ABSTRACT Soybean aphid [Aphis glycines Matsumura (Hemiptera: Aphididae)] is the most damaging insect pest of soybean [Glycine max (L.) Merr.] in the Upper Midwest of the United States and is primarily controlled by insecticides. Soybean aphid resistance [i.e., Rag genes] has been documented in some soybean accessions but more sources of resistance are needed. Incorporation of the resistance into marketed varieties has also been slow. Genome-wide association mapping can aid in identifying resistant accessions by correlating phenotypic data with single nucleotide polymorphisms (SNPs) across a genome. Aphid population measures from 2366 soybean accesses were collected from published studies screening cultivated soybean (G. max) and wild soybean (Glycine soja Siebold & Zucc.) with aphids exhibiting Biotype 1, 2, or 3 characteristics. Genotypic data were obtained from the SoySNP50K high-density genotyping array previously used to genotype the USDA Soybean Germplasm Collection. Significant associations between SNPs and soybean aphid counts were found on 18 of the 20 soybean chromosomes. Significant SNPs were found on chromosomes 7, 8, 13, and 16 with known Rag genes. SNPs were also significant on chromosomes 1, 2, 4 to 6, 9 to 12, 14, and 17 to 20 where Rag genes have not yet been mapped, suggesting that many Rag genes remain to be discovered. These SNPs can be used to determine accessions that are likely to have novel aphid resistance traits of value for breeding programs.

SOYBEAN is an important field crop in the United States with 36,207,629 ha harvested in 2017 (USDA-NASS, 2017). Soybean aphid, A. glycines, is a damaging pest of soybean, especially in the upper Midwest of the United States, which reduces yield by directly removing photosynthate and inhibiting photosynthesis (Macedo et al., 2003; Ragsdale et al., 2004). In addition, soybean aphid can vector plant viruses, provide a substrate for sooty mold growth by secreting honeydew, and facilitate soybean cyst nematode infection (Ragsdale et al., 2011; Tilmont et al., 2011; McCarville et al., 2014a). Injury caused by soybean aphid infestations can cost soybean growers $2.4 billion annually (Song et al., 2006). Threshold-based applications of broad-spectrum foliar insecticides are the primary tactic used to manage

Abbreviations: GWAS, genome-wide association study; MAF, minor allele frequency; MG, maturity group; QTL, quantitative trait locus; SNP, single nucleotide polymorphism
soybean aphid (Ragsdale et al., 2011; Hodgson et al., 2012). However, dependence on broad-spectrum insecticides can result in the development of pest resistance to insecticides (Hanson et al., 2017), outbreaks of secondary pests, resurgence of the target pest, and environmental contamination (Pedigo and Rice, 2009).

Host-plant resistance is a management strategy under development for the soybean aphid to reduce the likelihood of aphid populations causing economically significant yield loss (Hill et al., 2012; Hesler et al., 2013). Host-plant resistance is a heritable decrease in plant susceptibility to pests (Painter, 1951; Smith, 2005). Resistance of plants to insect pests can be divided into three categories that may act independently or in conjunction (Smith, 2005). More specifically, resistant plants can impact pest developmental time, survival, or fecundity (i.e., antibiosis) or behavioral avoidance such as reduced oviposition and colonization (i.e., antixenosis) (Painter, 1951; Li et al., 2004; Smith, 2005). In addition, some resistant plants can withstand large pest populations without experiencing economic damage (i.e., tolerance) (Smith, 2005). Soybean aphid host-plant resistance was first documented in PI548633 (i.e., 'Dowling') and PI548657 (i.e., 'Jackson') (Hill et al., 2004). Known sources of soybean aphid resistance have most often been categorized as involving antibiosis, antixenosis, or both (e.g., Diaz-Montano et al., 2006; Hesler and Dashiel, 2011; Enders et al., 2014), but tolerance has also been documented in select cases (Pierson et al., 2010; Prochaska et al., 2013; Marchi-Werle et al., 2014).

Of the more than 3500 soybean accessions that have been screened for resistance to soybean aphid, at least 39 accessions have been found to exhibit resistance (Cooper et al., 2015; Hanson et al., 2016). Linkage mapping involving these resistant accessions has been used to identify several large-effect quantitative trait loci (QTLs) (Table 1) that confer one or more categories of resistance to soybean aphid (reviewed by Hill et al., 2012; Hesler et al., 2013). Management of soybean aphid with host-plant resistance is complicated by the fact that soybean aphid biotypes that are virulent to aphid-resistant plants continue to be discovered in North America (Kim et al., 2008; Hill et al., 2012). Biotype 1 aphids cannot colonize plants with any known Rag genes, Biotype 2 aphids can colonize plants with Rag1 but not Rag2, Biotype 3 aphids can colonize plants with Rag2 (Hesler et al., 2013), and Biotype 4 aphids can colonize plants with either Rag1, Rag2, or both genes (Alt and Ryan-Mahmutagic, 2013). Pyramiding multiple resistance genes further increases efficacy against soybean aphid (McCarrave et al., 2014b; Chandrasona et al., 2015; Ajayi-Oyetunde et al., 2016). Additional sources of resistance will be needed to manage soybean aphid virulence to aphid-resistant plants (Michel et al., 2011; Hesler et al., 2013). To date, at least 10 Rag genes have been described as conferring vertical resistance or biotype-specific resistance (Van Der Plank 1966), but the underlying mechanisms of the host–aphid interactions with these resistance genes are still being examined (e.g., Li et al., 2008; Chiozza et al., 2010; Bansal et al., 2014).

Soybean is an autogamous crop species with limited genetic diversity among elite lines; therefore, elite lines are often outcrossed to plant introductions to introgress new resistance traits (Chung and Singh, 2008). Identification of soybean accessions carrying new aphid-resistance traits requires the screening of many accessions, most of which are susceptible (e.g., Bansal et al., 2013; Hesler, 2013; Bhusal et al., 2013, 2014; Hanson et al., 2016). Furthermore, the use of linkage mapping to determine the genetic basis of a resistance trait requires the creation of specialized mapping populations by crossing known resistant and susceptible accessions, requiring considerable time and effort (Zhang et al., 2013).

Many of the accessions previously used in screening studies were obtained from the USDA’s Soybean Germplasm Collection (Urbana, IL). The collection contains approximately 18,480 cultivated soybean accessions (G. max) and 1168 wild soybean (G. soja) accessions (Song et al., 2015). Therefore, only about 18% of the collection has been screened and described in peer-reviewed literature to date. These 19,648 G. max and G. soja accessions underwent a genotype analysis (i.e., the SoySNP50K array) that identified 42,509 SNPs across the soybean genome for each accession in the collection (Song et al., 2013, 2015). With multiple screening studies for soybean aphid resistance and the genotype data provided by Song et al. (2015), it is possible to further explore the genetic basis of soybean aphid resistance without needing to develop mapping populations for each resistant accession. A genome-wide association mapping study (GWAS) relies on historical recombination events, as opposed to recombination events within individual biparental populations (Myles et al., 2009). This approach can provide higher mapping resolution and surveys the full allelic diversity present within a pool of germplasm (Myles et al., 2009).
Multiple GWASs in soybean have identified SNP associations with agronomic traits of soybean including protein and oil content (Vaughn et al., 2014; Bandillo et al., 2015); flowering, maturity, and plant height (Zhang et al., 2015a); seed weight (Zhang et al., 2016); and resistance to soybean sudden death syndrome (Bao et al., 2015; Zhang et al., 2015b), Sclerotinia stem rot (Iquira et al., 2015), and soybean cyst nematode (Vuong et al., 2015). Genome-wide association studies have also been used for aphid pests. Qin et al. (2017) examined cowpea [Vigna unguiculata (L.) Walp.] resistance to cowpea aphid [Aphis craccivora Koch (Hemiptera: Aphididae)], where 338 cowpea accessions were included in multiple studies for a total of 2504 aphid evaluations. Some accessions with known Rag genes included in the selected studies had been genotyped. Bansal et al. (2013, 2014) included PI243540 (Rag1), PI567598B (Rag1b), and PI 567543C (Rag3), and PI 567541B (rag4 and rag1c). These accessions were also used by Bhushal et al. (2014) in addition to PI567598B (rag1b). The country of origin and maturity group (MG) designations of each accession were obtained from the USDA-GRIN database (USDA-ARS, 2016).

In studies listed as using Biotype 1 aphids, aphids came from a confirmed laboratory strain; those listed as Biotype 2 or 3 were field-collected colonies that exhibited Biotype 2 or 3 reactions overall in resistance assays, but may still contain other biotypes. Experiments by Bhushal et al. (2013, 2014) were performed in a greenhouse, and the remaining studies were performed in growth chambers. Direct aphid counts were not available from these studies, with the exception of Hanson et al. (2016). Reported aphid ratings were rounded to the nearest integer and converted to the midpoint of the respective scale’s range to account for interval-censoring and to standardize the phenotype data across studies. For example, a rating encompassing 25 to 100 aphids would be converted to 62.5 aphids. Right-censored ratings (e.g., greater than 500 aphids) were approximated as 640 aphids for Bhushal et al. (2013, 2014), 500 aphids for Bansal et al. (2013), 175 aphids for Hesler et al. (2013), and 275 aphids for Hesler et al. (2017). Approximated aphid counts were transformed via natural log-transformation to adjust for non-normality and for stabilizing variances across different levels of aphid counts (Martin et al., 2009; Schwantes-An et al., 2016). The log-transformed values were used as the response variable in downstream analyses of aphid counts.

### MATERIALS AND METHODS

**Phenotype Data**

The literature was searched for studies that screened for soybean aphid resistance in accessions contained in the USDA Soybean Germplasm Collection. Screening methodologies were highly variable among studies and therefore studies that used similar methodologies were chosen so that data could be more easily combined across studies. The methods for these studies generally consisted of potted plants of different soybean accessions (early vegetative growth stages) being placed in close proximity to one another to allow aphids to freely move between plants. Aphid populations were allowed to grow for 14 d. Such studies were chosen, because both antibiosis and antixenosis could influence aphid populations in these assays (Hanson et al., 2016). Though accessions exhibiting tolerance have been identified (Pierson et al., 2010; Bansal et al., 2013; Prochaska et al., 2013), such studies were not included in this analysis because too few accessions have been screened for tolerance to provide a robust GWAS. In addition, tolerance studies measure differences in yield between infested and uninfested plants instead of measuring aphid densities (Prochaska et al., 2013).

Aphid resistance data were compiled from six studies (Table 2). A total of 2366 unique soybean accessions were evaluated across these six studies; 135 accessions were included in multiple studies for a total of 2504 aphid evaluations. Some accessions with known Rag genes included in the selected studies had been genotyped. Bansal et al. (2013) included PI243540 (Rag2), Bhushal et al. (2013) included PI548663 (Rag1), PI243540 (Rag2), PI567543C (Rag3), and PI 567541B (rag4 and rag1c). These accessions were also used by Bhushal et al. (2014) in addition to PI567598B (rag1b). The country of origin and maturity group (MG) designations of each accession were obtained from the USDA-GRIN database (USDA-ARS, 2016).

In studies listed as using Biotype 1 aphids, aphids came from a confirmed laboratory strain; those listed as Biotype 2 or 3 were field-collected colonies that exhibited Biotype 2 or 3 reactions overall in resistance assays, but may still contain other biotypes. Experiments by Bhushal et al. (2013, 2014) were performed in a greenhouse, and the remaining studies were performed in growth chambers. Direct aphid counts were not available from these studies, with the exception of Hanson et al. (2016). Reported aphid ratings were rounded to the nearest integer and converted to the midpoint of the respective scale’s range to account for interval-censoring and to standardize the phenotype data across studies. For example, a rating encompassing 25 to 100 aphids would be converted to 62.5 aphids. Right-censored ratings (e.g., greater than 500 aphids) were approximated as 640 aphids for Bhushal et al. (2013, 2014), 500 aphids for Bansal et al. (2013), 175 aphids for Hesler et al. (2013), and 275 aphids for Hesler et al. (2017). Approximated aphid counts were transformed via natural log-transformation to adjust for non-normality and for stabilizing variances across different levels of aphid counts (Martin et al., 2009; Schwantes-An et al., 2016). The log-transformed values were used as the response variable in downstream analyses of aphid counts.

### Table 2: Studies used for aphid resistance phenotype data (assessed at 14 d after infestation) that included accessions with single nucleotide polymorphism data.

<table>
<thead>
<tr>
<th>Species</th>
<th>Study</th>
<th>Lines</th>
<th>Maturity Group</th>
<th>Biotype</th>
<th>State</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. max</td>
<td>Suppl. Table 1 from Bansal et al. (2013)</td>
<td>873</td>
<td>II–IV</td>
<td>1</td>
<td>Ohio</td>
<td>1–5 rating scale: 1: &lt; 25, 2: 25–100, 3: 101–200, 4: 201–400, and 5: &gt; 400 aphids per plant</td>
</tr>
<tr>
<td></td>
<td>Table 3 from Bhushal et al. (2013)</td>
<td>334</td>
<td>I</td>
<td>2†</td>
<td>South Dakota</td>
<td>Scale similar to Bansal et al. (2013) except with 4: 201–500 and 5: &gt; 500.</td>
</tr>
<tr>
<td></td>
<td>Suppl. Table 1 from Bhushal et al. (2014)</td>
<td>341</td>
<td>00–0</td>
<td>3‡</td>
<td>South Dakota</td>
<td>Same scale as Bhushal et al. (2013)</td>
</tr>
<tr>
<td></td>
<td>Fig. 1 from Hanson et al. (2016)</td>
<td>74</td>
<td>000–I</td>
<td>1</td>
<td>Almoesota</td>
<td>Aphid counts per plant</td>
</tr>
<tr>
<td>G. soja</td>
<td>Table 1 from Hesler (2013)</td>
<td>137</td>
<td>0–III</td>
<td>1</td>
<td>South Dakota</td>
<td>1–4 rating scale: 1: &lt; 51, 2: 51–100, 3: 101–150, and 4: &gt; 150 aphids per plant</td>
</tr>
</tbody>
</table>

† A field-collected population that was virulent on a Rag1 line but avirulent on Rag2.
‡ A field-collected population that was virulent on a Rag2 line but avirulent on Rag1.
Population Structure Analysis

Allelic scores for 42,291 SNPs for each accession were accessed from data by Song et al. (2015) at https://soybase.org/snps (accessed 22 June 2018). Single nucleotide polymorphism genotypes were numerically coded as 1 for homozygous major alleles, -1 for homozygous minor alleles, and 0 for heterozygous alleles. We excluded unanchored sequence scaffold SNPs from our analysis (Song et al., 2013). All statistical analyses were conducted in R version 3.2.4 (R Core Team, 2013). The underlying population structure shared among accessions, such as soybean species, MG, and country of origin, was evaluated using principal component analysis implemented in the FactoMineR package (Lê et al., 2008; Kumar et al., 2014). The number of principal components used in the analysis was determined from a scree plot and by visually assessing the component number at which the rate of eigenvalue decrease began to plateau. ANOVA was used to join the tree based on Euclidean distances was constructed with the ape package in R (Saitou and Nei, 1987; Paradis et al., 2004). The neighbor-joining tree was exported to the Interactive Tree of Life website (Letunic and Bork, 2016) to allow more detailed viewing and searching. Accessions found to be resistant to the soybean aphid were color-coded based on their resistance to individual biotypes to visualize the distribution of biotype resistance across this set of accessions (Letunic and Bork, 2016).

Genome-Wide Association Mapping

Genome-wide association mapping was performed for Biotypes 1, 2, and 3 separately, and on all biotypes combined. Single nucleotide polymorphisms were filtered to ensure that at least 10 individuals contained the minor allele for any analysis to reduce the likelihood of false positive associations with aphid resistance while also avoiding the exclusion of rare alleles (Tabangin et al., 2009; Bergfelder-Drüing et al., 2015). Associations between SNPs and approximated aphid counts were analyzed with the package rrBLUP for genome-wide association mapping (Endelman 2011). The GWAS model was fitted as:

\[ y = Xb + Wm + Zu + e, \]

where \( y \) is the approximated aphid count, \( b \) is a vector of study and population structure effects fitted as fixed effects, \( X \) is an incidence matrix relating \( b \) to \( y \), \( m \) is a vector of fixed SNP allele effects, \( W \) is an incidence (marker) matrix relating \( m \) to \( y \); \( u \) is a vector of random polygenic effects where \( u \sim MVN(0,K_{pc}) \), \( K \) is a realized additive relationship matrix calculated from the markers with the \( A.mat \) function in the rrBLUP package; \( Z \) is an incidence matrix relating \( u \) to \( y \); and \( e \) is a vector of random residuals where \( e \sim MVN(0,K_e) \) and I is and identity matrix. The columns of X corresponding to study effects consisted of 0s and 1s, indicating which elements of \( y \) belonged to each study. The columns of \( X \) corresponding to population structure effects consisted of principal component loadings with the principal component analysis conducted on each set of accessions used for GWAS on each biotype separately, as well as the GWAS on all biotypes combined.

The \( rrBLUP \) package provided a p-value score (\(-\log(p)\)) for each SNP, where \( p \)-values below a multiple-comparisons-corrected false-detection threshold of \( \alpha = 0.05 \) indicated SNPs where genotype was significantly associated with differences in log-transformed aphid counts. Estimates of allelic effects were outputted from the \texttt{mixed.solve} function as in Rosyara et al. (2016). The allelic effect was the natural-log back-transformed \( n \)-fold change per minor allele or:

\[ \#\text{aphids} = x e^{\log(\Delta x)} , \]

where, for a given SNP, \( x \) is the estimated number of aphids on a homozygous major accession, \( a \) is the back-transformed allelic effect, and \( m \) is the number of minor alleles present for that SNP in an accession. For each individual biotype and the combined biotype analysis, the significant SNP with the highest \(-\log(p)\) score within a 500-kb window was selected for further analysis. The 500-kb window size was selected on the basis of the soybean linkage disequilibrium decay analyzed by Song et al. (2015). ANOVA was used to examine the relative contribution of each significant SNP to aphid resistance in this analysis after adjusting for study effects. Terms were included for each SNP as a main effect to measure the variance attributed to SNPs across all measured biotypes, in addition to a SNP x biotype interaction effect determine the importance of this source of variation.

RESULTS

Within each study, few accessions had relatively low aphid counts compared with the number of accessions with higher aphid counts (Fig. 1). For data pooled across all experiments, the first five principal components accounted for 24.3% of the SNP variance, with the first two components accounting for 9.4 and 6.3% of total variance, respectively (Supplemental File S1). Among species (\( G. \max \) versus \( G. \soja \)), country of origin, and MG, the variance explained by the first principal component was most strongly correlated with species \( (r^2 = 0.74) \). The second principal component was more strongly associated with country of origin \( (r^2 = 0.49) \) than with MG \( (r^2 = 0.35) \), which is in agreement with Bandillo et al. (2015). ANOVA of aphid counts and the first five principal components indicated that the model including principal components accounted for relatively little additional variation compared with the study term alone \( (R^2_{pc} = 0.36; R^2_{pc} = 0.37) \). For individual biotypes, five,
nine, and seven significant principal components were found for Biotypes 1, 2, and 3, respectively.

Individual resistance to Biotypes 1, 2, and 3 appeared to occur among distantly related accessions but also sometimes clustered among more closely related accessions, especially for resistance to multiple biotypes (Fig. 2). Biotype 1 resistance was also frequently found in G. soja accessions, and G. max accessions that were more closely related to G. soja frequently had resistance to at least one of the three biotypes. Closely related accessions with resistance included PI071506 with Biotype 2 resistance, PI567598B with Biotype 3 resistance, and PI567597C and PI567543C with Biotype 2 and 3 resistance, each of which are MG III or IV accessions originating from China (Fig. 2). Chinese accessions PI430491 (MG 00), PI567250A (MG I), and PI603712 (MG 0) were another group of adjacent leaves, which had Biotypes 2, Biotype 3, and Biotype 2 and 3 resistance, respectively (Fig. 2).

Though they were not as closely related as previously mentioned groups, PI157492 (Japan, MG IV) and PI605765B (Vietnam, MG unknown) with Biotype 1 resistance, PI567541B (China, MG III) with Biotype 1 and 3 resistance, and PI603587A (China, MG I) with Biotype 2 and 3 resistance occurred within a set of higher-level branches (Fig. 2). Two closely related clusters of resistant individuals also occurred in less G. soja-like accessions with resistance to one of the three biotypes. One cluster included PI639537 (Russia, MG I) with Biotype 1 resistance as well as PI437075 (Russia, MG I), PI378663 (Russia, MG I), and PI189946 (France, MG I) with Biotype 3 resistance (Fig. 2). The second cluster of PI639534A (Russia, MG I), PI464911 (China, MG 0), and PI153214 (Belgium, MG I) had Biotypes 1, 2, and 3 resistance, respectively.

**Genome-Wide Association Mapping**

Manhattan plots for Biotype 1 aphids on G. max and G. soja accessions showed SNPs that were highly associated with soybean aphid population density on chromosomes 2, 7, and 13 (Fig. 3). Significantly associated SNPs were also present on chromosomes 5, 9 to 11, and 16 to 20, where no aphid resistance genes have been documented to date (Fig. 3). One significant SNP on chromosome 7 fell within the reported *rag1c* interval (Table 1, Fig. 3), but the other significant SNPs on chromosome 7 occurred outside reported *Rag* gene intervals. Two associations were found on chromosome 13 where multiple *Rag* genes have been reported (Fig. 3). The most significant SNP on chromosome 13 fell within the intervals of *Rag2* and *Rag5* but a second strong association was observed approximately 45 Mbp away from this region (Fig. 3).

Significant SNP associations were also found for putative Biotype 2 aphid resistance (Bhusal et al., 2013) on chromosomes 1, 4, 6, 8, 10, 12, 13, and 14 (Fig. 4a). Significant SNPs on chromosomes 7 and 13 fell within previously reported *Rag* gene intervals (Fig. 4a). The significant SNP on chromosome 7, located at 5,062,637 bp, was close to the reported intervals of *Rag1* and *Rag1b* and was also within the relatively wide interval for *raglc*; a second SNP on chromosome 7 was at least 28.6 Mbp from these genes (Table 1). For aphids exhibiting Biotype 3 characteristics according to Bhusal et al. (2014), significant SNPs on chromosomes without known soybean aphid resistance genes were found on chromosomes 5, 8, 10, and 19 (Fig. 4b). Significant SNPs were also found on chromosome 13 outside the range of the reported *Rag* gene intervals (Fig. 4b). The SNP on chromosome 8 at 41,031,762 bp was only 424 kbp from *Rag6*.

In the combined analysis of all biotypes used in the studies we examined, significant SNPs were found associated with either high or low aphid densities across biotypes on chromosomes 1, 5, 6, 18, and 19 where *Rag* genes have not been reported (Fig. 5). A significant SNP was also found on chromosome 13 within the range of *Rag2* and *Rag5*; another was approximately 2.5 Mbp outside this range (Fig. 5). Several genomic regions were common across different biotypes in terms of encompassing SNPs with significant associations with aphid resistance, such as the *Rag1* and *Rag2* regions on chromosomes 7 and 13, as well as a novel region on chromosome 10 (Table 3; Fig. 4 and Fig. 5). Many regions, however, were unique to the different biotypes.

Effect sizes also varied across significant SNPs and biotypes. Decreased aphid counts (i.e., effect size < 1) were typically associated with the minor allele for each significant SNP (Table 3). For Biotype 1, ss715583602 on chromosome 2, ss715606645 on chromosome 10, ss715609271 on chromosome 11, and ss715616609 on chromosome 13 had the strongest allelic effects where aphid counts were nearly halved on homozygous minor compared with homozygous major accessions (Table 3). These resistant alleles were also relatively rare, ranging from 0.4 to 0.7% minor allele frequency (MAF). Conversely, the minor
alleles at six loci were associated with increased aphid counts (Table 3). For Biotype 1, ss715614803 occurred with a 8.5% MAF and associated with a 1.1-fold change in aphid counts per minor allele, which indicated that the major allele was associated with a corresponding decrease in aphid counts. For Biotype 2, ss715617401 on chromosome 14 occurred with a 15% MAF and was associated with a 1.6-fold increase in aphid counts per minor allele. A similar effect was also found for significant Biotype 3 SNPs ss715615352 and ss715616460 on chromosome 13 and ss715635565 on chromosome 19, as well as ss715635693 on chromosome 19 for the combined biotype analysis.

A linear model was constructed that included the effects of study, selected significant SNP representing QTLs pooled across biotype analyses, and all selected SNP × biotype interactions. This analysis indicated that the significant selected SNPs, as a main effect, accounted for 12.9% of the variance in aphid counts. The SNP × biotype interaction effects accounted for 15.0% of the variance in aphid counts. Of the variance accounted for by the SNP main effects, SNPs within the intervals of known Rag genes accounted for 13.8% of this variation, which was primarily caused by SNPs within the Rag2 and Rag5 ranges (Supplemental File S2). Single
nucleotide polymorphisms on chromosomes 6, 10, and 20 each explained more than an additional 5% variation and are not in close proximity to known Rag genes (Supplemental File S2). Of the variance accounted for by the SNP × biotype interactions or biotype-specific variation, SNPs within the regions of known Rag genes accounted for 14.1% of this variance. Single nucleotide polymorphisms on chromosomes 1, 5, 6, and 16, where no known Rag genes exist, accounted for 26.1% of biotype-specific variation (Supplemental File S2).

**DISCUSSION**

Researchers have screened large numbers of diverse soybean accessions held in the USDA Soybean Germplasm Collection in their search for genetic resistance to the most economically damaging insect pest of soybean in North America (reviewed by Hill et al., 2012; Hesler et al., 2013). These extensive time- and resource-intensive screenings have discovered many Rag genes of tremendous value in the ongoing effort to develop new soybean varieties with robust resistance. Now that the entire USDA Soybean Germplasm Collection has been genotyped with the SoySNP50K panel of genetic markers, these data have renewed value for genome-wide association mapping to ultimately identify additional QTLs conferring aphid resistance and to better understand the distribution of soybean aphid resistance across the known soybean germplasm pool.

Genome-wide association mapping can directly use available genotype and phenotype data, thus saving tremendous amounts of time and resources that would be required for the development of specialized linkage mapping populations. Here, we leveraged data from six published screenings of soybean aphid resistance, amounting to a total of 2366 unique soybean accessions, to gain a better understanding of the genetic architecture underlying the variation in soybean aphid resistance within the USDA Soybean Germplasm Collection. We found that soybean aphid resistance is broadly distributed throughout the collection, yet some groups of accessions that clustered by genomic relationships were enriched for aphid resistance. Moreover, we identified several genomic regions that had strong associations with aphid resistance and have not been identified before in the...
literature. The regions of currently characterized \textit{Rag} genes only explained approximately 14\% of the variation in aphid counts reported by the studies listed in Table 2, suggesting that a large portion of genes associated with soybean aphid resistance possibly remain undescribed. Our phylogenetic analysis based on genomic relationships measured in the SoySNP50K marker panel found that resistance to individual soybean aphid biotypes is broadly distributed across the soybean germplasm pool (Fig. 2). In some cases, resistance to multiple biotypes
occurred in closely related groups of individuals (Fig. 2), suggesting that ancestors of those groups could have undergone selection pressure to develop resistance to multiple biotypes. These multi-biotype-resistant clusters should serve as starting points for future linkage mapping experiments to identify novel Rag genes (e.g., Zhang et al., 2013; Bhusal et al., 2017), or for assessing how such resistance develops in conjunction with soybean aphid biotype evolution (Michel et al., 2011). Accessions that have not yet been phenotyped for aphid resistance but are closely related to those with documented resistance in Fig. 2 may also be targets for additional screening. Most accessions included in this study had only been tested against one biotype and thus most accessions not marked as resistant in Fig. 2 still need to be screened against the remaining biotypes. Targeted testing of accessions for resistance to multiple biotypes should be pursued, and the results from this study could help facilitate targeted evaluations.

In the GWAS analysis, 45 SNPs were significantly associated with changes in soybean aphid resistance on 18 of the 20 soybean chromosomes. Significant SNPs were found on chromosomes 7, 8, 13, and 16 with known Rag genes (Table 1; Table 3). Single nucleotide polymorphisms were also significant on chromosomes 1, 2, 4, 5, 6, 9, 10, 11, 12, 14, 17, 18, 19, and 20, where Rag genes have not yet been mapped (Table 3). Rag1, raglb, and raglc have been mapped to similar regions on chromosome 7 (Zhang et al., 2009; Kim et al., 2010a; Bales et al., 2013). Significant SNPs from this analysis also fell either within or close to these regions. However, ss715598285 was also significant against Biotype 2 populations (Table 3), which may indicate the presence of a non-Rag1 gene associated with this region on chromosome 7. For Biotype 3 aphids, significant SNPs were not found within the range of Rag2 on chromosome 13.

Allelic effect size also varied across SNPs and biotypes. Larger effect sizes were typically found for Biotype 2 and 3 aphids. This may have partly been caused by the larger sample size of the Biotype 1 analysis (n = 1829), whereas Biotype 2 and 3 analyses included ~300 accessions, where only larger effect sizes would be detectable because of their smaller sample size. Single nucleotide polymorphisms with large effect sizes would be especially useful for highlighting accessions with potential strong resistance, such as ss715634601 on chromosome 19 (which had a 64% reduction in aphid counts for each ss715634601 minor allele present). After initial infestation, a susceptible soybean homozygous for the ss715634601 major allele would be expected to reach high aphid densities, such as an average of 674 aphids per plant, which is the point at which aphids cause significant economic damage (Ragsdale et al., 2007). However, soybean accessions that were homozygous for the resistant minor allele at this SNP would be expected to have ~90 aphids per plant under the same conditions.

On chromosome 13, ss715614932 accounted for a large amount of variation across biotypes. This region is associated with multiple known Rag genes, especially Rag5, which has no known virulent biotype (Table 1). Screening studies have also often found accessions with resistance to be associated with this region for Rag2 (e.g., Fox et al., 2014), which may explain the high statistical significance of SNPs in this overall region. Other SNPs, such as ss715590206 on chromosome 5, did not account for a large amount of variation overall but instead accounted for a large amount of variation for specific biotypes. On chromosome 1, ss715578827 was only significant when all biotypes were analyzed jointly (Table 3). This SNP also had a small effect size with a 3.8% MAF (Table 3). The power of GWAS to detect alleles of low frequency is low, especially when samples sizes are modest; therefore, this SNP was probably not detected in individual biotype analyses with smaller sample sizes (Spencer et al., 2009; Lettre, 2011; Park et al., 2011). This may indicate weak but broad non-specific resistance (i.e., horizontal resistance), as it was a significant SNP across biotypes (Van Der Plank, 1966).

Horizontal resistance would be more difficult to detect in soybean aphid screening assays because of its lower effect size. Only vertical resistance (i.e., biotype-specific resistance) has been documented to date (Hesler, 2013). Horizontal resistance has been found in other aphid species (e.g., Nielson and Kuehl, 1982). A benefit of our combined biotype analysis (Fig. 5) is the increased sample size and the power to detect these potential effects that may not be detected in the single biotype analyses. In this analysis, potential horizontal resistance could be associated with the significant SNPs found in the combined analysis that accounted for a relatively high amount of main effect genetic variability with low variance for biotype interactions (Supplemental File S2). If horizontal resistance can be found for soybean aphid, it would be a valuable tool in maintaining effective host-plant resistance to combat biotypes (Smith and Chuang, 2014).

Chang and Hartman (2017) also performed a GWAS on soybean aphid resistance and other soybean insect pests in the USDA Soybean Germplasm Collection with 2395 accessions but found only one SNP for soybean aphid (ss71559614 on chromosome 7) that was significantly associated with resistance to soybean aphid. This is a marked difference from the number of SNPs we detected. The datasets we analyzed, excluding Bhusal et al. (2013) for Biotype 2 aphids (Fig. 4a), were not examined by Chang and Hartman (2017), so it is difficult to make direct comparisons between the two studies. However, Chang and Hartman (2017) used categorical phenotypic data instead of using aphid densities or rating scales reported in the literature. This would reduce the resolution of their phenotypic data and power to detect underlying resistance. Chang and Hartman (2017) also used studies with different biotypes but did not indicate if biotype or study effects were accounted for in the GWAS model. Not accounting for experiment or biotype as part of a meta-analysis could also obscure the significance of individual SNPs (Cornelis et al., 2010).

Few GWASs of soybean insects or other pests such as nematodes have been carried out, and these studies have generally found few SNPs associated with resistance
Although many accessions in the USDA Soybean Germplasm Collection have not yet been screened for resistance, many accessions that do have confirmed resistance have not undergone mapping experiments. Resistance to soybean aphid (Rag) genes could be prioritized for future linkage mapping experiments (e.g., Hill et al., 2006; Zhang et al., 2013). Our identification of SNPs associated with soybean aphid resistance should provide a new resource to guide researchers in future soybean aphid screening and mapping experiments and will hopefully expedite the discovery and integration of additional soybean aphid resistance genes into available soybean varieties.

**Supplemental Information**

Supplemental File S1. Principal component analysis categorized by (a) soybean species and (b) country or region of origin.

Supplemental File S2. Genetic variation accounted for by all significant SNPs and interactions with soybean aphid biotypes. Single nucleotide polymorphisms within reported ranges of known Rag genes are in bold.

**Conflict of Interest Disclosure**

The authors declare that there is no conflict of interest.

**ACKNOWLEDGMENTS**

We thank the Minnesota Soybean Research and Promotion Council, the Minnesota Invasive Terrestrial Plants and Pests Center, and the Minnesota Department of Agriculture for providing funding for this study. This research was supported in part by the University of Minnesota’s Doctoral Dissertation Fellowship. We would like to acknowledge those who assisted in gathering soybean aphid phenotypic data in the studies used in this analysis. We also give special thanks to Erin Gilbert for assistance on Glycma 2.0 marker positioning and Dr. Ralph Holzenthal and Dr. Amitrpal Singh for providing advice on phylogenetic tree analysis and presentation.

**REFERENCES**


Diaz-Montano, J., J.C. Reese, W.T. Schapaugh, and L.R. Campbell. 2006. Characterization of antibiotic and antixenosis to the soybean aphid...


