. A balance needs to be maintained between levels sufficient for metabolic processes and avoidance of iron toxicity (Rouault and Klausner, 1997).

The GO terms listed in Table 4 such as iron-sulfur cluster binding (GO:0051536) and 4 iron, 4 cluster (GO: 0051539), common from bacteria to humans, indicate metallo co-factors that are part of proteins involved in electron transport, enzymatic catalysis and regulation and also have important roles in cellular and mitochondrial iron balance. Mitochondrial aconitase (GO:0003994; aconitate hydratase activity) contains a 4Fe-4S cluster, and one iron atom of this cluster facilitates the dehydration-hydration reaction that converts citrate to isocitrate as part of the citric acid cycle, a crucial metabolic process (Rouault and Tong, 2005).

Repeating GO terms were observed in *Rag1* and *Rag1+2* varieties, the harshest environments of the three varieties tested. We hypothesize that GO terms associated with iron related pathways are enriched as a result of a perturbation of these processes in the aphids by *Rag1* and *Rag1+2* mechanisms of plant resistance.

4. Conclusion

This study is comprised of a high-quality draft genome sequence assembly and gene annotation of *Ap. glycines* B1, a culture established shortly after the introduction of this species to North America. As such it represents the closest approximation to the invasive genotype. The companion papers in this special issue have benefited from the *Ap. glycines* B1 genome sequence assembly and gene annotation. Among other findings, the analysis of this genome has shown that the duplicated portion of *Ap. glycines* proteome is mostly comprised of genes related to apoptosis, indicative of possible adaptations to plant chemical defenses. These duplicated genes, in turn may serve as pre-adaptations that facilitate aphids' ability to surmount anthropogenic stressors such as pesticides and resistant plant varieties. The duplicated genes appear critical, as one-third are duplicated in parallel in other aphid species. The sequence of this genome has brought to the fore that a comparative genomic approach to the study of aphid pest species is crucial. This is evident in the difference in the level of genes duplicated in *Ap. glycines*, that have less

than three percent in parallel duplication in *Ap. gossypii*, suggestive of different strategies to overcome environmental stressors. The world-wide population analysis suggests that the place of origin of the North American invasive population of *Ap. glycines* is likely to be China or South Korea. Genetic variation of North American soybean aphids has decreased through time and appears not correlated with geography, implying a high degree of dispersal capacity for this species. The genomic resources provided in this study will facilitate future research in the identification of specific genes, pathways and mechanisms involved in the adaptation of the soybean aphid and other pests to the North American agricultural landscape, leading to sustainable and non-polluting measures for their control.

Acknowledgments

We thank Rosa Alfaro for growing soybean plants. Alvaro G. Hernandez, Chris L. Wright and the staff at the DNA Services Lab, Roy J. Carver Biotechnology Center, University of Illinois at Urbana Champaign, for their excellent sequencing support. Clark W. Bailey, Daniel Guyot, T. Kikuchi, Masafumi Kobayashi, Helen Thompson Robert C. Bellm and Scott Berolo for assistance in obtaining aphid specimens. Adam Morris from SAS for statistical support. Rebekah D. Wallace, Center of Invasive Species and Ecosystem Health, University of Georgia for help with *R. cathartica* map.

This work was supported by generous grants from the U.S. Mid-West farmers through the checkoff program funds from the United Soybean Board (USB), Illinois Soybean Association (ISA), and the North Central Soybean Research Program (NCSRP).

Appendix

*Soybean aphid research community:

Tatsiana Akraiko[1], Andrew Aschwanden[2], Arian Avalos[3], Mark Band[4], Bryony Bonning[5], Julie Breault[6], Hugh Brier[7], Olga Chiesa[8], Anitha Chirumamilla[9], Brad S. Coates[10], Giuseppe Cocuzza[11], Eileen Cullen[12], Peter Desborough[13], Brian Diers[14], Christina DiFonzo[15], Dana Gagnier[16], John Gavloski[17], Mary Gebhardt[18], Ronald B. Hammond[19], George Heimpel[20], Ames Herbert[21], Theresa Herman[22], David Hogg[23], Yongping Huanga[24], Doug Johnson[25], Janet Knodel[26], Chiun-Cheng Ko[27], Christian H. Krupke[28], Genevieve Labrie[29], Doris Lagos-Kutz[30], Brian Lang[31], Joon-Ho Lee[32], Seunghwan Lee[33], Mauro Mandrioli[34], Gian Carlo Manicardi[35], Eric L. Maw[36], Emanuele Mazzoni[37], Michael McCarville[38], Giulia Melchiori[39], Andy Michel[40], Ana Micijevic[41], Nick Miller[42], Robin Mittenthal[43]†, Tamotsu Murai[44], Andy Nasruddin[45], Brian A. Nault[46], Matthew E. O'Neal[47], Michela Panini[48], Massimo Pessino[49], Deirdre Prischmann-Voldseth[50], G. Quesnel[51], David W. Ragsdale[52], Hugh H. Robertson[53], Tiana Schuster[54], Liu Sijun[55], Hojun Song[56], James F. Stimmel[57], Shigeru Takahashi[58], Kelley Tilmon[59], John Tooker[60], Sarah Wilson[61], Kongming Wu[62], Shuai Zhan[63], Ying Zhang[64]

[1]Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign, IL, USA

[2]Pennsylvania State University, University Park, PA, USA

[3]USDA, Agriculture Research Services, Baton Rouge, LA, USA

[4]Roy J. Carver Biotechnology Center, University of Illinois, Urbana, IL; Institute of Evolution, University of Haifa, Israel

[5]Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA

[6]MAPAQ - Ministère de l'Agriculture, des Pêcheries et de l'Alimentation, Quebec, Canada

[7]Department of Agriculture and Fisheries, Kingaroy, Australia

[8]Dipartimento di Scienze delle Produzioni Vegetali Sostenibili, Universitá Cattolica del Sacro Cuore, Piacenza, Italy

[9]Department of Entomology, North Dakota State University, Fargo, ND, USA

[10]Department of Entomology, Iowa State University, Ames, IA, USA

[11]Università degli Studi di Catania, Dipartimento di Agricoltura, Alimentazione e Ambiente, Catania, Italia

[12]Department of Entomology, University of Wisconsin, Madison, WI, USA

[13]NSW Department of Primary Industries, Orange, New South Wales, Australia

[14]University of Illinois, College of Agricultural, Consumer and Environmental Sciences, Urbana, IL, USA

[15]Department Entomology, Michigan State University, East Lansing, MI, USA

[16]Department of Agriculture and Agri-Food Canada, Harrow, Ontario, Canada

[17]Manitoba Agriculture, Food and Rural Development, Carman, Canada

[18]Department of Entomology, North Dakota State University, Fargo, ND, USA

[19]Department of Entomology, Ohio State University, Wooster, OH, USA

[20]Department of Entomology, University of Minnesota, St. Paul, MN, USA

[21]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA

[22]Department of Crop Sciences, University of Illinois, Urbana, IL. USA

[23]Department of Entomology, University of Wisconsin-Madison, Madison, WI, USA

[24]Key Laboratory of Insect Developmental and Evolutionary Biology, CAS Center for Excellence in Molecular Plant Science, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

[25]Department of Entomology, University of Kentucky, Princeton, KY, USA

[26]Department of Plant Pathology, North Dakota State University, Fargo, ND, USA

[27]Department of Entomology, National Taiwan University, Taipei, Taiwan

[28]Department Entomology, Purdue University, West Lafayette, IN, USA

[29]Centre de Recherche sur les Grains Inc. (CÉROM), Québec, Canada

[30]USDA-ARS, Urbana, IL, USA

[31]Extension and Outreach, Iowa State University, Decorah, IA, USA

[32]College of Agriculture and Life Sciences, Seoul National University, Seoul, Rep. Of Korea

[33]Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Rep. of Korea

[34]Dipartimento di Scienze della Vita, Universitá di Modena e Reggio Emilia, Modena, Italy

[35]Dipartimento di Scienze della Vita, Universitá di Modena e Reggio Emilia, Via Campi 213/D, Modena, Italy

[36]Agriculture and Agri-Food Canada, Ottawa Research and Development Centre and Canadian National Collection of Insects, Ottawa, Ontario, Canada

[37]Dipartimento di Scienze delle Produzioni Vegetali Sostenibili, Universitá Cattolica del Sacro Cuore, Piacenza, Italy

[38]Department of Entomology, Ohio State University, Wooster, OH, USA

[39]Dipartimento di Scienze della Vita, Universitá di Modena e Reggio Emilia, Modena, Italy

[40]Department of Entomology, Ohio State University, Wooster, OH, USA

[41]Department of Agronomy, Horticulture & Plant Science, South Dakota State University, Brookings, SD, USA

[42]Illinois Institute of Technology, Chicago, IL, USA

[43]Department of Nutritional Sciences at the College of Agricultural and Life Sciences (CALS), University of Wisconsin-Madison, Madison, WI, USA. † Passed away in 2017

[44]Department of Bioproductive Science, Utsunomiya University, Tochigi, Japan

[45]Agroteknologi, Universitas Hasanuddin, Makassar, Indonesia

[46]Department of Entomology, Cornell Entomology, Ithaca, NY, USA

[47]Department of Entomology, Iowa State University, Ames, IA, USA

[48]Dipartimento di Scienze delle Produzioni Vegetali Sostenibili, Universitá Cattolica del Sacro Cuore, Piacenza, Italy

[49]Department of Entomology, University of Illinois, Urbana, IL, USA

[50]Department of Entomology, North Dakota State University, Fargo, ND, USA

[51]Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA), Kemptville, Ontario, Canada

[52]Department of Entomology, Texas A&M University, Galveston, TX, USA

[53]Department of Entomology, University of Illinois, Urbana, IL, USA

[54]South Dakota State University, Brookings, SD, USA

[55]Department of Entomology, Iowa State University, Ames, IA, USA

[56]Department of Entomology, Texas A&M University, College Station, TX, USA

[57]Bureau of Plant Industry, PA Department of Agriculture, PA, USA

[58]Faculty of Agriculture, Utsunomiya University, Utsunomiya, Japan

[59]Department Entomology, Ohio State University, Columbus, OH, USA

[60]College of Ag. Sciences, Penn State, University Park, PA, USA

[61]North Dakota State University, Department of Entomology, Fargo, ND, USA

[62]Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China.

[63]Key Laboratory of Insect Developmental and Evolutionary Biology, CAS Center for Excellence in Molecular Plant Science, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

[64]Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China

References

Andersen, S.O., 2005. Cuticular sclerotization and tanning. Comprehensive Mol Insect Sci 4, 145-170.

Albrechtsen, A., Nielsen, F.C., Nielsen, R., 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27(11), 2534-2547.

Ajayi-Oyetunde, O.O., Diers, B.W., Lagos-Kutz, D.M, Hill, C.B., Hartman, G.L., Reuter-Carlson, U., Bradley, C.A., 2016. Differential reactions of soybean isolines with combinations of aphid resistance genes *Rag1*, *Rag2*, and *Rag3* to four soybean aphid biotypes. J Econ Entomol 109, 1431–1437.

Alleman, R.J., Grau, C.R., Hogg, D.B., 2002. Soybean aphid host range and virus transmission efficiency, in: Cooperative Extension, University of Wisconsin-Extension; College of Agricultural and Life Sciences, University of Wisconsin—Madison (Eds.), Proceedings of the Wisconsin Fertilizer, Aglime, and Pest Management Conference, Wisconsin, USA, Vol. 41-42. https://soilsextension.triforce.cals.wisc.edu/wp-content/uploads/sites/68/2016/07/Alleman-Conf-2002.pdf (accessed 29 April 2019)

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, E. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37, 420-423.

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D., Dopazo, J., 2007. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res 35 (2), w91-w96.

1639 Alt, J., Ryan-Mahmutagic, M., 2013. Soybean aphid biotype 4 identified. Crop Sci 53,
1640 1491–1495.
1641
1642 Andersen, S.O., Hojrup, P., Roepstorff, P., 1995. Insect cuticular proteins. Insect
1643 Biochem. Mol. Biol. 25, 153-176.
1644
1645 Aronson, B.D., Fisher, A.L., Blechman, K., Caudy, M., Gergen, J.P., 1997. Groucho-
1646 dependent and -independent repression activities of Runt domain proteins. Mol Cell
1647 Biol 17, 5581–5587
1648
1649 Blackman, R.L., Eastop, V.F., 1984. Aphids on the World's Crops, an Identification and
1650 Information Guide. John Wiley and Sons, Chichester, New York, Brisbane, Toronto,
1651 Singapore.
1652
1653 Blackman, R.L, Eastop, V.F., 2000. Aphids on the World's Crops: an identification and
1654 information Guide. The Natural History Museum. John Wiley and Sons, Ltd., New York.
1655
1656
1657 Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for
1658 Illumina Sequence Data. Bioinformatics 30(15), 2114-2210.
1659
1660 Buels, R., Yao, E., Diesh, C.M., Hayes,R.D., Munoz-Torres, M., Helt, G., Goodstein,
1661 D.M., Elsik ,C.G., Lewis, S.E., Stein, L., Holmes, I.H., 2016. JBrowse: a dynamic web
1662 platform for genome visualization and analysis. Genome Biol 17, 66–90.
1663
1664 Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C, Sanchez
1665 Alvarado, A., Yandell, M., 2008. MAKER: An easy-to-use annotation pipeline designed
1666 for emerging model organism genomes. Genome Res 18(1), 188-196.
1667
1668 Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T., 2009. trimAl: a tool for
1669 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25
1670 (15), 1972-1973.
1671
1672 Carletto, J., Lombaert, E., Chavigny, P., Brevault, T., Lapchin, L., Vanlerberghe-Masutti,
1673 F., 2009. Ecological specialization of the aphid *Aphis gossypii* Glover on cultivated host
1674 plants. Mol Ecol 18, 2198–2212.
1675
1676 Chacón, J., Landis, D., Heimpel, G. 2008. Potential for biotic interference of a classical
1677 biological control agent of the soybean aphid. Biol Control 46, 216-225.
1678
1679 Clark, A.J., Perry, K.L., 2002. Transmissibility of field isolates of soybean viruses by
1680 *Aphis glycines*. Plant Dis 86, 1219-1222.
1681
1682 Close, R.C., Tomlinson, A.I., 1975. Dispersal of the grain aphid *Macrosiphum*
1683 *miscanthi* from Australia to New Zealand. N Z Entomol. 6, 62–65.
1684

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18), 3674–3676.

Cornman, R.S., Willis, J.H., 2008. Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes. Insect Biochem Mol Biol 38, 661e676.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T, Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27 (15), 2156–2158.

Davis, J.A., Radcliffe, E.B., Ragsdale, D.W., 2005. Soybean aphid, *Aphis glycines* Matsumura, a new vector of Potato virus Y in potato. Am J *Potato* Res 82 (3),197-201.

Davis, J.A., Radcliffe, E.B., 2008. The importance of an invasive aphid species in vectoring a persistently transmitted potato virus: *Aphis glycines* is a vector of potato leafroll virus. Plant Dis 92, 1515-1523.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Dombrovsky, A., Sobolev, I., Chejanovsky, N, Raccah, B., 2007. Characterisation of RR-1 and RR-2 cuticular proteins from *Myzus persicae*. Comparative Biochemistry and Physiology Part B: Biochemistry and Evolution 146, 256-264.

Domier, L.L., Latorre, I.J., Steinlage, T.A., McCoppin, N., Hartman, G.L., 2003. Variability and transmission by *Aphis glycines* of North American and Asian Soybean mosaic virus isolates. Arch Virol 148, 1925-1941.

Dubnicoff, T., Valentine, S.A., Chen, G., Shi, T., Lengyel, J.A., Paroush, Z., Courey, A.J., 1997. Conversion of dorsal from an activator to a repressor by the global corepressor groucho. Genes Dev 11, 2952-2957.

Duncan, R.P., Feng, H., Nguyen, D.M., Wilson, A.C.C., 2016. Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. Genome Biol Evol 8 (3), 753-764.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32 (5), 1792–1797.

Falcon, S., Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23 (2), 257–258.

1731 Fisher, A.L., Ohsako, S., Caudy, M., 1996. The WRPW motif of the hairy-related basic
1732 helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and
1733 protein-protein interaction domain. Mol Cell Biol 16, 2670-2677.
1734
1735 Fletcher, M. J., Desborough, P., 2000. The soybean aphid, *Aphis glycines*, present in
1736 Australia https://www.dpi.nsw.gov.au/biosecurity/plant/insect-pests-and-plant-
1737 diseases/soybean-aphid (accessed 24 May 2019)
1738
1739 Gabaldón, T., 2008. Comparative genomics-based prediction of protein function.
1740 Methods Mol Biol 439, 387–401.
1741
1742 Gallot, A., Rispe, C., Leterme, N., Gauthier, J.P., Jaubert-Possamai, S., Tagu, D. 2010.
1743 Cuticular proteins and seasonal photoperiodism in aphids. Insect Biochem. Mol Biol 40,
1744 235-240.
1745
1746 Gillespie, J.H., 2000. Genetic drift in an infinite population. The pseudohitchhiking
1747 model. Genetics 155, 909-919.

1748 Gillespie, J.H., 2001. Is the population size of a species relevant to its
1749 evolution? Evolution 55, 2161–2169.

1750 Gordon, A., Hannon, G.J., 2010. FASTX-Toolkit, FASTQ/A short-reads pre-processing
1751 tools. http://hannonlab.cshl.edu/fastx_toolkit/ (accessed 24 May 2019)
1752
1753 Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large
1754 phylogenies by maximum likelihood. Syst Biol 52 (5), 696–704.
1755
1756 Gildow, F.E., Shah, D.A, Sackett, W.M., Butzler, T., Nault, B.A.m, Fleisher, S.J., 2008.
1757 Transmission efficiency of Cucumber mosaic virus by aphids associated with virus
1758 epidemics in snap bean. Phytopathology 98 (11): 1233-1241.
1759
1760 Gouy, M., Guindon, S., Gascuel, O., 2009. SeaView version 4: A multiplatform graphical
1761 user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol
1762 27(2), 221-224.
1763
1764 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I.,
1765 Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N.,
1766 Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K.,
1767 Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-seq data
1768 without a reference genome. Nat. Biotechnol 29 (7), 644-52.
1769
1770 Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large
1771 phylogenies by maximum likelihood. Syst Biol 52 (5), 696–704.
1772
1773 Gunderson, K.L., Kruglyak, S., Graige, M.S., Garcia, F., Kermani, B.G., Zhao, C., Che,
1774 D., Dickinson, T., Wickam, E., Bierle, J., Doucet, D., Milewski M., Yang, R., Siegmund,

C., Haas, J., Zhou, L., Oliphant, A., Fan, J-B., Barnard, S., Chee, M.S. 2004. Decoding randomly ordered DNA arrays. Genome Res14, 870-877.

Hanson, A.A., Menger-Anderson, J., Silverstein, C., Potter, B.D., MacRae, I.V., Hodgson, E.W., Koch, R.L,. 2017. Evidence for soybean aphid (Hemiptera: Aphididae) resistance to pyrethroid insecticides in the upper midwestern United States. J Econ Entomol 110, 2235–2246.

Hartman, G.L., Domier, L.L., Wax, L.M., Helm, C.G., Onstad, D.W., Shaw, J.T., Solter, L.F., Voegtlin, D.J., D'Arcy, C.J., Gray, M.E., Steffy, K.L., Orwick, P.L., 2001. Occurrence and distribution of *Aphis glycines* on soybean in Illinois in 2000 and its potential control. Online. Plant Health Progr.

Heimpel, G.E., Ragsdale, D.W., Venette, R., Hopper, K.R., O'Neil, R.J., Rutledge, C.E., Wu, Z.S., 2004. Prospects for importation biological control of the soybean aphid: Anticipating potential costs and benefits. Ann Entomol Soc Am 97, 249-258.

Hesler, L.S., Chiozza, M.V., O'Neal, M.E., MacIntosh, G.C., Tilmon, K.J., Chandrasena, D.I., Tinsley, N.A., Cianzio, S.R., Costamagna, A.C., Cullen, E.M., DiFonzo, C.D., Potter, B.D., Ragsdale, D.W., Steffey, K., Koehler, K.J., 2013. Performance and prospects of *Rag* genes for management of soybean aphid. Entomol Exp Appl 147, 201–216.

Hill, J.H., Alleman, R., Hogg, D.B., Grau, C.R., 2001. First report of transmission of *Soybean mosaico virus* and *Alfalfa mosaico virus* by *Aphis glycines* in the New world. Plant Dis 85, 561.

Hill, C.B., Li, Y., Hartman, G.L., 2004a. Resistance to the soybean aphid in soybean germplasm. Crop Sci 44, 98–106.

Hill, C.B., Y. Li, and G.L. Hartman. G.L., 2004b. Resistance of *Glycine* species and various cultivated legumes to the soybean aphid (Homoptera : Aphididae). J Econ Entomol 97, 1071-1077.

Hill, C.B., Li, Y., Hartman, G.L., 2006a. A single dominant gene for resistance to the soybean aphid in the soybean cultivar Dowling. Crop Sci 46, 1601-1605.

Hill CB, Li, Y., Hartman, G.L., 2006b. Soybean aphid resistance in soybean Jackson is controlled by a single dominant gene. Crop Sci 46, 1606-1608.

Hill, C.B., Kim, K-S., Crull, L., Diers, B.W., Hartman, G.L., 2009. Inheritance of resistance to the soybean aphid in soybean PI 200538. Crop Sci 49, 1193-1200.

Hill, C.B., Crull, L., Herman, T.K., Voegtlin, D.J., Hartman, G.L., 2010. A new soybean aphid (Hemiptera: Aphididae) biotype identified. J Econ Entomol 103, 509-515.

Hill, C.B., Chirumamilla, A., Hartman, G.L., 2012. Resistance and virulence in the soybean-*Aphis glycines* interaction. Euphytica 186, 635–646.

Hill, C.B., Shiao, D., Fox, C.M., Hartman, G.L., 2017. Characterization and genetics of multiple soybean aphid biotype resistance in five soybean plant introductions. Theor Appl Genet 130, 1335-1348.

Hodgson, E.W., McCornack, B.P., Tilmon, K., Knodel, J.J., 2012. Management recommendations for soybean aphid (Hemiptera: Aphididae) in the United States. J Integ Pest Mngmt 3 (1), E1-E10.

Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010a. ETE: a python Environment for Tree Exploration. BMC bioinformatics 11 (24).

Huerta-Cepas, J., Marcet-Houben, M., Pignatelli, M., Moya, A., Gabaldón, T., 2010b. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. Insect Mol Biol 19, 13–21.

Huerta-Cepas, J., Gabaldón, T., 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27 (1), 38–45.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., Gabaldon, T.,Gabaldón, T., 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res 39, D556-D560.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Marcet-Houben, M., Gabaldon, T., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., Gabaldón, T., 2014. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res 42, D897-D902.

Hunt, D., Footit, R., Gagnier, D., Baute, T., 2003. First Canadian records of Aphis glycines (Hemiptera: Aphididae). Can Entomol 135, 879–881.

Ioannidou, Z.S., Theodoropoulou, M.C., Papandreou, N.C., Willis, J.H., Hamodrakas S.J., 2014. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile Hidden Markov Models. Insect Biochem Mol Biol 52, 51-59.

Irwin, M.E., Kampmeier, G., Weisser, W., 2007. Aphid movement: process and consequences, in: van Emden, H., Harrington, R. (Eds.), Aphids as Crop Pests. CABI, Wallingford, UK.

Iwaki, M., Roechan, M., Hibino, H., Tochihara, H., Tantera, D.M., 1980. A persistent aphid borne virus of soybean, Indonesian Soybean dwarf virus transmitted by *Aphis glycines*. Plant Dis 64, 1027–1030.

Jimenez, G., Paroush, Z., Ish-Horowicz, D., 1997. Groucho acts as a corepressor for a subset of negative regulators, including hairy and engrailed. Genes Dev 11, 3072-3082.

Johnson, C.G., 1954. Aphid migration in relation to weather. Biol Rev 29, 87–118.

Julca, I., Marcet-Houben, M., Cruz, F., Vargas-Chavez, C., Johnston, J.S., Gómez-Garrido, J., Frias, L., Corvelo, A., Loska, D., Cámara, F., Gut, M., Alyotto, T., Latorre, A., Gabaldón, T. (in press). Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of aphidomorpha.

Jun, T-H., Michel, A.P., Wenger, J.A., Kang, S-T., Rouf Mian, M.A., 2013. Population genetic structure and genetic diversity of soybean aphid collections from the USA, South Korea, and Japan. Genome 56, 345-350.

Kamangar, S.B., Christiaens, O., Taning, C.N.T., De Jonghe, K., Smagghe, G., 2019. The cuticle protein MPCP2 is involved in Potato virus Y transmission in the green peach aphid *Myzus persicae*. J Plant Dis Protect 126 (4), 351–357

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33 (2), 511–518.

Kim, H., Hoelmer, K.A., Lee, W., Kwon Y-D, Lee, S., 2010. Molecular and Morphological identification of the soybean aphid and other Aphis species on the primary host *Rhamnus davurica* in Asia. Ann Entomol Soc Am 103(4), 532-543.

Kim, K.-S., Hill, C.B., Hartman, G.L., Rouf Mianc, M.A., Diersa, B.W., 2008. Discovery of soybean aphid biotypes. Crop Sci 48, 923-928.

Kindler, S.D., Springer, T.L., 1989. Alternate hosts of Russian wheat aphid (Homoptera: Aphididae). J Econ Entomol 82, 1358-62.

Koch, R.L., Potter, B.D., Glogoza, P.A., Hodgson, E.W., Krupke, C.H., Tooker, J.F., DiFonzo, C.D., Michel, A.P., Tilmon, K.J., Prochaska, T.J. et al. 2016. Biology and economics of recommendations for insecticide-based management of soybean aphid. Plant Health Prog 17, 265–269.

Korf, I., 2004. Gene finding in novel Genomes. BMC Bioinformatics 5, 59.

Krupke, C.H., Obermeyer, J.L., Bledsoe, L.W., 2005. Soybean aphid, E-217-W Purdue Extension, Purdue University, Indiana, USA.
https://extension.entm.purdue.edu/publications/E-217.pdf (accessed 30 April 2019)

1911 Lachance, J., Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic
1912 analyses: Why it is important, and how to correct it. BioEssays 35, 780-786.
1913
1914 Laetsch, D.R., Blaxter, M.L., 2017. BlobTools: Interrogation of genome assemblies
1915 [version 1; referees: 2 approved with reservations]. F1000Research 6, 1287.
1916 https://doi.org/10.12688/f1000research.12232.1
1917
1918 Lagos, D.M., Voegtlin, D.J., Coeur d'acier, A., Giordano, R., 2014. *Aphis* (Hemiptera:
1919 Aphididae) species groups found in the Midwestern United States and their contribution
1920 to the phylogenetic knowledge of the genus. Insect Sci 21(3), 374-91.
1921
1922 Lassmann, T., Sonnhammer, E.L.L., 2005. Kalign-an accurate and fast multiple sequence
1923 alignment algorithm. BMC bioinformatics 6, 298.
1924
1925 Li, Y., C.B. Hill, S. Carlson, B.W. Diers, and G.L. Hartman., 2007. Soybean aphid
1926 resistance genes in the soybean cultivars Dowling and Jackson map to linkage group M.
1927 Mol Breed 19, 25-34.
1928
1929 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
1930 Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The
1931 Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078-9.
1932
1933
1934 Liang, Y., Gao X-W. 2017. The cuticle protein gene MPCP4 of *Myzus persicae*
1935 (Homoptera: Aphididae) plays a critical role in Cucumber mosaic virus acquisition. J
1936 Econ Entomol 110, 848-853.
1937
1938 Llewellyn, K.S., Loxdale, H.D., Harrington, R., Brookes, C.P., Clark, S.J., Sunnucks, P.,
1939 2003. Migration and genetic structure of the grain aphid (*Sitobion avenae*) in Britain
1940 related to climate and clonal fluctuation as revealed using microsatellites. Mol Ecol
1941 12(1), 21-34.
1942
1943 Loxdale, H.D., Hardie, J., Halbert, S., Foottit, R., Kidd, N.A.C., Carter, C.I., 1993. The
1944 relative importance of short- and long-range movement of flying aphids. Biol Rev 68,
1945 291–311.
1946
1947 Loxdale, H.D., Lushai, G., 1999. Slaves of the environment: the movement of
1948 herbivorous insects in relation to their ecology and genotype. Philos Trans R Soc Lond B
1949 Biol Sci 354, 1479–1498.
1950
1951 Ma, Y.H., 1984. Development of soybean genetic and breeding research in China, in: S.
1952 Wong (Ed.), Proceedings of the 2nd U.S.-China soybean symposium, 28 July-2 August
1953 1984, Changchun, Jilin, China, pp. 15-19.
1954
1955 Macedo, T. B., Bastos, C. S., Higley, L. G., Ostlie, K. R., Madhavan, S., 2003.
1956 Photosynthetic responses of soybean to soybean aphid (Homoptera: Aphididae) injury. J

Econ Entomol 96, 188-193.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., Lopez, R., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47(W1):W636–W641.

Magalhaes, L.C., Hunt, T.E., Siegfried, B.D., 2008. Development of methods to evaluate susceptibility of soybean aphid to imidacloprid and thiamethoxam at lethal and sublethal concentrations. Entomol Exp Appl 128, 330-336.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen. Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg. J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Mathers, T.C., Chen, Y., Kaithakottil, G., Legeai, F., Mugford, S.T., Baa-Puyoulet, P., Bretaudeau, A., Clavijo, B., Colella, S., Collin, O., Dalmay, T., Derrien, T., Feng, H., Gabaldón, T., Jordan, A., Julca, I., Kettles, G.J., Kowitwanich, K., Lavenier, D., Lenzi, P., Lopez-Gomollon, S., Loska, D., Mapleson, D., Maumus, F., Moxon, S., Price, D.R.G., Sugio, A., van Munster, M., Uzest, M., Waite, D., Jander, G., Tagu, D., Wilson, A.C.C., van Oosterhout, C., Swarbreck, D., Hogenhout, S.A., 2017. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonize diverse plant species. Genome Biol 18 (1), 27.

Mahmood, Q., Bilal, M., Jan, S., 2014. Herbicides, Pesticides, and plant tolerance: An overview, in: P. Ahmad (Eds.), Emerging Technologies and Management of Crop Stress Tolerance, Vol.1, 423-448.

Malumphy, C.P., 1997. Morphology and anatomy of honeydew eliminating organs, in: Ben-Dov, A.Y., Hodgson, C.J. (Eds.), Soft Scale Insects: Their Biology, Natural enemies and Control, Vol. 7. Elsevier Science B.V., Amsterdam, The Netherlands, pp. 269-274.

McCarville, M.T., O'Neal, M.E., Pecinovsky,K.T., 2014. Evaluation of soybean aphid-
resistant soybean lines. Iowa State Research Farm Progress Reports. 2034. https://lib.dr.iastate.edu/farms_reports/2034/ (accessed 30 April 2019)

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol 17(1), 122.

McTavish, E.J., Hillis, D.M., 2015. How do SNP ascertainment bias schemes and population demographics affect inferences about population history. BMC Genomics 16, 266.

Mezey, Á., Szalay-Marzsó, L., 2001. Host plant preference of *Diuraphis noxia* (Kurdj.) (Hom., Aphididae). J Pest Science 74, 17-21.

Michel, A.P., Zhang, W., Jung, J.K., Kang, S-T., Rouf Mian, M.A., 2009. Population genetic structure of *Aphis glycines*. Mol Ecol Evol 38, 1301-1311.

Moran, A. N., 1988. The evolution of host-plant alternation in aphids: evidence for specialization as a dead end. Amer Nat 132, 681-706.

Morgan, M., Falcon, S., Gentleman, R., 2019. GSEABase: Gene set enrichment data structures and methods. https://rdrr.io/bioc/GSEABase/ (accessed 24 May 2019)

Mueller, F.P., 1985. Biotype Formation and Sympatric Speciation in Aphids (Homoptera: Aphidinea). Entomol Gen 10, 161-181.

Mueller, E.E., Frost, K.E., Esker, P., Gratton, C. 2010. Seasonal phenology of *Aphis glycines* (Hemiptera:Aphididae) and other aphid species in cultivated bean and non-crop habitats in Wisconsin. J Econ Entomol 103 (5), 1670-1681.

Myers, S.W., Hogg, D.B., Wedberg, J.L., 2005. Determining the optimal timing of a foliar insecticide applications for control of soybean aphid (Hemiptera: Aphididae) on soybean. J Econ Entomol 98, 2006-2012.

Nichol, H., Law, J.H., Winzerling, J., 2002. Iron metabolism in insects. Annu Rev Entomol 47, 535-559.

Nicholson, S.J., Nickerson, M.L., Dean, M., Song, Y., Hoyt, P.R., Rhee, H., Kim, C., Puterka, G.J., 2015. The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC genomics 16 (1), 429.

Nielsen, C., Hajek, A.E., 2005. Control of invasive soybean aphid, *Aphis glycines* (Hemiptera: Aphididae), populations by existing natural enemies in New York State, with emphasis on entomopathogenic fungi. Environ Entomol 34, 1036-1047.

Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168, 2373–2382.

Nielsen, R., 2005. Molecular signatures of natural selection. Annu Rev Genet 39, 197-218.

2048 Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005.
2049 Genomic scans for selective sweeps using SNP data. Genome Res 15, 1566-11575.
2050
2051 Nováková, E., Hypša, V., Klein, J., Foottit, R.G., von Dohlen, C.D., Moran, N.A., 2013.
2052 Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the
2053 obligate symbiont *Buchnera aphidicola*. Mol Phylogenetics Evol 68 (1), 42–54.
2054
2055 Paroush, Z., Finley, R.L., Kidd, T., Wainwright, S.M., Ingham, P.W., Brent, R.,
2056 Ishhorowicz, D., 1994. Groucho is required for *Drosophila* neurogenesis, segmentation,
2057 and sex determination and interacts directly with hairyrelated bHLH proteins. Cell 79,
2058 805–815.
2059
2060 Parry, H.R., 2013. Cereal movement: general principles and simulation modeling. Mov
2061 Ecol 1, 14.
2062
2063 Pedigo, L. P., Rice, M. E., 2009. Entomology and pest management, 6th ed. Prentice
2064 Hall, Upper Saddle River, NJ.
2065
2066 Powell, G., Tosh, C., Hardie, J., 2006. Host plant selection by aphids: Behavioral,
2067 evolutionary, and applied perspectives. Annu Rev Entomol 51, 309-30.
2068
2069 Quan, Q., Hu, X., Pan, B., Zeng, B., Wu, N., Fang, G., Cao, Y., Chen, X., Li, X., Huang,
2070 Y.,Zhan, S., 2019. Draft genome of the cotton aphid *aphis gossypii*. Insect Biochem Mol
2071 Biol 105, 25-32.
2072
2073 Quinlan, A., Hall, I., 2010. BEDTools: a flexible suite of utilities for comparing genomic
2074 features. Bioinformatics 26 (6), 841-842.
2075
2076 Rakauskas, R., 2000. Experimental ybridisation between *Aphis grossulariae* and *Aphis*
2077 *triglochinis* (Sernorrhyncha: Aphididae). Eur J Entomol 97, 377-386.
2078
2079 Ragsdale, D.W., Voegtlin, D.J., O'Neil, R.J., 2004. Soybean aphid biology in North
2080 America. Ann Entomol Soc Am 97(2), 204-208.
2081
2082 Ragsdale, D.W., Landis, D.A., Brodeur, J., Heimpel, G.E., Desneux, N., 2011. Ecology
2083 and management of the soybean aphid in North America. Annu Rev Entomol 56, 375-
2084 399.
2085
2086 Ragsdale, D.W., McCornack, B.P., Venette, R.C., Potter, B.D., Macrae, I.V., Hodgson,
2087 E.W., O'Neal, M.E., Johnson, K.D., O'Neil, R.J., DiFonzo, C.D., Hunt, T.E., Glogoza,
2088 P.A., Cullen, E.M., 2007. Economic threshold for soybean aphid (Hemiptera: Aphididae).
2089 J Econ Entomol 100, 1258-1267.
2090
2091 Rebers, J.E., Riddiford, L.M.,1988. Structure and expression of a *Manduca sexta* larval
2092 cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol 203, 411-423.
2093

2094 Rebers, J.E., Willis, J.H., 2001. A conserved domain in arthropod cuticular proteins binds
2095 chitin. Insect. Biochem. Mol. Biol. 31, 1083-1093.
2096
2097 Rispe, C., Legeai, F., Arora, A.K., Baa-Puyoulet, P., Barberà, M.M, Bouallègue, M.,
2098 Bretaudeau, A., Brisson, J.A., Calevro, F., Capy, P., Catrice, O., Chertemps, T., Couture,
2099 C., Douglas, A.E., Dufault-Thompson, K., Escuer, P., Feng, H., Fernández, R., Gabaldón,
2100 T., GenoTOOL platform, Guigó, R., Hilliou, F., Hinojosa, S., Hsiao, Y-M.,
2101 Hudaverdian, S., Jacquin-Joly, E., James, E., Johnston, S., Joubard, B., Le Goff, G., Le
2102 Trionnaire, G., Liu, S., Lu, H-L., Maïbèche, M., Martínez-Torres, D., Montagné, N.,
2103 Moran, N., Makni, M., Marcet-Houben, M., Meslin, C., Nabity, P., Papura, D., Parisot,
2104 N., Rahbé, Y., Robin, S., Roux, P., Rozas, J., Ripoll, A., Sánchez-Gracia, A., Sánchez-
2105 Herrero, J.F., Santesmasses, D., Tang, M., Thompson, K., Tian, W., van Munster, M.,
2106 Wemmer, J., Wilson, A.C.C., Zhang, Y., Zhao, C., Zhao, J., Zhao, S., Zhou, X.,
2107 International Aphid Genomics Consortium, Delmotte, F., Tagu, D. 2019. (in press).
2108 Insights on the genome evolution and invasion routes of grape phylloxera. Molecular
2109 Biology and Evolution
2110
2111 Rouault, T., Klausner, R., 1997. Regulation of iron metabolism in eukaryotes. Curr Top
2112 Cell Regul 35, 1-19.
2113
2114 Rouault, T., Tong, W-H., 2005. Iron-sulphur cluster biogenesis and mitochondrial iron
2115 homeostasis. Nat Rev Mol Cell Biol 6(4), 345-351.
2116
2117 Rutledge C.E., O'Neil R.J., 2005. *Orius insidiosus* (Say) as a predator of the soybean
2118 aphid, *Aphis glycines* Matsumura. Biol Control 33 (1), 56-64.
2119
2120 Sama, S., Saleh, K.M., van Halteren, P., 1974. Research reports 1969–1974, in: Varietal
2121 screening for resistance to the aphid, *Aphis glycines*, in soybean. Agricultural
2122 Cooperation, Indonesia-the Netherlands, pp. 171–172.
2123
2124 Sass, M.E., Navarro, F.M., German, T.L., Nienhuis, J., 2004. The search for resistance to
2125 the soybean aphid virus complex in snap beans. Annual Report Bean Improvement
2126 Cooperative 47, 65-66.
2127
2128 Shufran, K.A., Payton, T.L., 2009. Limited genetic variation within and between Russian
2129 wheat aphid (Hemiptera: Aphididae) biotypes in the United States. J Econ Entomol
2130 102(1):440-5.
2131
2132 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015.
2133 BUSCO: assessing genome assembly and annotation completeness with single-copy
2134 orthologs. Bioinformatics 31(19), 3210-2.
2135
2136 Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R.,
2137 McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast,
2138 scalable generation of high-quality protein multiple sequence alignments using Clustal
2139 Omega. Mol Syst Biol 7, 539.

Song, F.S., Swinton, S.M., DiFonzo, C., O'Neal, M., Ragsdale, D.W., 2006. Profitability analysis of soybean aphid control treatments in three north-central states. Michigan State University Department of Agricultural Economics: Staff Paper, 2006-24.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9), 1312–1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Res 34, W435-W439.

Takahashi, S., Inaizumi, M., Kawakami, K., 1993. Life cycle of the soybean aphid *Aphis glycines* Matsusmura in Japan. Jap J Appl Entomol Zool 37, 207-212.

The International Aphid Genomics Consortium, 2010. Genome sequence of the Pea Aphid *Acyrthosiphon pisum*. PLoS Biology 8 (2), e1000313.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9),1105-11.

Uzest, M., Gargani, D., Drucker, M., Hébrard, E., Garzo, E., Candresse, T., Fereres, A., Blanc, S., 2007. A protein key to plant virus transmission at the tip of the insect vector stylet. PNAS 104, 17959–17964.

van Emden, H.F., Harrington, R., 2007. Aphids as Crop Pests. CABI Publishing, London, UK.

Venette, R.C., Ragsdale, D.W., 2004. Assessing the invasion by soybean aphid (Homoptera: Aphididae): Where will it end? Ann Entomol Soc Am 97, 219-226.

Voegtlin, D. J., O'Neil, R.J., Graves, W.R., 2004. Tests of suitability of overwintering hosts of *Aphis glycines*: identification of a new host association with *Rhamnus alnifolia* L'Héritier. Ann Entomol Soc Am 97, 233-234.

Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C., 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34 (6), 1692–1699.

Wang, C. L., Siang, N.J., Chang, G.S., Chu, H.F., 1962. Studies on the soybean aphid, *Aphis glycines* Mastumura. Acta Entomol Sinica 11, 31-44.

Wang, C.P., Zhao, H.Y., Zhang, G.S., Luo, K., 2011b. Location of Sa1 genes for resistance to wheat aphid. Crop Science Society of China.

Wang, L., Zhang, S., Luo, J.Y., Wang, C.Y., Lv, L.M., Zhu, X.Z., Li, C.H., Cui, J.J.,

2184 2016. Identification of *Aphis gossypii* Glover (Hemiptera: Aphididae) biotypes from
2185 different host plants in north China. PLoS One 11, e0146345.
2186
2187 Ward, S.A., Leather, S.R., Pickup, J., Harrington, R., 1998. Mortality during dispersal
2188 and the cost of host specificity in parasites: how many aphids find hosts? J Anim
2189 Ecol 67, 763–773.
2190
2191 Webster, C.G., Pichon, E., van Munster, M., Monsion, B., Deshoux, M., Gargani, D.,
2192 Calevro, F., Jimenez, J., Moreno, A., Krenz, B., Thompson, J.R., Perry, K., Fereres, A.,
2193 Blanc, S., Uzest, M., 2018. Identification of plant virus receptor candidates in the stylets
2194 of their aphid vectors. J Virol 92(14), e00432-18.
2195
2196 Webster, C.G., Thillier, M., Pirolles, E., Cayrol, B., Blanc, S., Uzest, M., 2017.
2197 Proteomic composition of the acrostyle: Novel approaches to identify cuticular proteins
2198 involved in virus-insect interactions. Insect Sci 24 (6), 990-1002.
2199
2200 Wenger, J.A., Cassone, B.J., Legeai, F., Johnston, J.S., Bansal, R., Yates, A.D., Coates,
2201 B.S., Pavinato, V.A.C., Michel, A. 2017. (in press). Whole genome sequence of the
2202 soybean aphid, *Aphis glycines*. Insect Biochem Mol Biol
2203 https://doi.org/10.1016/j.ibmb.2017.01.005
2204
2205 Willis, J.H, Iconomidou, V.A., Smith, R.F., Hamodrakas, S.J., 2005. Cuticular proteins,
2206 in: L.I. Gilbert, K. Iatrou,, S.S. Gill (Eds.), Comprehensive Molecular Insect Science,
2207 Vol. 4., 79-109.
2208
2209 Willis, J.H., 2010. Structural cuticular proteins from arthropods: Annotation,
2210 nomenclature, and sequence characteristics in the genomics era. Insect Biochem Mol Biol
2211 40: 189-204.
2212
2213 Wu, Z.S., Schenk-Hamlin, D., Zhan, W.Y., Ragsdale, D.W., Heimpel, G.E., 2004. The
2214 soybean aphid in China: A historical review. Ann Entomol Soc Am 97, 209-218.
2215
2216 Wyckhuys, K.A.G., Hopper, K.R., Wu, K-M., Straub, C., Gratton, C., Heimpel, G.E.,
2217 2007. Predicting potential ecological impact of soybean aphid biological control
2218 introductions. Biocontrol News and Information 28, 30-34.
2219
2220 Xi, J., Pan, Y., Bi, R., Gao, X., Chen, X., Peng, T., Zhang, M., Zhang, H., Hu, X., Shang,
2221 Q., 2015. Elevated expression of esterase and cytochrome P450 are related with lambda-
2222 cyhalothrin resistance and lead to cross resistance in *Aphis glycines* Matsumura. Pest
2223 Biochem Physiol 118, 77–81.
2224
2225 Zhang, G. X., and T. S. Zhong. 1982. Experimental studies on some aphid life-cycle
2226 patterns. Sinozoologia 2, 7-17.
2227
2228 Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The
2229 MaSuRCA genome assembler. Bioinformatics 29(21), 2669-77.

Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Yorke, J.A., Dvorak, J., Salzberg, S., 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. Genome Res 27(5), 787-792.

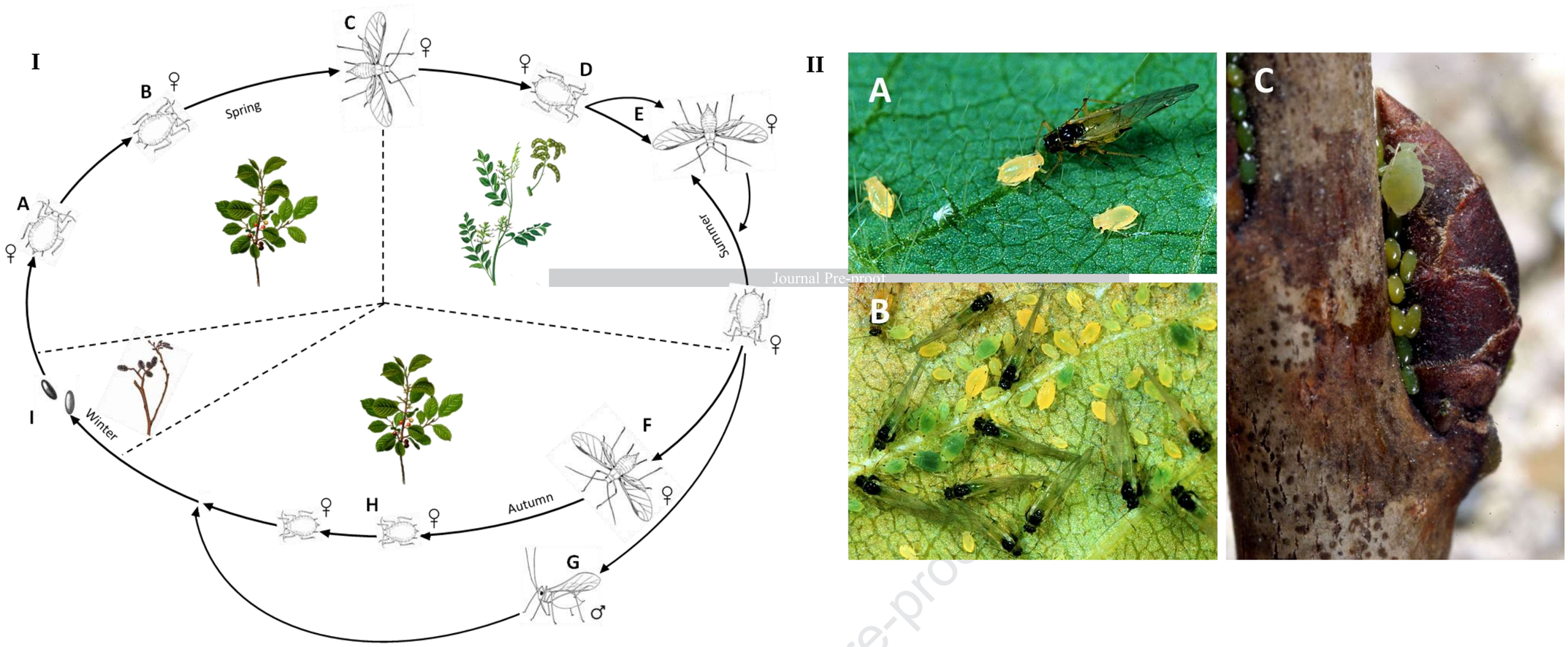


Fig. 1. Life cycle of the soybean aphid (*Aphis glycines* Matsumura). (A) Fundatrix on Rhamnus spp.; (B) Apterous viviparous female on Rhamnus spp.; (C) Alate viviparous female, spring migrant from Rhamnus spp. to soybean; (D) Apterous viviparous female on soybean; (E) Alate viviparous female, summer migrant; (F) Gynopara, fall migrant from soybean to Rhamnus spp.; (G) Male migrates from soybean to Rhamnus spp.; (H) Ovipara, on Rhamnus spp.; (I) Overwintering egg on Rhamnus spp. (II) Representations of different life stages of *A. glycines* on their summer and overwintering hosts. A. Alate and nymphs on a soybean leaf. B. Gynoparae and abundance of nymphs that will develop into ovipara on a leaf of *Rhammus catharctica.* C. Ovipara and eggs adjacent to a bud of *R. cathartica. (*Photo credits David Voegtlin).

Table 1 Comparison of assembly statistics for currently available aphid genomes. Entries with an asterix (*) indicate genome sequence assemblies not available at GenBank but at AphidBase.

Statistics	A. glycines Bt1	A. glycines Field Pop.	A. gossypii	M. persicae	M. cerasi	A. pisum	D. noxia	M. sacchari	R. maidis	R. padi	S. graminum	S. flava
GenBank Accession	NA*	NA*	GCF_004010815.1	GCF_001856785.1	NA*	GCF_000142985.2	GCF_001186385.1	GCF_002803265.2	GCA_003676215.3	NA*	GCA_003264975.1	GCF_003268045.1
# Scaffolds	3,224	8,397	4,718	4,021	49,286	23,925	5,637	1,347	220	15,587	7,859	1,923
Genome (Scaffolds) Size Mb	308	303	294	347	406	542	395	300	326	319	385	353
Longest Scaffold Mb	23.00	1.00	5.00	2.00	0.26	3.00	2.00	26.00	94.00	0.62	13.00	8.00
Shortest Scaffold nt	60	2000	889	959	1001	200	928	1662	1096	1001	1004	1000
# Scaffolds > 500 nt	3,209 (99.5%)	8,397 (100.0%)	4,718 (100.0%)	4,021 (100.0%)	49,286 (100.0%)	23,451 (98.0%)	5,637 (100.0%)	1,347 (100.0%)	220 (100.0%)	15,587 (100.0%)	7,859 (100.0%)	1,923 (100.0%)
# Scaffolds > 1K nt	3,208 (99.5%)	8,397 (100.0%)	4,487 (95.1%)	4,017 (99.9%)	49,286 (100.0%)	12,914 (54.0%)	5,613 (99.6%)	1,347 (100.0%)	220 (100.0%)	15,587 (100.0%)	7,859 (100.0%)	1,922 (99.9%)
# Scaffolds > 10K nt	410 (12.7%)	2,716 (32.3%)	1,574 (33.4%)	1,845 (45.9%)	9,745 (19.8%)	2,355 (9.8%)	2,941 (52.2%)	808 (60.0%)	155 (70.5%)	3,832 (24.6%)	2,425 (30.9%)	860 (44.7%)
# Scaffolds > 100K nt	121 (3.8%)	968 (11.5%)	683 (14.5%)	788 (19.6%)	178 (0.4%)	1,106 (4.6%)	902 (16.0%)	161 (12.0%)	8 (3.6%)	940 (6.0%)	325 (4.1%)	318 (16.5%)
# Scaffolds > 1M nt	55 (1.7%)	1 (0.0%)	33 (0.7%)	38 (0.9%)	0 (0.0%)	89 (0.4%)	34 (0.6%)	78 (5.8%)	4 (1.8%)	0 (0.0%)	93 (1.2%)	122 (6.3%)
Mean Scaffold size Kb	95	36	62	86	8	22	70	223	1,481	20	49	183
Median Scaffold size Kb	3	4	3	7	3	1	10	12	20	3	7	9
N50 Scaffold Length Mb	6.00	0.10	0.44	0.44	0.02	0.50	0.40	3.00	93.00	0.12	1.29	1.68
L50 Scaffold Count	15	512	195	224	4472	280	281	25	2	782	71	67
Scaffold %A	35.77	36.08	34.47	34.82	35.04	32.41	26.57	36.18	36.15	36.09	33.71	34.45
Scaffold %C	13.4	13.88	12.88	14.94	14.93	13.73	10.89	13.22	13.85	13.88	12.98	14.8
Scaffold %G	13.41	13.87	12.9	14.93	14.93	13.73	10.89	13.22	13.84	13.89	12.99	14.8
Scaffold %T	35.73	36.03	34.31	34.78	35.05	32.4	26.57	36.2	36.15	36.12	33.7	34.46
Scaffold %N	1.68	0.14	5.44	0.53	0.05	7.71	24.94	1.17	0.01	0.02	6.63	1.49
Scaffold N Mb	5.18	0.42	16	1.84	0.2	41.78	98.53	3.52	0.05	0.05	25.52	5.26
% Assembly in Scaffolded Contigs	0.831	0.267	0.959	0.761	0.196	0.951	0.99	0.915	0.984	0.224	0.842	0.924
% Assembly in Unscaffolded Contigs	0.169	0.733	0.041	0.239	0.804	0.049	0.01	0.085	0.016	0.776	0.158	0.076
Average Length of Ns Between Contigs	14284	316	2152	941	93	1139	2185	3785	100	99	4839	3132
# Contigs	3,587	9,610	12,144	5,971	51,353	60,594	50,723	2,276	689	16,133	13,128	3,599
# Contigs in Scaffolds	530	2,223	9,224	3,020	3,858	41,082	48,794	1,180	473	998	6,404	2,084
# Contigs not in Scaffolds	3,057	7,387	2,920	2,951	47,495	19,512	1,929	1,096	216	15,135	6,724	1,515
Contigs Size Mb	303	303	278	345	405	500	296	298	326	319	360	348
Longest Contig Mb	7.7	0.88	0.71	1.5	0.21	0.42	0.17	2.4	42.51	0.57	0.78	2
Shortest Contig	60	0	415	1	1001	200	60	81	1096	1001	48	146

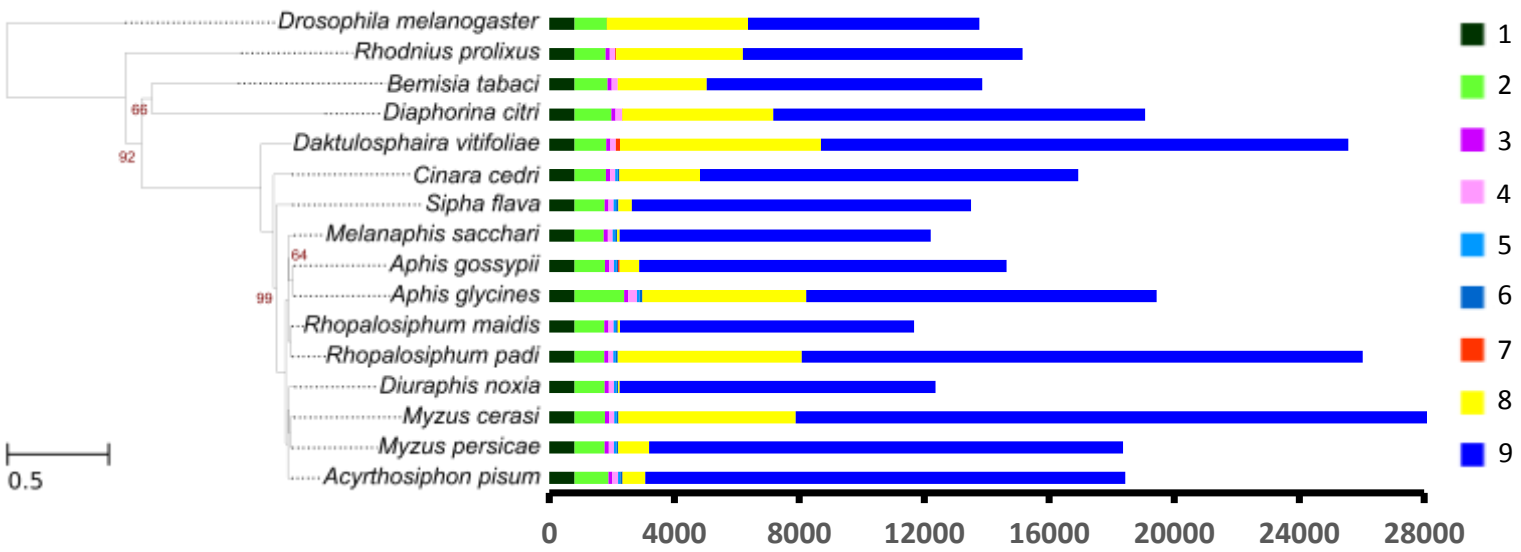


Fig. 2. Species tree obtained from the concatenation of 67 widespread single-gene families using *D. melanogaster* as the outgroup. Bootstrap values below 100% are indicated in red, the rest are not shown. Bars on the right represent relationships of orthologous genes among different taxa used in the analysis. 1) 811 single copy genes present in all taxa; 2) Multi copy genes present in all taxa (range: 935-1,589); 3) 130 single copy Hemiptera-specific genes; 4) Multi copy Hemiptera-specific genes (range: 155-276); 5) 81 single copy aphid-specific genes; 6) Multi copy aphid-specific genes (range: 52-93); 7) Single copy species-specific genes (range: 1-140); 8) Multi copy species-specific genes (range: 56-6,426); 9) Remaining genes not included in the previous categories. The genomic resources for *C. cedri* and *D. vitifoliae* are not publicly accessible, and were kindly made available prior to publication by Toni Gabaldon and Denis Tagu.

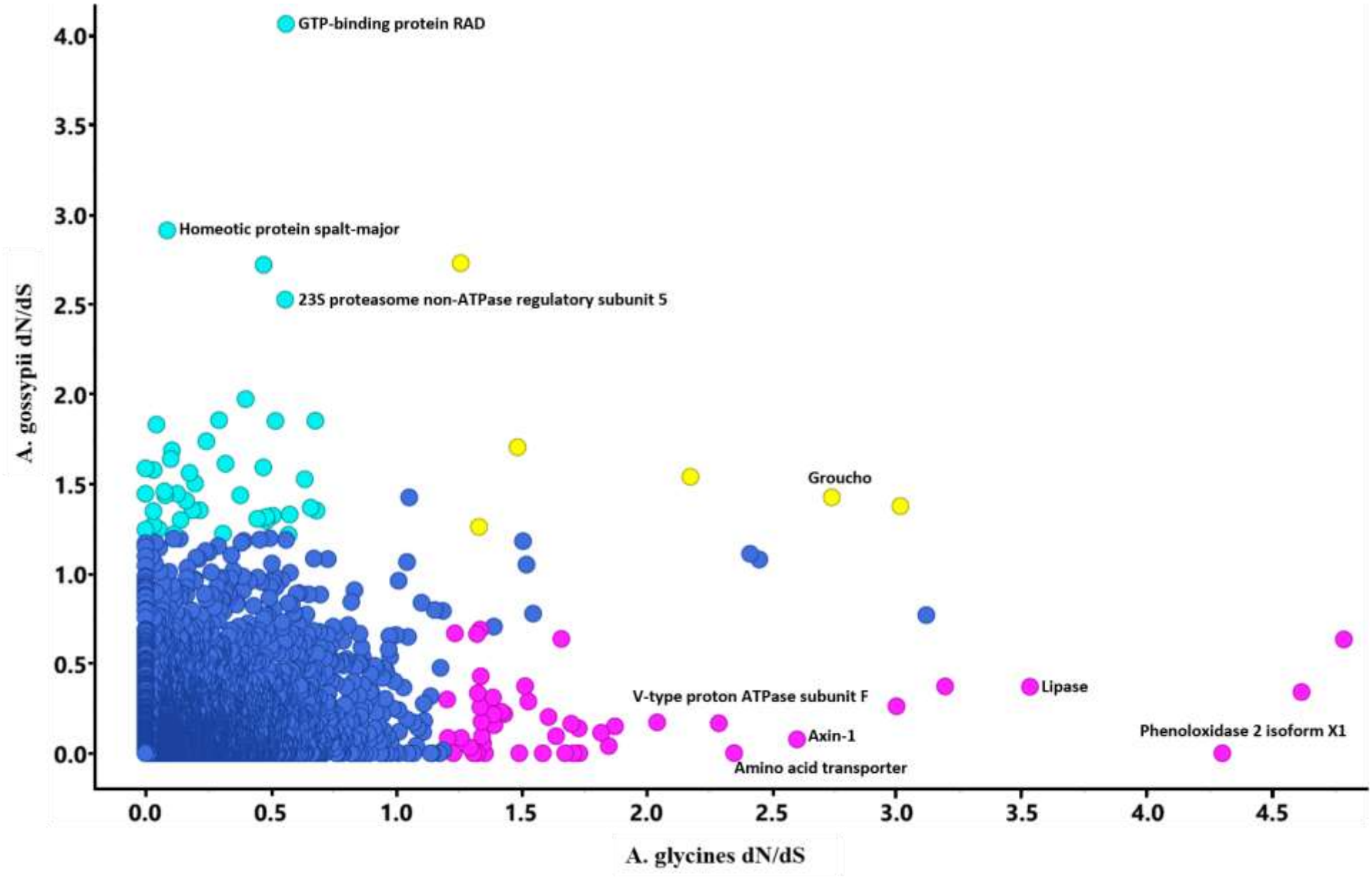

GTP-binding protein RAD
Homeotic protein spalt-major
23S proteasome non-ATPase regulatory subunit 5
Groucho
Lipase
V-type proton ATPase subunit F
Axin-1
Amino acid transporter
Phenoloxidase 2 isoform X1
A. gossypii dN/dS
A. glycines dN/dS
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

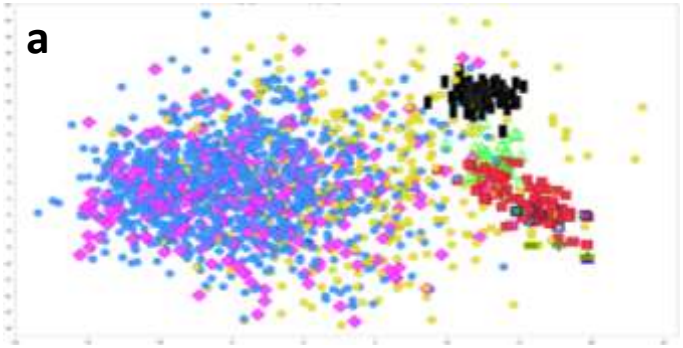
Fig. 3. dN/dS ratios for 3,825 one to one orthologous genes between A. *glycines* and A. *gossypii* that passed the filtering cutoffs (see Materials and Methods for cutoff values). Genes under selection with dN/dS values >2 and with available annotations are labeled with the specific name of the gene. Those under selection in both A. *glycines* and A. *gossypii* are in yellow circles, those in A. *glycines* are indicated in pink, and those in A. *gossypii* are represented in aqua marine.

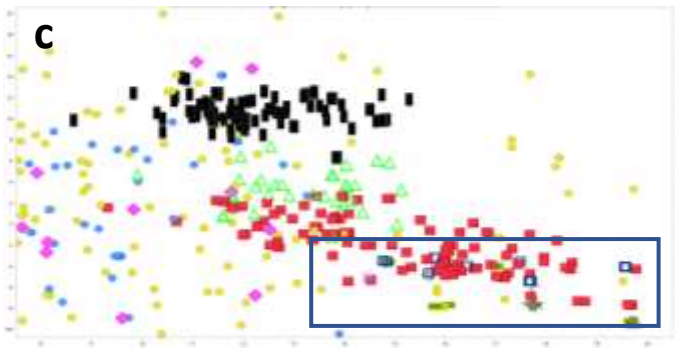
Table 2 Tally of all SBA samples used in the population structure analysis. Samples are listed by country, region, and year collected. A total of 3791 samples were collected and analyzed. Samples indicated with asterix (*) represent those collected from experimental Rag plots. Abbreviations used are as follows: Australia (NSW, New South Wales; QLD, Queensland); Canada (MB, Manitoba; ON, Ontario; QC, Quebec); USA (IA, Iowa; IL, Illinois; IN, Indiana; KY, Kentucky; MI, Michigan; MN, Minnesota; MO, Missouri; NY, New York; ND, North Dakota; OH, Ohio; PA, Pennsylvania; SD, South Dakota; VA, Virginia; WI, Wisconsin)

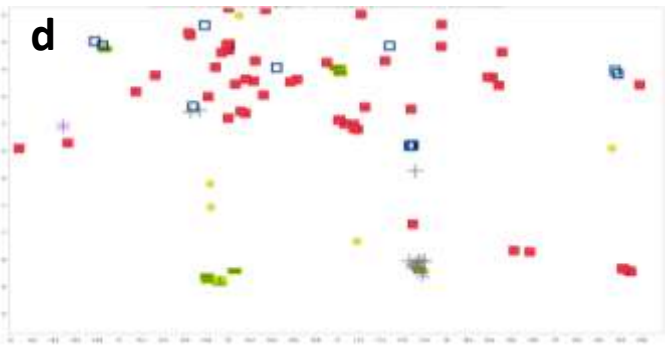
Location	Year	# of Samples
Asia		***656***
China		**167**
Guangxi	2008	10
Hebei	2008	7
	2010	11
Hei long jiang	2001	12
	2008	10
Hubei	2007	10
Jiangsu	2010	24
Jilin	2001	12
	2010	24
Shanxi	2008	12
	2010	24
Zhejiang	2008	11
Indonesia		**112**
Cianjur	2013	57
Lombok	2010	5
Majalengka	2013	10
Malang	2010	24
Maros	2012	15
Sakabumi	2013	1
Japan		**244**
Aomori	2008	9
	2010	24
Furukawa	2001	11
Ibaraki	2001	12
Iwate	2008	5
Unknown loc	2001	12
Morioka	2001	12
Nagano	2010	24
Shimane	2010	6
Tochigi	2001	12
	2008	22
	2011	48
Yamagata	2001	12
Yamaguchi	2008	11
	2010	24
Myanmar		**48**
Shan	2013	48

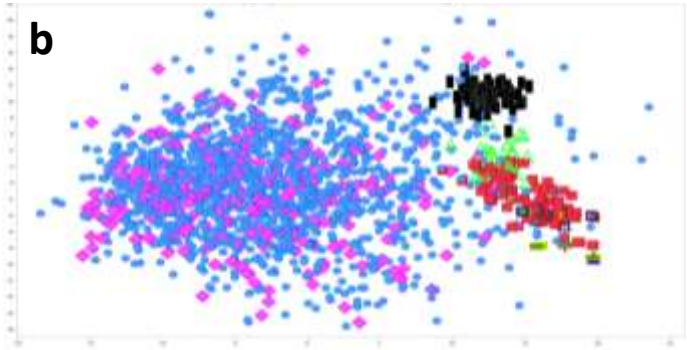
Location	Year	# of Samples
North America		***3104***
Canada		**457**
MB	2011	167
ON	2003	28
	2011	85
QC	2004	55
	2011	88
	2012	34
USA		**2647**
IA	2010	9
	2011	147
	2012	73
	2013	93*
IL	2001	5
	2005	34
	2008	17
	2009	55
	2010	35
	2011	23
IN	2011	22
KY	2001	23
MI	2001	96
	2006	15
MN	2001	13
	2005	7
	2009	12
	2010	92
	2011	192
	2012	339
	2013	113*
MO	2001	10
ND	2009	21
	2011	34
	2013	76*
NY	2011	38
	2012	72
OH	2001	132
	2010	12
	2013	91*

South Korea		**50**
Asan	2012	12
Cheonan	2011	14
Muan	2012	12
Suwon	2011	12
Taiwan		**18**
Kao-Usuing	2003	6
	2011	12
Thailand	**2011**	**17**
Australia		***31***
NSW	2004	7
QLD	2012	24
PA	2001	108
	2010	12
	2011	23
SD	2008	23
	2009	20
	2011	96
	2012	96
	2013	83*
VA	2009	16
WI	2009	19
	2010	109*
	2011	141





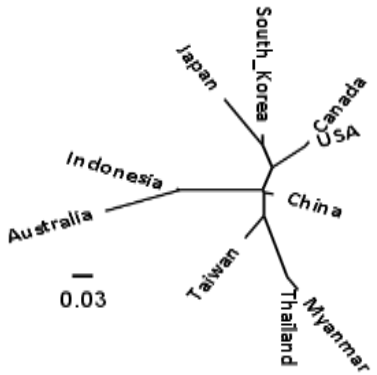




e

Weir & Cockerham Weighted Fst	All Asian countries	USACanada	USA	Canada	South Korea	China	Japan	Taiwan	Thailand	Indonesia	Myanmar	Australia
USACanada	0.10213				0.13017	0.12839	0.1656	0.18899	0.21864	0.21015	0.26436	0.28606
USA				0.0090418	0.12746	0.12594	0.1638	0.18604	0.21583	0.2073	0.26177	0.28312
Canada			0.00904		0.15431	0.15058	0.19	0.21551	0.24499	0.24257	0.29567	0.31758
South Korea		0.13017	0.12746	0.15431		0.07867	0.1124	0.17832	0.23015	0.24096	0.31432	0.34074
China		0.12839	0.12594	0.15058	0.07867		0.176	0.09318	0.14692	0.14837	0.21823	0.24743
Japan		0.16562	0.16383	0.19001	0.1124	0.17596		0.28112	0.31976	0.32598	0.37107	0.40174
Taiwan		0.18899	0.18604	0.21551	0.17832	0.09318	0.2811		0.11611	0.18492	0.22003	0.34211
Thailand		0.21864	0.21583	0.24499	0.23015	0.14692	0.3198	0.11611		0.21841	0.02663	0.38457
Indonesia		0.21015	0.2073	0.24257	0.24096	0.14837	0.326	0.18492	0.21841		0.29461	0.11873
Myanmar		0.26436	0.26177	0.29567	0.31432	0.21823	0.3711	0.22003	0.02663	0.29461		0.44988
Australia	0.21471	0.28606	0.28312	0.31758	0.34074	0.24743	0.4017	0.34211	0.38457	0.11873	0.44988	

f



g

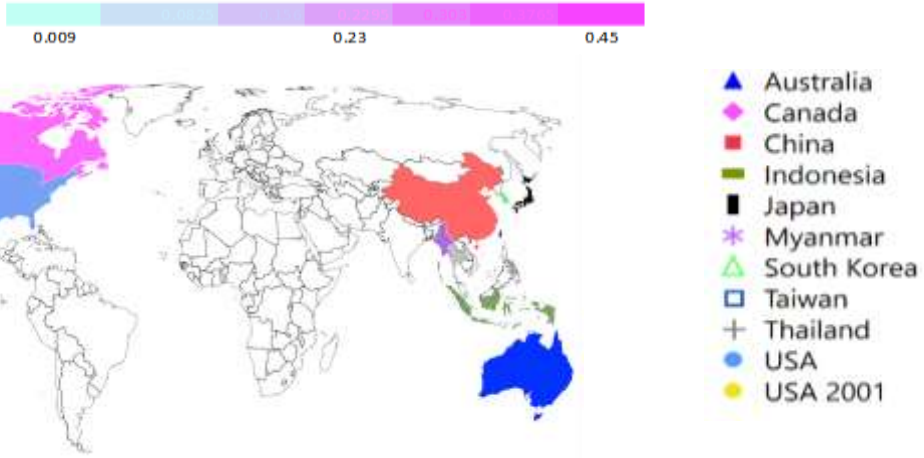
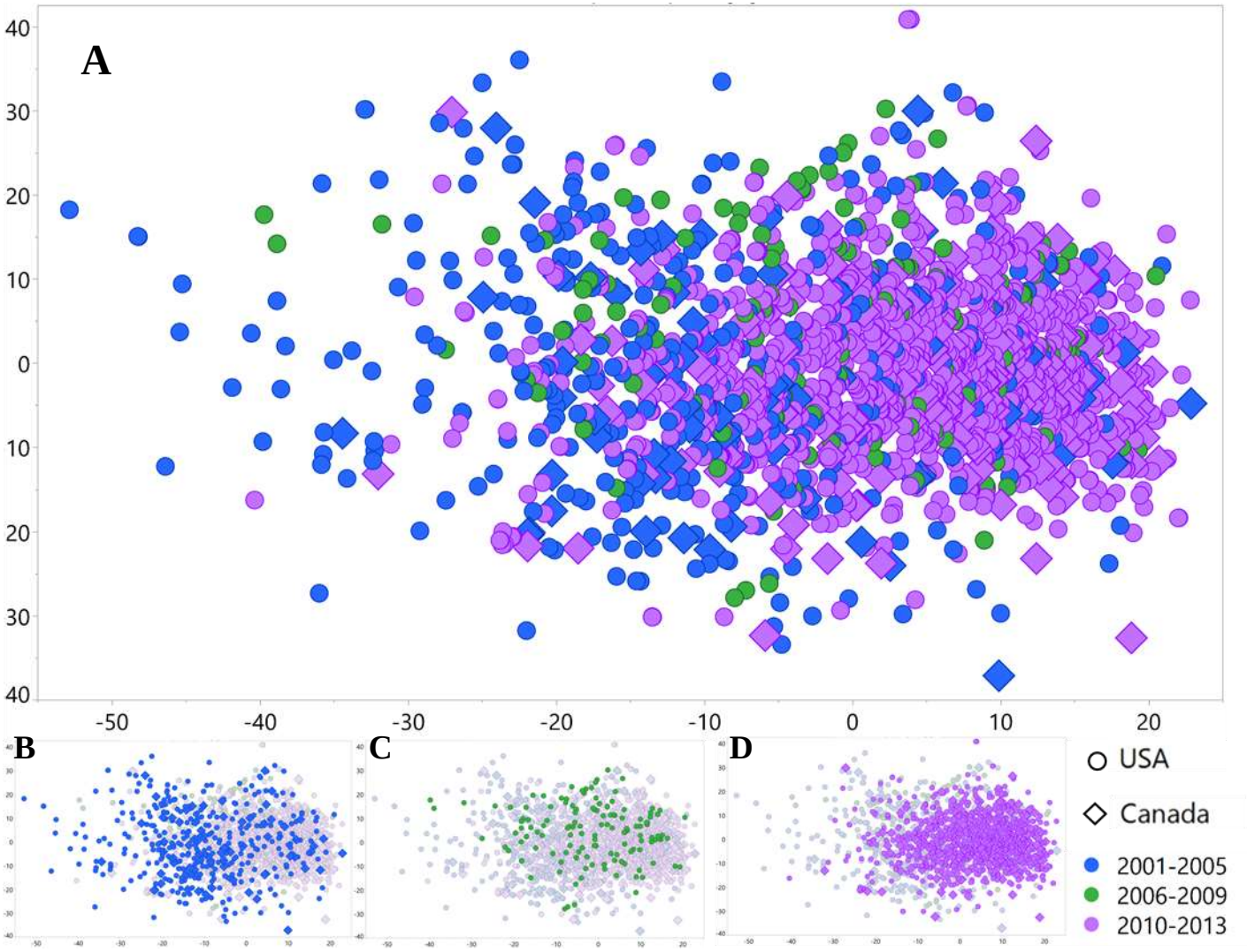


Fig. 4. Population structure analysis of the SBA world wide geographic distribution, Asia, Australia and North America, using 926 SNPs with minimized ascertainment bias. (a) PCA of samples for all populations for all years with 2001 US samples in yellow (X-axis PC1; Y-axis PC2); (b) PCA of samples for all populations for all years; (c) Enlargement of Asian and Australian populations indicated in rectangle in (a); (d) Enlargement of Australian and Indonesian populations indicated in square in (c). (e) F_{st} values for all pairwise comparisons of populations used in this study calculated according to Weir and Cockerham (1984). Color scale under the table indicates relationship between color and F_{st} level; (f) Neighbor Joining tree for all populations generated using F_{st} values as distances using the program QuickTree; (g) World map indicating the countries whose SBA populations were sampled. Colors in map correspond to the colors usen in the PCA plots.



	CA 2003	CA 2004	US 2005	US 2006	US 2008	US 2009	US 2010	US 2011	US/CA 2011	US 2012	US/CA 2012	US 2013
US 2001	0.006	0.011	0.115	0.009	0.022	0.034	0.033	0.028	0.032	0.041	0.029	0.035

Fig. 5. PCA of all samples from Canada and U.S. from 2001 to 2013 divided by three time periods (X-axis PC1; Y-axis PC2): 2001-2005; 2006-2009; 2010-2013 generated using 2,380 SNPs. (A) All time periods combined. (B) Same as A but with 2001-2005 period highlighted. (C) Same as A but with 2006-2009 period highlighted. (D) Same as A but with 2010-2013 highlighted. Table at the bottom of the figure shows F_{st} values for comparisons between 2001 and each year of sample collection for U.S. and Canada.

Fig. 6. Heatmap of F_{st} values calculated by comparing the allele frequencies for all 2,380 SNPs for SBA samples collected in 2001 against those collected yearly from 2003 to 2012 and represented in their respective columns. Similar to the Manhattan plot (Fig. 11), scaffolds are sorted by lengths with the longest one at the top of the column, SNPs within scaffolds are sorted in ascending order of their coordinates on the scaffolds. Each row represents the same SNP across the years sampled. Intensity of color indicates level of F_{st} value as represented in the scale bar on the top right corner.

Table 3 List of enriched Gene Ontology (GO) terms that were identified repeatedly in more than one year, for genes overlapping with SNPs having F_{st} values greater than or equal to 0.1 for the comparison between U.S. 2001 and those from years with the highest number of samples (2009, 2010, 2011, 2012). P-values are from overrepresentation analysis. GO class abbreviations: BP= Biological Processes, CC = Cellular Components, MF = Molecular Function

GO Class	GO ID	Term	2009		2010		2011		2012	
			p-value	#sign genes	p-value	#sign genes	p-value	#sign genes	p-value	#sign genes
BP	GO:0070887	cellular response to chemical stimulus	0.035	2						
	GO:0051716	cellular response to stimulus					0.049	15		
	GO:0051716	cellular response to stimulus							0.031	10
	GO:0007166	cell surface receptor signaling pathway			0.023	6				
	GO:0007166	cell surface receptor signaling pathway					0.026	5		
	GO:0050794	regulation of cellular process					0.003	24		
	GO:0050794	regulation of cellular process							0.005	15
	GO:0065007	biological regulation					0.017	25		
	GO:0065007	biological regulation							0.028	15
	GO:2001141	regulation of RNA biosynthetic process					0.026	7		
	GO:2001141	regulation of RNA biosynthetic process							0.023	5
	GO:0006355	regulation of transcription, DNA-templated					0.026	7		
	GO:0006355	regulation of transcription, DNA-templated							0.023	5
	GO:0023052	signaling					0.029	14		
	GO:0023052	signaling							0.032	9
	GO:0007154	cell communication					0.029	14		
	GO:0007154	cell communication							0.032	9
	GO:0019219	regulation of nucleobase-containing compound metabolic process					0.031	7		
	GO:0019219	regulation of nucleobase-containing compound metabolic process							0.026	5

	GO:0016459	myosin complex			0	5				
	GO:0016459	myosin complex					0.014	3		
	GO:0016459	myosin complex							0.036	2
CC	GO:0098802	plasma membrane receptor complex			0.015	2				
	GO:0098802	plasma membrane receptor complex					0.009	2		
	GO:0098803	plasma membrane receptor complex							0.003	2
	GO:0005887	integral component of plasma membrane					0.034	3		
	GO:0005888	integral component of plasma membrane							0.006	3
	GO:0042578	phosphoric ester hydrolase activity	0.01	5						
	GO:0008081	phosphoric diester hydrolase activity					0.002	4		
	GO:0032555	purine ribonucleotide binding	0.02	19						
	GO:0032555	purine ribonucleotide binding			0.038	28				
	GO:0097367	carbohydrate derivative binding	0.02	19						
	GO:0097367	carbohydrate derivative binding			0.034	30				
MF	GO:0004114	3',5'-cyclic-nucleotide phosphodiesterase activity	0.024	2						
	GO:0004114	3',5'-cyclic-nucleotide phosphodiesterase activity			0.007	3				
	GO:0004114	3',5'-cyclic-nucleotide phosphodiesterase activity					0.003	3		
	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides			0.02	15				
	GO:0016787	hydrolase activity			0.023	35				
	GO:0016788	hydrolase activity, acting on ester bonds					0.035	8		

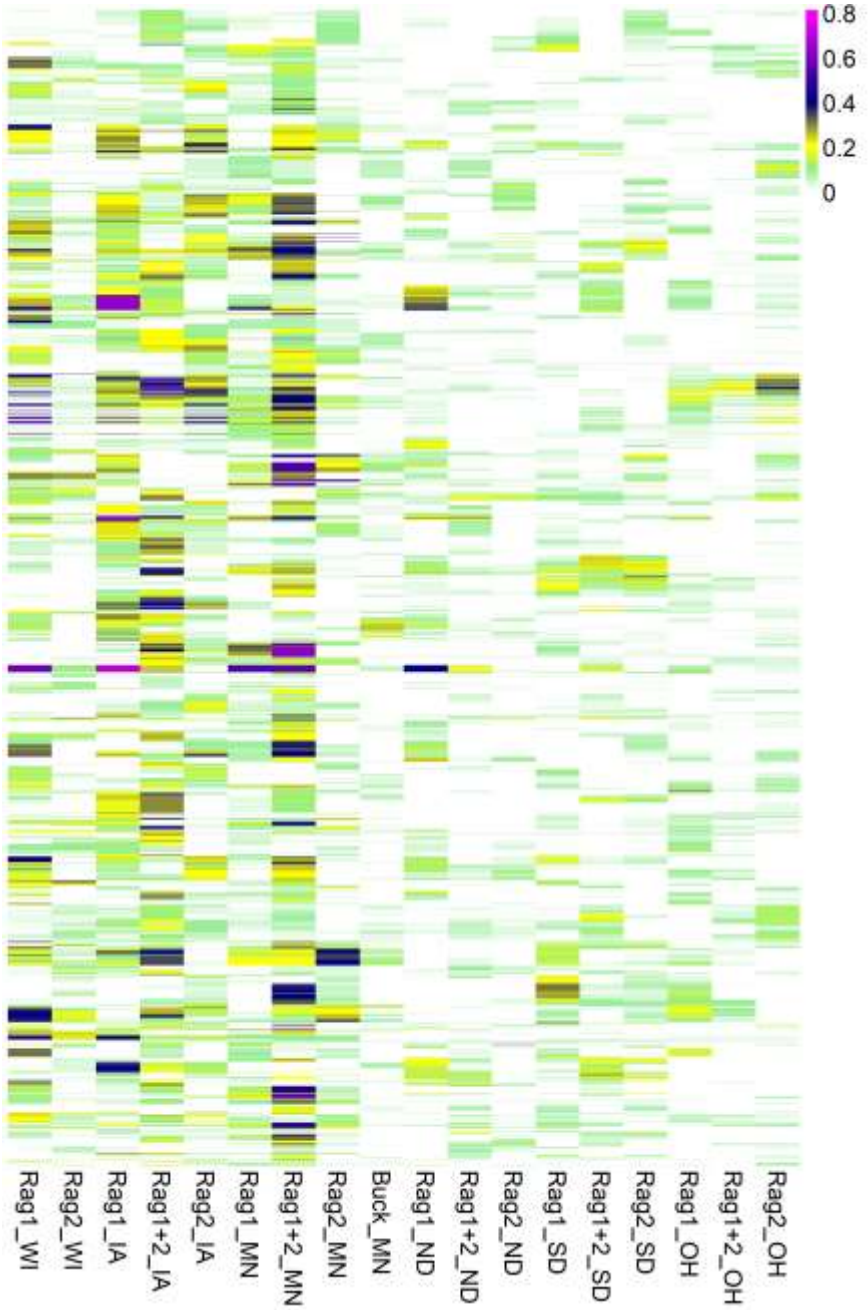


Fig. 7. Heat map of F_{st} values calculated by comparing the allele frequencies for all 2,380 SNPs for SBA field Rag experimental samples against SBA susceptible. Each row is a SNP, intensity of color indicates level of F_{st} value as represented in the scale bar on the top right corner. WI samples were collected in 2010, all other samples were collected in 2013. Abbreviations used are as follows: Buck, Buckthorn; IA, Iowa; MN, Minnesota; ND, North Dakota; SD, South Dakota; WI, Wisconsin.

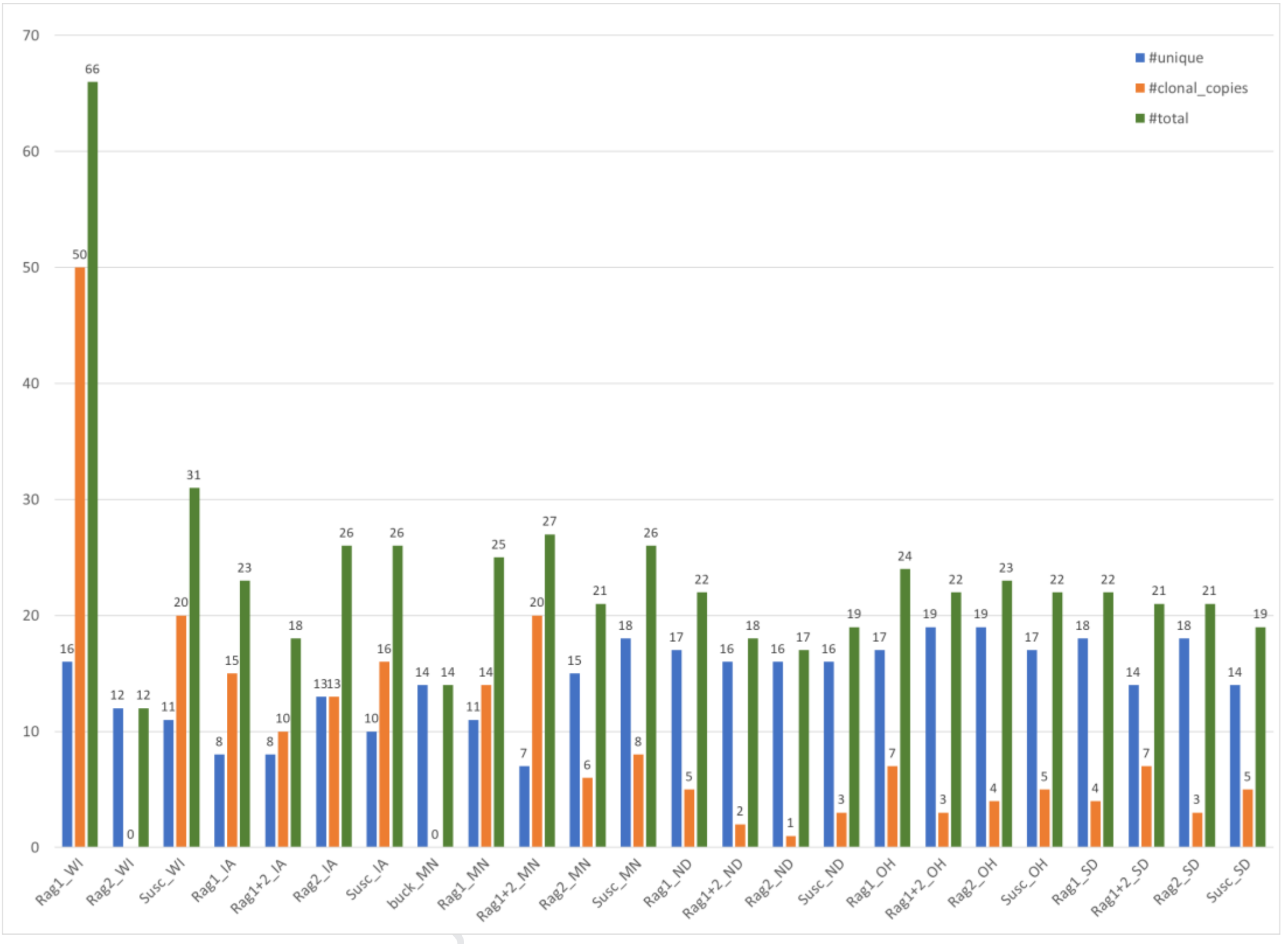


Fig. 8. Histogram of total aphid numbers (Y-axis) sampled for localities sampled and their respective Rag varieties and buckthorn plants (X-axis) and their corresponding unique and clonal individuals. Sample locations are indicated as WI, Wisconsin; IA, Iowa; MN, Minnesota; ND, North Dakota; OH, Ohio; SD, South Dakota.

Table 4 List of enriched Gene Ontology (GO) terms for genes overlapping with SNPs having F_{st} values greater than or equal to 0.1 for the comparison between Rag and susceptible plant varietiesfor WI, 2010; IA and MN 2013. P-values are from overrepresentation analysis. GO class abbreviations: BP= Biological Processes, CC = Cellular Components, MF = Molecular Function

GO Class	GO ID	Term	2010 WI Rag1		2013 IA Rag1		2013 IA Rag1+2		2013 MN Rag1		2013 MN Rag1+2		2013 MN Rag2	
			p-value	#sign genes	p-value	#sign genes	p-value	#sign genes	p-value	#sign genes	p-value	#sign genes	p-value	#sign genes
BP	GO:0007600	sensory perception	0.023	4										
	GO:0007605	sensory perception of sound							0.004	2				
	GO:0018205	peptidyl-lysine modification	0.029	3										
	GO:0018193	peptidyl-amino acid modification											0.03	2
	GO:0016192	vesicle-mediated transport			0.04	8								
	GO:0016192	vesicle-mediated transport					0.032	10						
	GO:0010038	response to metal ion			0.03	3								
	GO:0010038	response to metal ion							0.025	2				
	GO:0001505	regulation of neurotransmitter levels			0.03	2								
	GO:0001505	regulation of neurotransmitter levels					0.045	2						
CC	GO:0031010	ISWI-type complex	0.025	2										
	GO:0031010	ISWI-type complex			0.03	2								
MF	GO:0050660	flavin adenine dinucleotide binding	0.017	5										
	GO:0050660	flavin adenine dinucleotide binding	0.02	5										
	GO:0016705	oxidoreductase activity	0.02	4										
	GO:0016614	oxidoreductase activity, CH-CH donors					0.038	5						
	GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors									0.015	5		
	GO:0051536	iron-sulfur cluster binding					0.034	4						
	GO:0051539	4 iron, 4 sulfur cluster binding									0.024	2		
	GO:0003994	aconitate hydratase activity					0.041	2						
	GO:0003994	aconitate hydratase activity							0.004	2				

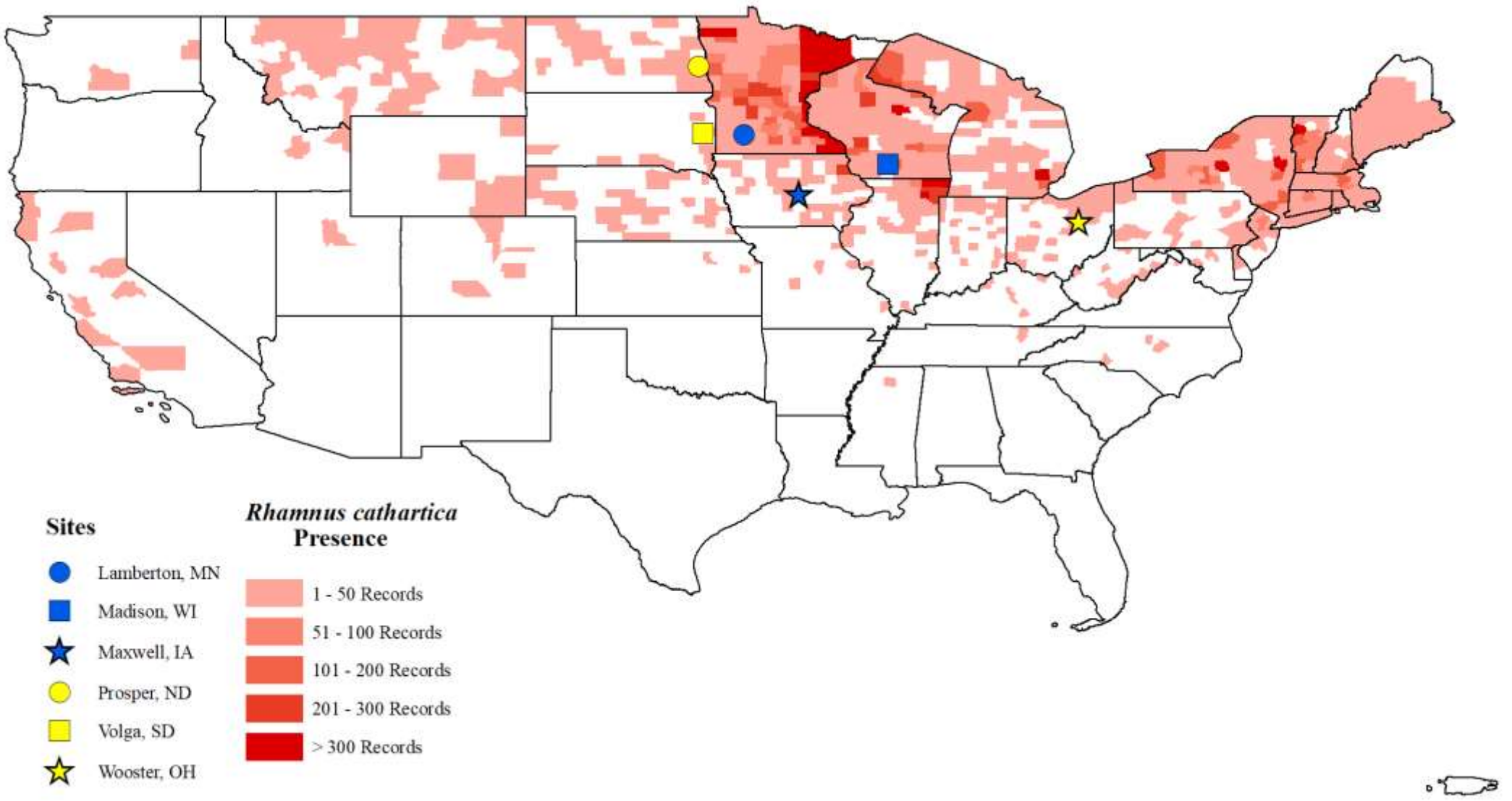


Fig. 9. Distribution of *Rhamnus cathartica* in the U.S. Presence levels of *R. cathartica* are indicated by degree of shading. Blue (Group1) and yellow (Group2) symbols indicate localities where soybean aphid samples were collected from experimental plots of *Rag* and susceptible soybean varieties. The map projection is in World Geodetic System, 1984 (WGS84) and was made using ArcGIS 10.5.

- Draft genome of *Aphis glycines* Biotype 1, a culture established in 2001, the first year subsequent to its discovery in the U.S.

- The duplicated portion of the *Ap. glycines* proteome mainly contains genes involved in apoptosis, a possible adaptation to plant chemical defenses.

- SNP based population analysis indicates China and South Korea as likely sources of the invasive U.S. soybean aphid population.

- *Ap. glycines* genetic diversity in North America has decreased over the sampled time period.

- *Ap. glycines* samples collected from *Rag* plants in Minnesota, Iowa, and Wisconsin, but not in Ohio, North Dakota, and South Dakota, show a higher frequency of specific alleles of genes associated with iron metabolism compared to aphids on susceptible plants.