Soybean aphid biotype 1 genome: Insights into the invasive biology and adaptive evolution of a major agricultural pest

Rosanna Giordano,a,b,1, Ravi Kiran Donthu,a,b,1, Aleksey V. Zimin,c, Irene Consuelo Julca Chavez,a,b,1, Toni Gabaldon,d,e,f, Manuella van Munster,g, Lawrence Hon,h, Richard Hall,i, Jonathan H. Badger,k, Minh Nguyen,l, Alejandra Flores,m, Bruce Potter,n, Tugrul Giray,o, Felipe N. Soto-Adames,p, Everett Weber,q, Jose A.P. Marcelino,q,r, Christopher J. Fields,s, David J. Voegtlín,s, Curt B. Hill,t, Glen L. Hartman,u, Soybean aphid research community (Tatsiana Akraiko,v, Andrew Aschwanden,w, Arian Avalos,x, Mark Band,y,z, Bryony Bonning,g, Anthony Bretaud.eau,b, Olga Chiesa, Chirumamilla,b, Brad S.Coates,ab, Giuseppe Cocuzza,ac, Eileen Cullen,ad, Peter Desborough,ae, Brian Diers,af, Christina DiFonzo,ag, George E. Heimpel,ah, Theresa Herman,ai, Yongping Huangaj, Janet Knodelak, Chiun-Cheng Koal, Genevieve Labrie,an, Doris Lagos-Kutz,ao, Joon-Ho Lee,ap, Seunghwan Lee,ao, Fabrice Legeai,ab, Mauro Mandrioli,ba, Gian Carlo Manni,ca, Emanuele Mazzoni,cb, Giulia Melchiori,ap, Ana Mićijević,ar, Nick Miller,as, Andi Nasuddin,at, Brian A. Nault,au, Matthew E. O’Neal,ab, Michela Panini,av, Massimo Pessino,aw, Deirdre Prischmann-Voldseth,ax, Hugh M. Robertson,ay, Sijun Liu,az, Hojun Song,ba, Kelley Tilmon,bx, John Tooker,by, Kongming Wu,bz, Shuai Zhan,bz)

a Puerto Rico Science, Technology and Research Trust, San Juan, PR, USA
b Know Your Bee, Inc. San Juan, PR, USA
c Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
d Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
e Barcelona Supercomputing Centre (BSC-CNS), Barcelona, Spain
f Institute for Research in Biomedicine, Barcelona, Spain
g Institución Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
h Color Genomics, Burlingame, CA, USA
i Pacific Biosciences, Menlo Park, CA, USA
j Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institute of Health, DCCS, Bethesda, MD, USA
k Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA
l College of Liberal Arts and Sciences, School of Molecular and Cellular Biology, University of Illinois, Urbana, IL, USA
m University of Minnesota, Southwest Research and Outreach Center, Lamberton, MN, USA
n USDA, Agricultural Research Services, Baton Rouge, LA, USA
o USDA, Agricultural Research Services, Baton Rouge, LA, USA
p USDA, Agricultural Research Services, Baton Rouge, LA, USA
q USDA-ARS and Department of Crop Sciences, University of Illinois, Urbana, IL, USA
r Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA
s Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA
t Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign, IL, USA
u Department of Entomology, Pennsylvania State University, University Park, PA, USA
v Department of Entomology, Pennsylvania State University, University Park, PA, USA
w USDA, Agriculture Research Services, Baton Rouge, LA, USA
x USDA, Agriculture Research Services, Baton Rouge, LA, USA
y Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign, IL, USA
z Dipartimento di Scienze delle Produzioni Vegetali Sostenibili, Università Cattolica del Sacro Cuore, Piacenza, Italy
aa Department of Entomology, North Dakota State University, Fargo, ND, USA
ab Department of Entomology, Iowa State University, Ames, IA, USA
ac INRA, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
ad Agricen Sciences, Pilot Point, TX, USA
ae USDA-ARS and Department of Crop Sciences, University of Illinois, Urbana, IL, USA
af Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA
ag Department of Entomology and Nematology, University of Florida, Gainesville, FL, USA
ah Department of Entomology, Pennsylvania State University, University Park, PA, USA
ai Department of Entomology, Pennsylvania State University, University Park, PA, USA
aj USDA, Agriculture Research Services, Baton Rouge, LA, USA
ak Department of Entomology, Pennsylvania State University, University Park, PA, USA
al USDA, Agriculture Research Services, Baton Rouge, LA, USA
am Genevieve Labrie,am, Doris Lagos-Kutz,am, Joon-Ho Lee,am, Seunghwan Lee,am, Fabrice Legeai,am, Mauro Mandrioli,an, Gian Carlo Manni,an, Emanuele Mazzoni,an, Giulia Melchiori,an, Ana Mićijević,an, Nick Miller,an, Andi Nasuddin,an, Brian A. Nault,an, Matthew E. O’Neal,an, Michela Panini,an, Massimo Pessino,an, Deirdre Prischmann-Voldseth,an, Hugh M. Robertson,an, Sijun Liu,an, Hojun Song,an, Kelley Tilmon,an, John Tooker,an, Kongming Wu,an, Shuai Zhan,an)

* Corresponding author. Puerto Rico Science, Technology and Research Trust, San Juan, PR, USA.
** Corresponding author. Puerto Rico Science, Technology and Research Trust, San Juan, PR, USA.

E-mail addresses: rgiordano@prscientestrust.org, rgiordano500@gmail.com (R. Giordano), rkiran@prscientestrust.org, donthuanalyst@gmail.com (R.K. Donthu).

1 These authors contributed equally to this work.


Academic Editor: S. Brian Lovell
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 *A. 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 *A. gossypii*, its closest related species, has the lowest number of these duplicated genes. An Illumina GoldenGate assay of 2380 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 *Fst* 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 2500 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, *A. glycines* has a life cycle during which both sexual and asexuall morphs are produced (holocyclic) on alternating plant hosts (heteroecous). *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., 2004) 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 produce the sexual generation (*gy-noparae*). 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).
Large aphid populations reduce soybean production (Henderson et al., 2012; Hodgson et al., 2011; Nielsenn and Hajek, 2005; Rutledge and O’Neil, 2005; Wu 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; Pedigo and Ragsdale, 2011). 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.ksueu.issu/aphids/report/html/citations.html]). 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 population. 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, myrmeleontids, 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 $24 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 Raddifie, 2008), and bean, Phaseolus spp., (Mueller et al., 2010).

While there have been efforts to establish environmentally sound biological controls methods (Chacin 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, 2011; Hodgson et al., 2012; Koch et al., 2016). 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...
# Table 1

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genbank Accession</td>
<td>NA*</td>
<td>NA*</td>
<td>GCF_0004010815.1</td>
<td>GCF_0001856785.1</td>
<td>NA*</td>
<td>GCF_0001429815.2</td>
<td>GCF_0001186385.1</td>
<td>GCF_0002803265.2</td>
<td>GCA_003676215.3</td>
<td>NA*</td>
<td>GCA_003264975.1</td>
<td>GCF_000268045.1</td>
</tr>
<tr>
<td># Scaffolds Genome (Scaffolds) Size Mb</td>
<td>3224</td>
<td>8397</td>
<td>4718</td>
<td>4021</td>
<td>49,286</td>
<td>3,925</td>
<td>5637</td>
<td>1347</td>
<td>220</td>
<td>15,587</td>
<td>7859</td>
<td>1923</td>
</tr>
<tr>
<td>Longest Scaffold Mb</td>
<td>23.00</td>
<td>1.00</td>
<td>5.00</td>
<td>2.00</td>
<td>959</td>
<td>1.00</td>
<td>200</td>
<td>2.00</td>
<td>928</td>
<td>1662</td>
<td>1096</td>
<td>1001</td>
</tr>
<tr>
<td>Shortest Scaffold nt</td>
<td>60</td>
<td>200</td>
<td>889</td>
<td>959</td>
<td>1001</td>
<td>200</td>
<td>26.00</td>
<td>94.00</td>
<td>6.22</td>
<td>10.00</td>
<td>1004</td>
<td></td>
</tr>
<tr>
<td># Scaffolds &gt; 500 nt</td>
<td>3209</td>
<td>8397</td>
<td>4718 (100.0%)</td>
<td>4021 (100.0%)</td>
<td>49,286 (100.0%)</td>
<td>23,451 (98.0%)</td>
<td>5637 (100.0%)</td>
<td>1347 (100.0%)</td>
<td>220 (100.0%)</td>
<td>15,587 (100.0%)</td>
<td>7859 (100.0%)</td>
<td>1923 (100.0%)</td>
</tr>
<tr>
<td># Scaffolds &gt; 1K nt</td>
<td>3208</td>
<td>8397</td>
<td>4487 (95.1%)</td>
<td>4017 (99.9%)</td>
<td>49,286 (100.0%)</td>
<td>12,914 (54.0%)</td>
<td>5613 (99.6%)</td>
<td>1347 (100.0%)</td>
<td>220 (100.0%)</td>
<td>15,587 (100.0%)</td>
<td>7859 (100.0%)</td>
<td>1923 (100.0%)</td>
</tr>
<tr>
<td># Scaffolds &gt; 10K nt</td>
<td>410</td>
<td>2716</td>
<td>1574 (33.4%)</td>
<td>1845 (45.9%)</td>
<td>9745 (19.8%)</td>
<td>2355 (9.8%)</td>
<td>2941 (52.2%)</td>
<td>808 (60.0%)</td>
<td>155 (70.5%)</td>
<td>3032 (24.6%)</td>
<td>2425 (30.9%)</td>
<td>860 (44.7%)</td>
</tr>
<tr>
<td># Scaffolds &gt; 100K nt</td>
<td>121</td>
<td>968</td>
<td>683 (14.5%)</td>
<td>780 (19.6%)</td>
<td>170 (4.0%)</td>
<td>1106 (4.6%)</td>
<td>902 (12.0%)</td>
<td>161 (12.0%)</td>
<td>8 (3.6%)</td>
<td>940 (6.0%)</td>
<td>325 (4.1%)</td>
<td>318 (16.5%)</td>
</tr>
<tr>
<td># Scaffolds &gt; 1M nt</td>
<td>55</td>
<td>1 (0.0%)</td>
<td>33 (0.7%)</td>
<td>38 (0.9%)</td>
<td>0 (0.0%)</td>
<td>34 (0.6%)</td>
<td>78 (5.8%)</td>
<td>4 (1.8%)</td>
<td>0 (0.0%)</td>
<td>93 (1.2%)</td>
<td>122 (6.3%)</td>
<td></td>
</tr>
<tr>
<td>Mean Scaffold size Kb</td>
<td>95</td>
<td>36</td>
<td>62</td>
<td>86</td>
<td>8</td>
<td>22</td>
<td>70</td>
<td>223</td>
<td>1481</td>
<td>20</td>
<td>49</td>
<td>183</td>
</tr>
<tr>
<td>Median Scaffold size Kb</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>12</td>
<td>20</td>
<td>3</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>NS0 Scaffold Length Mb</td>
<td>6.00</td>
<td>0.10</td>
<td>0.44</td>
<td>0.44</td>
<td>0.02</td>
<td>0.50</td>
<td>0.40</td>
<td>3.00</td>
<td>93.00</td>
<td>0.12</td>
<td>1.29</td>
<td>1.68</td>
</tr>
<tr>
<td>L50 Scaffold Count</td>
<td>15</td>
<td>512</td>
<td>195</td>
<td>224</td>
<td>4472</td>
<td>280</td>
<td>281</td>
<td>25</td>
<td>2</td>
<td>782</td>
<td>71</td>
<td>67</td>
</tr>
<tr>
<td>Scaffold %A</td>
<td>35.77</td>
<td>36.08</td>
<td>34.47</td>
<td>34.82</td>
<td>35.04</td>
<td>32.41</td>
<td>26.57</td>
<td>36.18</td>
<td>36.15</td>
<td>36.09</td>
<td>33.71</td>
<td>34.45</td>
</tr>
<tr>
<td>Scaffold %G</td>
<td>35.73</td>
<td>36.03</td>
<td>34.31</td>
<td>34.78</td>
<td>35.05</td>
<td>32.41</td>
<td>26.57</td>
<td>36.15</td>
<td>36.12</td>
<td>36.13</td>
<td>33.71</td>
<td>34.46</td>
</tr>
<tr>
<td>Scaffold %T</td>
<td>1.68</td>
<td>0.14</td>
<td>5.44</td>
<td>0.53</td>
<td>0.05</td>
<td>7.71</td>
<td>24.94</td>
<td>1.17</td>
<td>0.01</td>
<td>0.02</td>
<td>6.63</td>
<td>1.49</td>
</tr>
<tr>
<td>Scaffold N Mb</td>
<td>5.18</td>
<td>0.42</td>
<td>16</td>
<td>1.94</td>
<td>0.2</td>
<td>41.78</td>
<td>98.53</td>
<td>3.52</td>
<td>0.05</td>
<td>25.52</td>
<td>5.26</td>
<td></td>
</tr>
<tr>
<td>% Assembly in Scaffolds Contigs</td>
<td>0.831</td>
<td>0.267</td>
<td>0.959</td>
<td>0.761</td>
<td>0.196</td>
<td>0.951</td>
<td>0.99</td>
<td>0.915</td>
<td>0.984</td>
<td>0.224</td>
<td>0.842</td>
<td>0.924</td>
</tr>
<tr>
<td>% Assembly in Unassembled Contigs</td>
<td>0.169</td>
<td>0.733</td>
<td>0.041</td>
<td>0.239</td>
<td>0.804</td>
<td>0.049</td>
<td>0.01</td>
<td>0.085</td>
<td>0.016</td>
<td>0.776</td>
<td>0.158</td>
<td>0.076</td>
</tr>
<tr>
<td>Average Length of Ns Between Contigs</td>
<td>14284</td>
<td>316</td>
<td>2152</td>
<td>941</td>
<td>93</td>
<td>1139</td>
<td>2185</td>
<td>3785</td>
<td>100</td>
<td>99</td>
<td>4839</td>
<td>3132</td>
</tr>
<tr>
<td># Contigs</td>
<td>3587</td>
<td>9610</td>
<td>12,144</td>
<td>5971</td>
<td>51,353</td>
<td>60,594</td>
<td>50,723</td>
<td>2276</td>
<td>689</td>
<td>16,133</td>
<td>13,128</td>
<td>3599</td>
</tr>
<tr>
<td>Genome Size Mb</td>
<td>2023</td>
<td>9,082</td>
<td>7,244</td>
<td>1,723</td>
<td>3,205</td>
<td>9,624</td>
<td>689</td>
<td>16,133</td>
<td>13,128</td>
<td>3599</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longest Contig Mb</td>
<td>7.7</td>
<td>0.88</td>
<td>0.71</td>
<td>1.5</td>
<td>0.21</td>
<td>0.42</td>
<td>0.17</td>
<td>2.4</td>
<td>42.51</td>
<td>0.57</td>
<td>0.78</td>
<td>2</td>
</tr>
<tr>
<td>Shortest Contig</td>
<td>60</td>
<td>0</td>
<td>415</td>
<td>1</td>
<td>1001</td>
<td>200</td>
<td>60</td>
<td>81</td>
<td>1096</td>
<td>1001</td>
<td>48</td>
<td>146</td>
</tr>
</tbody>
</table>
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 (Voegtltn 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, *Rag* (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’Neill, R. and Voegtltn, D.J., Personal communication). Previous observations and laboratory tests had shown that the *Ap. glycines* gyno-parae (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 (Voegtltn 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 *Rag*+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 *Rag*+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. gossypii* and *Ap. trilochinius* 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). 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://biapa.genouest.org/is/aphidbase/), *Acrystosiphon 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://biapa.genouest.org/is/aphidbase/), *Schizaphis graminum*, and *Sipha stava* (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 the recently-sequenced genomes of the cedar aphid *Cinara cedri* (Julca et al., in press) and of the phyloxeran *Daktulosphaira vignae* (Rüpe 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; Garletto 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 answer to the differences in evolution; 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& domain (Cornman and Willis, 2009).

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 × 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 ~100 μl 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 5 μl of 20 mg/ml of Proteinase K (−20 °C) 2). 30 μl of 3M KAc was added to the tubes, mixed gently, and placed on ice for 30 min. Tubes were centrifuged in a refrigerated microfuge for 10 min after which the supernatant was removed. 3) An equal volume (500 μl) of Tris equilibrated phenol (ChCl3:Phenol) was added and the tubes mixed by hand. Tubes were then spun for 5 min at room temperature. The upper aqueous phase (475 μl) 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 min at RT, the aqueous phase retrieved and placed into new tubes. 5) 1 μl of 32 mg/ml of RNaseA (~−20°C) (Sigma R4642) was added to tubes, which were mixed and incubated at 37 °C for 15 min. 6) 100-95% ethanol, in a volume of two times the amount of supernatant, (700–800 μl) 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 1 ml of ethanol and stored at −20 °C. 8) Tubes were spun in refrigerated centrifuge for 5 min then dried in an incubator at 39 °C while not allowing the DNA to get overly dry to facilitate re-suspension. 9) 20 μl 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 20 μg of total RNA using Oligotex (Qiangen, 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 adapted cDNA was amplified for 10 cycles and normalized with the Trimmer Direct kit (EvoGen, 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 1 kb DNA ladder (Thermo Fisher).

RNaseq libraries were constructed using the TruSeq RNA Sample Preparation Kit (Illumina, CA). Briefly, messenger RNA was selected from 1 μg 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 10 nM. The individually barcoded libraries were pooled in equimolar concentration. The pooled libraries were further quantified 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 100 nt. The raw.fastq 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 180 μl of lysis solution and 5 μl 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 s 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 1 μl of RNAase (R-4642 Sigma-Aldrich) ~24 mg/ml. Tubes were briefly vortexed and incubated at room temperature (25 °C) for 10 min. Tubes were centrifuged for 30 s 200 μl of buffer AL was added and tubes mixed briefly by pulse-vortex. Tubes were incubated at 70 °C for 5 min to dissolve precipitate, vortexed briefly at low speed, and incubated for an additional 3 min or until all precipitate was dissolved. Tubes were spun briefly at low speed to bring down condensation on the inside of the caps, and cooled for 5 to 10 min. 200 μl of cold (-20 °C) ethanol (96-100%) was added and tubes vortexed briefly after which they were placed at 4 °C overnight to allow DNA to precipitate. Tubes were briefly centrifuged and the entire lysate transferred to Promega columns (Wizard SV Minicolumns Part# A129B) without wetting the rim, and centrifuged at 8000 rpm for 1 min. The flow through was discarded and the column membrane washed with 500 μl Buffer AW1, centrifuged at 8000 rpm for 1 min and rewarshed again with 500 μl Buffer AW2 and centrifuged at 8000 rpm for 1 min. A final centrifuge step at 12,000 rpm for 3 min was used to dry the membrane completely. The column was then placed in a clean, labeled, 1.5 ml Eppendorf tube and 50 μl of Sigma tissue culture water was added to the center of the membrane and allowed to saturate the membrane for 3 min. Membrane was centrifuged at 12,000 rpm for 3 min to elute the DNA. Tubes with eluted DNA were incubated in a heat block at 60 °C for ~1/2 h, to insure that all residual ethanol from the wash buffers evaporated which reduced the volume in tube to 30 μl ± 3 μl. Tubes were vortexed gently and spun down briefly. DNA was measured using a Qubit Fluorometer (Thermo Fisher Scientific, U.S.). As aphids used differed in size the DNA obtained with the above protocol ranged from ~230 to 650 ng of total DNA from a single aphid. Aphid specimens resulting in a concentration of 300-400 ng in a 7 μl volume were chosen for the downstream steps. DNA resulting in a concentration of 300-400 ng (~395 ng) in a 7-30 μl volume, was placed in individual wells of a 96 well plate. The plates were sealed and run in a SpinVac to dry without heat for 1 h. Plates were checked to confirm if dry, if not, the procedure was repeated for another 15 min 7 μl of water was added to wells in plated, covered with film and the DNA allowed to re-suspend overnight at 4 °C. If the plate was not run right away it was stored a -20 °C.

2.5. Sequencing of genome

An Illumina HiSeq 2000 and 454 Titanium system was used to generate Illumina and 454 sequences (NCBI SRA accessions: PRJNAS51277). Two types of libraries were prepared and sequenced with 454 Titanium platform: 1) random shotgun, in which genomic DNA was randomly sheared to a size of 600 nt to 1.2 kb and 2) paired-end, in which DNA was sheared to a size of 8 kb and 20 kb fragments. On Illumina HiSeq 2000 system, the shotgun library, with a fragment size of 200 bp, were sequenced from both ends (paired-end sequencing), each read being 100 nt in length. Mate-pair libraries with a jump size of 3 kb and 8 kb were sequenced at 35 nt from each end of the fragments. Using Pacific Biosciences (PacBio) RSII sequencing platform with C2 chemistry, we sequenced a 10K library on 8 SMRT cells which yielded a total of 193,586 sequences that passed quality filters (NCBI SRA accessions: PRJNAS51277). Mean length of these sequences was 4274 bases. Total number of bases in all the 193,586 sequences was 1,299, 749, 757.

2.6. Genome sequence assembly

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 3 kb and 10 kb target insert sizes. The 454 sequencing data contained mate pairs with target insert size of 8 Kb. 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 MGene 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 300 ng of high molecular weight DNA (HMW).

Using the Knickers software (v1.5.5), we determined that the best nicking enzyme for this genome was BsSI (New England BioLabs), with a labeling density of approximately 16 nicks per 100 kb (http://www. bmxinstall.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 h at 37 °C. This was subsequently combined with the labeling master mix and incubated for 1 h at 72 °C. A repair master mix was then added for 0.5 h 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, v8, 2016). The raw data output of 221.9 GB obtained from these scans was analyzed using IrysView software (v.2.5.1) and the protocol given in “IrysView v2.5.1 Software Training Guide” (Document # 30035, v.G, 2016). The filtered data output consisted of 1826 Gb.

Using the BioNano Genomics assembly pipeline, genomic maps of...
DNA molecules in bnx format were aligned against each other and assembles into BioNano Genome map contigs. There were 665 BioNano Genome Map (BNM) 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 Refaligener, 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 3261 to 3254 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 blob-plot.

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 Trimomatic (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 −PARALLELSTATS. Illumina reads thus normalized were assembled using Trinity version 2.1.1 (Grabherr et al., 2011) in genome guided mode with parameters −geneid 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 Trimomatic (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 3.0.1.1 (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 GO06 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 (Stanle 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-asserbler.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_optscf file, est2genome and protein2genome was set to 0. “snap hmm” 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 *A. 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 *A. 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 *A. 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 *A. glycines* B1 genome, CutProtFam annotation site (http://aiais.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 *A. 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 homologs 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 v1.3 (Capella-Gutierrez 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 downloading 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: *A. 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 viridifolia*. 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 (pfam 00379) was used for phylogenetic analyses, because they tend to be highly divergent and difficult to align along their full length. RR-2 proteins from *A. glycines*, *M. persicae*, *A. pisum*, *D. noxia*, *R. padi* and *D. viridifolia* 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: 15%.

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 (*A. 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 v8.11.17 (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 data from *A. 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 -so
leca1.3-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 with conditions of each bead type unique in order to assure the 2 sample basis 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 (Gundersen 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 250 ng, 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 3072 SNPs with best designability scores were selected for genotyping a total of 4421 samples collected from Australia, Canada, China, Indonesia, Japan, Myanmar, South Korea, Taiwan, Thailand and USA. Using Illumina genome studio, genotype clusters for all 3072 SNPs were manually examined and edited. Of 4421 samples, 212 were excluded because the call rate was less than 95% and 418 were excluded because they were lab culture samples. Of 3072 SNP clusters, 637 SNP genotype clusters were manually flagged as being poor quality and removed from the analysis. Of the remaining 2435 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 2380 SNP genotypes in 3791 samples (Table 2) that was used in the downstream analysis.

2.20. Annotation of genes overlapping SNPs

There were 1700 genes that overlapped with 2380 SNPs. Of the genes found to overlap, GO terms were obtained for 1185 genes, Eukaryotic Orthologous Groups (KOG) categories were obtained for 1025 genes, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway names were identified for 641 genes. To obtain KOG categories, RPSBLAST 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/). 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.

Table 2

<table>
<thead>
<tr>
<th>Location</th>
<th>Year of Samples</th>
<th>Year</th>
<th># of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guangxi</td>
<td>2008</td>
<td>10</td>
<td>167</td>
</tr>
<tr>
<td>Hebei</td>
<td>2008</td>
<td>7</td>
<td>2003</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>11</td>
<td>2011</td>
</tr>
<tr>
<td>Hei long jiang</td>
<td>2001</td>
<td>12</td>
<td>2004</td>
</tr>
<tr>
<td>Hubei</td>
<td>2007</td>
<td>10</td>
<td>2012</td>
</tr>
<tr>
<td>Jiangsu</td>
<td>2010</td>
<td>24</td>
<td>2647</td>
</tr>
<tr>
<td>Jilin</td>
<td>2001</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Shansi</td>
<td>2008</td>
<td>12</td>
<td>2012</td>
</tr>
<tr>
<td>Zhejiang</td>
<td>2008</td>
<td>11</td>
<td>2001</td>
</tr>
<tr>
<td>Indonesia</td>
<td>2012</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>Cianjur</td>
<td>2013</td>
<td>57</td>
<td>2008</td>
</tr>
<tr>
<td>Lombok</td>
<td>2010</td>
<td>5</td>
<td>2009</td>
</tr>
<tr>
<td>Malang</td>
<td>2013</td>
<td>10</td>
<td>2010</td>
</tr>
<tr>
<td>Maros</td>
<td>2012</td>
<td>15</td>
<td>2011</td>
</tr>
<tr>
<td>Sakabumi</td>
<td>2013</td>
<td>1</td>
<td>2001</td>
</tr>
<tr>
<td>Japan</td>
<td>2008</td>
<td>24</td>
<td>2012</td>
</tr>
<tr>
<td>Aomori</td>
<td>2009</td>
<td>9</td>
<td>2006</td>
</tr>
<tr>
<td>Furukawa</td>
<td>2001</td>
<td>11</td>
<td>2005</td>
</tr>
<tr>
<td>Ibaraki</td>
<td>2001</td>
<td>12</td>
<td>2009</td>
</tr>
<tr>
<td>Iwate</td>
<td>2008</td>
<td>5</td>
<td>2010</td>
</tr>
<tr>
<td>Morika</td>
<td>2001</td>
<td>12</td>
<td>2011</td>
</tr>
<tr>
<td>Naganoo</td>
<td>2010</td>
<td>24</td>
<td>2013</td>
</tr>
<tr>
<td>Shimane</td>
<td>2010</td>
<td>6</td>
<td>2001</td>
</tr>
<tr>
<td>Tochigi</td>
<td>2001</td>
<td>12</td>
<td>2009</td>
</tr>
<tr>
<td>Yamagata</td>
<td>2008</td>
<td>12</td>
<td>2011</td>
</tr>
<tr>
<td>Yamaguchi</td>
<td>2008</td>
<td>11</td>
<td>2012</td>
</tr>
<tr>
<td>Myanmar</td>
<td>2010</td>
<td>24</td>
<td>2013</td>
</tr>
<tr>
<td>Shan</td>
<td>2013</td>
<td>48</td>
<td>2013</td>
</tr>
<tr>
<td>South Korea</td>
<td>2012</td>
<td>50</td>
<td>2001</td>
</tr>
<tr>
<td>Asan</td>
<td>2012</td>
<td>12</td>
<td>2010</td>
</tr>
<tr>
<td>Cheonan</td>
<td>2011</td>
<td>14</td>
<td>2011</td>
</tr>
<tr>
<td>Muan</td>
<td>2012</td>
<td>12</td>
<td>2008</td>
</tr>
<tr>
<td>Sungun</td>
<td>2011</td>
<td>12</td>
<td>2009</td>
</tr>
<tr>
<td>Taiwan</td>
<td>2003</td>
<td>6</td>
<td>2012</td>
</tr>
<tr>
<td>Kao-Itsing</td>
<td>2011</td>
<td>12</td>
<td>2013</td>
</tr>
<tr>
<td>Thailand</td>
<td>2011</td>
<td>17</td>
<td>2009</td>
</tr>
<tr>
<td>Australia</td>
<td>2004</td>
<td>7</td>
<td>2010</td>
</tr>
<tr>
<td>NSW</td>
<td>2012</td>
<td>24</td>
<td>2011</td>
</tr>
<tr>
<td>QLD</td>
<td>2004</td>
<td>7</td>
<td>2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>109*</td>
</tr>
</tbody>
</table>

Abbreviations are: A, Australia; EUR, Europe; J, Japan; K, Korea; L, Latin America; N, North America; S, South America; U, USA; W, World; I, Indonesia; M, Malaysia; T, Taiwan; T, Thailand; B, Bangladesh; R, Russia; A, Argentina; C, Canada; D, Denmark; F, France; G, Germany; I, India; N, Netherlands; P, Pakistan; S, South Africa; T, Turkey; U, United Arab Emirates; U, United Kingdom; V, Vietnam; W, Australia.
2.21. Assessment and management of ascertainment bias

Our SNP discovery process is based on the alignment of sequence reads from U.S. samples against the reference genome of *U. p. glycinum*. 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 ascertainment bias, when comparing U.S./Canada and Asian/Australian populations, we resolved to restrict our analysis to the use of 926 SNPs that fell 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 2380 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 FST values

VCF tools version 0.1.15 (Danec et al., 2011) was used to calculate FST values according to the method described in Weir and Cockerham (1984). VCF file with 3791 samples and 2380 SNPs was given as input to the VCFtools using –weir-fst-pop option for each population in the pairwise comparison. FST values were calculated for all pairwise comparisons between all populations sampled: Australia, China, Japan, South Korea, Indonesia, Taiwan, Thailand, Myanmar, Canada and U.S. FST 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, FST 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 FST 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 FST value to the y-axis.

2.26. Heat maps

Comma delimited files with FST values were imported into R using the read_csv function. Heat maps were generated on the imported FST values using heatmap function of R package pheatmap (Kolde, 2015).

2.27. Over representation analysis

Over representation analysis was performed for multiple sets of genes that overlap with SNPs with FST values 1) >0.14 in a comparison between U.S. samples collected in 2001 and 2005 2) >0.1 in a comparison between U.S. samples collected in 2001 and 2009, 2010, 2011, and 2012 3) >0.2 in comparison between Rag (Rag1; Rag1+2; Rag2) and susceptible aphids. To identify the GO terms or KEGG pathways overrepresented among these two sets of genes, hypergeometric analysis was performed using the GOStats package (Falcon and Gentleman, 2007). The genes that overlapped with the 2380 SNPs used in this study were considered as “universe”. The readtable function was used to import input files into R. For the GO terms over representation analysis, GOALLFrame and GeneSetCollection data objects were created using GOALLFrame and GeneSetCollection functions of GSEABase package (Morgan et al., 2019). The GSEAGOHyperGParams and hyperGTest functions were used to perform hypergeometric test on GO terms, while GSEAKEGGHyperGParams and hyperGTest functions were used to perform hypergeometric test on KEGG pathway terms.

2.28. Identification of non-synonymous SNPs

To identify the non-synonymous SNPs among the 2380 SNPs, Ensembl Variant Effect Predictor (McLaren et al., 2016) was run on an input file with 2380 SNPs in VCF format using parameter “-f” along with the gene annotation file in GFF format with parameter “-gff” and the genome sequence in FASTA format using parameter “-fasta”.

2.29. Data availability

The genome sequence assembly scaffolds, gene annotation and functional annotation files are available at AphidBase (https://bipaa.genouest.org/is/aphidbase/). The genome sequence assembly and gene annotation was also deposited at NCBI GenBank under the accession VTZ001000000, GenBank assembly accession GCA_009761285.1; BioProject PRJNA551277; BioSample SAMN12143004. The raw sequence data was deposited at NCBI SRA database under accession PRJNA551277. The SNP genotype data was deposited at the European
Variation Archive under project PRJEB35243 and analyses ERZ1108186 (https://www.ebi.ac.uk/ena/data/view/PRJEB35243).

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, 3224 scaffolds and a N50 value of 6 Mbp making it the 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 Sapropilaraceae. 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-Marzó, 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 Aphids 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 Eastop, 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 (phylomed ID: 75890, http://www.phylomed.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 1848 are present in each species. Of these only 811 have single-copy orthologs present in all species (Fig. 2, Table S4). When Hemiptera 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 3972 soybean aphid proteins (20.4%) of the proteome) that have paralogs. These genes considered as in-paralogs can be assigned to 1028 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., 2018).

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 aphid and species-specific duplications (Mathers et al., 2017; Duncan et al., 2016; Huerta-Cepas et al., 2018).

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: 935–1589); 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: 56–6426); 9) Remaining genes not included in the previous categories. The genomic resources for C. cedri and D. vitifolii are not publicly accessible, and were kindly made available prior to publication by Toni Gabaldón and Denis Tagu.
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 *A. 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 *A. glycines*. There are 1621 (41%) *A. glycines* genes that are involved in 1028 gene expansion events, of these 372 occur in at least one other aphid species (Table S5). Unexpectedly, *A. 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 *A. 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 signaling 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 *A. glycines* and its most closely related species *A. gossypii*, we calculated the dN/dS ratios of 7502 single-copy orthologs of *A. glycines* and *A. gossypii* using *M. persicae* GO06 as the outgroup. Of these orthologs, 3825 passed the cut off filters (see Materials and Methods). Most of the genes (~90%) 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 3825 single copy orthologs, six proteins were identified as under positive selection in both *A. glycines* and *A. 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 *A. 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 *A. gossypii*, 24 have known functional annotations. Genes under this category also cover a wide variety of metabolic functions from Azurocidin, an antimicrobial protein, to optomotor-blind protein required for optic lobes and wing development, to the sodium channel protein Nach, in-volved in the clearance of tracheal liquid.

### 3.3. Cuticular proteins

The manual curation and annotation of *A. 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 representative 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 *A. 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).
Fig. 4. Population structure analysis of the SBA world wide geographic distribution, Asia, Australia and North America, using 926 SNP’s 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 (a); (d) Enlargement of Australian and Indonesian populations indicated in rectangle in (c). (e) Fst 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 Fst level; (f) Neighbor Joining tree for all populations generated using Fst 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 used in the PCA plots.

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 2380 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 Fst values for comparisons between 2001 and each year of sample collection for U.S. and Canada.
complexity signals the absence of specific duplication trends for this protein subfamily (Fig. S5A). An ortholog of Stylin 01, originally identified in *A. pism* and *M. persicae* was also found in *Ap. glycines* (AG6029153) (grey box, Fig. S5A). This RR-1 protein present at the tip of aphid stylenes 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; Glikow et al., 2008). Indeed Stylin 01, named previously Mpp4 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* (Corman 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 stylenes (Webster et al., 2017). However, it was not possible to precisely identify one specific candidate. Recently, Kamangar et al. (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 Mpp000169000 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 2380 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.

---

**Fig. 6.** Heatmap of *F*<sub>st</sub> values calculated by comparing the allele frequencies for all 2380 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. S10), 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*<sub>st</sub> value as represented in the scale bar on the top right corner.

R&R Consensus domain present in CPRs confer chitin-binding properties to these proteins and is involved in cuticle formation (Rebers and Riddiford, 1989). 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 cysteine 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
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. 4c and d).

The results and interpretations derived from the PGA analysis are in concordance with those derived from the pairwise $F_{ST}$ values calculated for all countries (Fig. 4e). The lowest $F_{ST}$ values were observed between the U.S. and Canada and these form a cluster in the PCA plot (Fig. 4a, 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. glycin* 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. glycin* 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 2380 SNPs (Fig. S6).

The same relationship between populations from different countries were seen using $F_{st}$ 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 2380 (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 2001 was 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 2380 SNP data set was utilized for subsequent analyses that pertained to this population. PCA plots generated using the total number of 2380 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 observed 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 to 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. 5). While apterous aphids move very short distances, historically it has been thought that aphid flight is common (Close and Tomlinson, 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 (Loudale 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 astronomical 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 2380 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 *A. 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. *A. glycines* reproduces clonally in the summer months and all samples tested were asexual 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, and 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
Fig. 9. Distribution of Rhamnus cathartica in the U.S. Presence levels of R. cathartica are indicated by degree of shading. Blue (Group 1) and yellow (Group 2) 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 and Early Detection and Distribution Mapping System 2020. The University of Georgia - Center for Invasive Species and Ecosystem Health http://www.edmaps.org/.

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 hydrolyase, 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 experimental 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 2003–2010 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 4
List of enriched Gene Ontology (GO) terms for genes overlapping with SNPs having $F_{GT}$ values greater than or equal to 0.1 for the comparison between Rag and susceptible plant varieties for WI, 2010; IA and MN 2013. P-values are from overrepresentation analysis. GO class abbreviations: BP= Biological Processes, CC = Cellular Components, MF = Molecular Function.

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

(Table S13) for genes overlapping with SNPs having $F_{GT}$ 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 aconitate (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 aphis 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 *A. glycines*, that have less than three percent in parallel duplication in *A. 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 *A. 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 an 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


in virus-insect interactions. Insect Sci. 24 (6), 990–1002.