Theor Appl Genet
DOI 10.1007/s00122-013-2079-8

ORIGINAL PAPER

Identification of high-quality single-nucleotide polymorphisms in *Glycine latifolia* using a heterologous reference genome sequence

Sungyul Chang · Glen L. Hartman · Ram J. Singh · Kris N. Lambert · Houston A. Hobbs · Leslie L. Domier

Received: 2 October 2012 / Accepted: 23 February 2013

Abstract Like many widely cultivated crops, soybean [*Glycine max* (L.) Merr.] has a relatively narrow genetic base, while its perennial distant relatives in the subgenus *Glycine* Willd. are more genetically diverse and display desirable traits not present in cultivated soybean. To identify single-nucleotide polymorphisms (SNPs) between a pair of *G. latifolia* accessions that were resistant or susceptible to *Sclerotinia sclerotiorum* (Lib.) de Bary, reduced-representations of DNAs from each accession were sequenced. Approximately 30 % of the 36 million 100-nt reads produced from each of the two *G. latifolia* accessions aligned primarily to gene-rich euchromatic regions on the distal arms of *G. max* chromosomes. Because a genome sequence was not available for *G. latifolia*, the *G. max* genome sequence was used as a reference to identify 9,303 *G. latifolia* SNPs that aligned to unique positions in the *G. max* genome with at least 98 % identity and no insertions and deletions. To validate a subset of the SNPs, nine TaqMan and 384 GoldenGate allele-specific *G. latifolia* SNP assays were designed and analyzed in F_2 *G. latifolia* populations derived from *G. latifolia* plant introductions (PI) 559298 and 559300. All nine TaqMan markers and 91 % of the 291 polymorphic GoldenGate markers segregated in a 1:2:1 ratio. Genetic linkage maps were assembled for *G. latifolia*, nine of which were uninterrupted and nearly collinear with the homoeologous *G. max* chromosomes. These results made use of a heterologous reference genome sequence to identify more than 9,000 informative high-quality SNPs for *G. latifolia*, a subset of which was used to generate the first genetic maps for any perennial *Glycine* species.

Abbreviations

ITS	Internal transcribed spacer
PCR	Polymerase chain reaction
PI	Plant introduction
QTL	Quantitative trait loci
SNP	Single-nucleotide polymorphism

Introduction

Soybean [*Glycine max* (L.) Merr.; $2n = 40$] is the sole domesticated member of 28 known species of the genus *Glycine* Willd. (Ratnaparkhe et al. 2011). The genus *Glycine* has been divided into two subgenera, *Glycine* Willd. and *Soja* (Moench) F. J. Hermann. The subgenus *Soja* contains the annuals *G. max* (cultigen) and its wild progenitor *G. soja* Sieb. et Zucc; both native to Northeast Asia. The subgenus *Glycine* contains 26 perennial species, including *G. canescens* Herm., *G. clandestina* Wendl., *G. latifolia* (Benth.) Newell & Hymowitz, and *G. tabacina* (Labill.) Benth. that are native to Australia and grow in a wide range of climatic conditions (Chung and Singh 2008). The genetically diverse perennial *Glycine* species possess potentially useful genes that so far have not been characterized or used to improve soybean. Useful traits that have

Communicated by J. Ray.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2079-8) contains supplementary material, which is available to authorized users.

S. Chang · G. L. Hartman · K. N. Lambert · H. A. Hobbs · L. L. Domier
Department of Crop Sciences, University of Illinois,
Urbana, IL 61801, USA

G. L. Hartman · R. J. Singh · L. L. Domier (✉)
United States Department of Agriculture, Agricultural Research Service, Urbana, IL 61801, USA
e-mail: leslie.domier@ars.usda.gov

Published online: 15 March 2013

been identified in the perennial species include genes for resistance to pathogens that cause brown spot (Lim and Hymowitz 1987), cyst nematode (Riggs et al. 1998), Sclerotinia stem rot (Hartman et al. 2000), and soybean rust (Burdon and Marshall 1981; Hartman et al. 1992). Fertile plants have been recovered from crosses between *G. max* and *G. tomentella* ($2n = 78$) Hayata by embryo rescue of F_1 immature seeds, colchicine-induced chromosome doubling, and repeated backcrossing (Singh 2010; Singh et al. 1998). Others have attempted to produce crosses between *G. max* and other perennial *Glycine* species using similar techniques, but failed to produce fertile lines (Grant 1990; Hammatt et al. 1992; Newell et al. 1987). As an alternative to wide crosses, it may be possible with appropriate genetic and genomic resources, to identify the desirable genes in the wild relatives by positional cloning and move them into highly adapted varieties through DNA-mediated transformation (Dinkins and Collins 2008; Rech et al. 2008).

Recent developments in techniques for high-throughput genomics have provided tools to characterize genetic diversity in perennial *Glycine* species that could be beneficial to soybean production especially when resistance is generally lacking, as in the case of soybean resistance to Sclerotinia stem rot. While the genome of *G. soja* has been sequenced (Kim et al. 2010), few molecular genetic tools are available for the perennial *Glycine* species, and no molecular markers have been described for *G. latifolia* to date. Only a few attempts to use markers developed for *G. max* with perennial *Glycine* species have been successful. For example, on average about one-third of the primers for *G. max* microsatellite markers amplified fragments from *G. clandestina*, *G. cyrtoloba* Tind. or *G. tomentella* DNAs (Hempel and Peakall 2003; Zou et al. 2004). To begin to address the lack of molecular markers for perennial *Glycine* species, Bronski et al. (2009) identified 13 microsatellite markers that were polymorphic among the A-genome perennials (*G. argyrea* Tind., *G. canescens*, *G. clandestina*, *G. latrobeana* (Meissner) Benth., *G. rubiginosa* Tind. & Pfeil, and *G. syndetika* Pfeil & Craven). However, these microsatellite markers have not been tested in *G. latifolia.* Recently, Hyten et al. (2010a) and Wu et al. (2010) reported the identification of single-nucleotide polymorphism (SNP) markers using high-throughput Illumina sequencing of reduced representation libraries of genomic DNA of *G. max* parental lines and used them to analyze corresponding mapping populations. In both studies, greater than 90 % of the SNPs predicted from multiple sequences and selected for analysis were verified, which demonstrated the quality of the Illumina sequencing data. If SNP frequencies are comparable in the perennial *Glycine* species with those in *G. max*, high-throughput next-generation DNA sequencing could be a powerful tool to rapidly develop molecular genetic tools for the characterization of those materials.

Sclerotinia stem rot (white mold), an important yield-reducing disease of soybean in the United States (Hartman et al. 1998), is caused by *Sclerotinia sclerotiorum* (Lib.) de Bary, a necrotrophic fungus that produces oxalic acid, which induces the death of plant tissues before colonization (Purdy 1979). Inheritance of low levels of resistance to *S. sclerotiorum* has been reported in *G. max* as being multigenic with relatively low heritability. Quantitative trait loci (QTLs) for resistance to Sclerotinia stem rot have been identified on 17 of the 20 *G. max* chromosomes (Arahana et al. 2001; Guo et al. 2008; Huynh et al. 2010; Kim et al. 2000; Li et al. 2010; Vuong et al. 2008). However, the associated QTLs have explained relatively low levels of the observed phenotypic variation, making marker-assisted selection for resistance challenging. In contrast, selected *G. latifolia* accessions showed higher levels of resistance to Sclerotinia stem rot (Hartman et al. 2000). In addition to enhanced resistance to Sclerotinia stem rot, *G. latifolia* accessions have been identified that are significantly more tolerant to the herbicide 2,4-D (2-4-dichlorophenoxyacetic acid) than *G. max* (Hart et al. 1991). Additionally, *G. latifolia* accessions have been shown to be tolerant to grazing, drought, and frost, as well as resistant to infection by *Alfalfa mosaic virus*, a common pathogen of soybean in the Midwestern United States (Giesler and Ziems 2006; Horlock et al. 1997; Jones et al. 1996; Mueller and Grau 2007). Hence, *G. latifolia* is a potential source of several agronomically relevant genes for *G. max*.

As a consequence of whole genome duplications, plant genomes contain large numbers of genes with very similar sequences in addition to interspersed highly repetitive transposable elements (Soltis et al. 2009). These repeated sequences pose a challenge for the use of high-throughput sequence data for SNP identification in plant species for which reference genome sequences are not available. Recent duplications, like that in the genus *Glycine*, are evidenced by the presence of duplicate copies of most genes, which are often maintained in blocks with similar gene organizations (Schmutz et al. 2010). In addition other genes, notably genes for resistance to plant pathogens, are members of highly conserved gene families (Innes et al. 2008). While methods for SNP identification without a reference genome sequence have been reported (Ratan et al. 2010), it can still be difficult to differentiate between nucleotide sequence variation among duplicated genes and allelic variation. In this study, we identified SNPs between *G. latifolia* accessions that were resistant or susceptible to *S. sclerotiorum* using the *G. max* genome sequence as a reference to help differentiate between highly conserved duplicated loci.

Materials and methods

Plant materials and population development

Glycine latifolia ($2n = 40$) accessions PI 559298 and PI 559300 were obtained from the USDA Soybean Germplasm Collection in Urbana, Illinois (http://www.ars-grin.gov/npgs/urbana.html). The responses of the two *G. latifolia* accessions to inoculation with *S. sclerotiorum* were confirmed using three to four plants for each accession grown in five 10-cm pots. Plants were inoculated with mycelial-agar plugs as described by Hartman et al. (2000) and evaluated for viability at 1-week intervals and reinoculated after 2 and 3 weeks. The experiment was terminated after 4 weeks and repeated once. The numbers of surviving plants were counted at the end of the experiment for each accession.

Reciprocal crosses were performed between the two *G. latifolia* accessions. Because *G. latifolia* accessions lacked obvious phenotypic markers that could be used to confirm that putative F_1 plants were true hybrids, DNA was extracted from each parent using a DNeasy Plant Mini Kit (Qiagen, Valencia, CA) and portions of their *DCL3* genes (Glyma04g06060) were amplified by polymerase chain reaction (PCR) using primers 5′-ccgagaaattcaagccctctgcca-3′ and 5′-aggggtcaccgctggatgtgt-3′. The products were treated with ExoSAP-IT (Affymetrix, Cleveland, OH) and directly sequenced with Big Dye fluorescent terminator sequencing reagents (Applied Biosystems, San Diego, CA) as recommended by the manufacturers and analyzed at the University of Illinois, W.M. Keck Center for Comparative and Functional Genomics. DNA from putative F_1 plants was similarly extracted, amplified and sequenced to confirm that the plants were true hybrids. The *DCL3* gene was selected because it is relatively large (>10 kb) and single copy in the *G. max* genome (Schmutz et al. 2010).

Library construction and sequencing

DNA was extracted from leaf tissue of PI 559298 and PI 559300 as described above. For each accession, 50 µg of DNA was digested with four methylation-insensitive restriction enzymes, *Hae*III, *Msl*I, *Pst*I, and *Ssp*I, and one enzyme, *Rsa*I, cleavage by which is blocked by some combinations of overlapping CpG methylation, as described by Hyten et al. (2010a). DNA fragments of 300 to 500 bp were excised from agarose gels and purified using a QIAquick Gel Extraction Kit (Qiagen). To control for bias in the distribution of *G. latifolia* sequence reads aligned to *G. max* chromosomes that may have been introduced by restriction enzyme digestion, DNA from PI 559298 was randomly fragmented by nebulization at 220 kPa for 1 min to produce fragments with an average size of 500 bp. Approximately 4 µg of each DNA sample was ligated to sequencing adapters and sequenced on Illumina sequencers (Illumina Inc., San Diego, CA) at the W. M. Keck Center. Single-end reads of restriction enzyme digested DNAs of PI 559298 and PI 559300 were sequenced in a single lane using Illumina Multiplexing Sample Preparation and Sequencing Primer kits. Sequence reads from the multiplexed data were assigned to each accession using Illumina Pipeline Analysis software. Paired-end reads of the randomly sheared DNA of PI 559298 were sequenced in a separate lane. Low-quality *G. latifolia* sequence reads and sequence reads that aligned to the *G. max* chloroplast genome sequence (Saski et al. 2005) were removed from the data sets using the programs fastq_quality_filter (http://hannonlab.cshl.edu/fastx_toolkit/) and Bowtie2 (Langmead and Salzberg 2012), respectively. To estimate the error rate in the sequencing reactions, filtered reads were aligned to four contig sequences (24–32 kb each) representing approximately 110 kb of the *G. latifolia* genome sequence that were assembled from the paired-end reads from randomly sheared DNA of PI 559298 using ABySS (Simpson et al. 2009). The filtered reads also were aligned to the *G. max* genome (Schmutz et al. 2010) and sequences of large-insert bacterial artificial chromosome (BAC) clones from a diploid *G. tomentella* accession (Wawrzynski et al. 2008) using Bowtie2 and the "-very-sensitive" preset, which allowed for insertions and deletions and up to 12 mismatches in comparisons of the *G. latifolia* sequence reads to *G. max* and *G. tomentella* sequences. *Glycine tomentella* is another perennial relative of *G. max*, and is more closely related to *G. latifolia* than to *G. max* (Ratnaparkhe et al. 2011). Consequently, a larger proportion of *G. latifolia* sequence reads were expected to align to *G. tomentella* than to *G. max* sequences. The filtered reads depleted of chloroplast sequences were also aligned to the predicted amino acid sequences of *G. max* and *G. tomentella* transposable elements (Wawrzynski et al. 2008) using USEARCH (Edgar 2010) and an E-value threshold of 1.0×10^{-3}. Sequence reads were deposited in the Sequence Read Archive (SRA052163.2, http://www.ncbi.nlm.nih.gov/Traces/sra) at the National Center for Biotechnology Information.

SNP discovery

Because a reference genome sequence was not available for *G. latifolia*, the *G. max* genome sequence was used to differentiate between informative SNPs and sequence variation in members of multigene families and between homoeologous genes that resulted from the whole-genome duplication that occurred in the genus *Glycine* between 5 × 10^6 and 1 × 10^7 years ago (Gill et al. 2009). Sequences from the two *G. latifolia* accessions that were polymorphic at a single position and were detected from 3 to 20 times from each accession were aligned to the *G. max* genome using Bowtie2. Potential SNP markers were selected from *G. latifolia* sequences that were at least 98 % identical to *G. max* sequences with no insertions or deletions, and that aligned at a single location in the *G. max* genome. The alignment parameters were selected to differentiate between homoeologous loci because in *G. max* homoeologs share from 54 to 97 % nucleotide sequence identity (Schlueter et al. 2007). Initially, four and five loci allele-specific real-time PCR assays were designed using the Custom TaqMan Assay Design Tool (Applied Biosystems, Foster City, CA) to SNPs that mapped to *G. max* chromosome 4 and satellite chromosome 13, respectively, and obtained from Applied Biosystems (Supplemental Table 1). Subsequently, 384 GoldenGate SNP assays were designed using Illumina DesignStudio for SNPs that mapped to all 20 *G. max* chromosomes at an average spacing between SNPs of 2.4 × 10^6 bp. SNPs were selected with an average design score of 0.94 (Supplemental Table 2). Markers were named for the *G. max* chromosome and the nucleotide position on the chromosome ($\times 10^{-6}$) to which the SNP-containing sequences aligned.

SNP confirmation and segregation

Two populations of 91 and 92 F_2 plants were derived from reciprocal crosses between PI 559298 and PI 559300. DNA was extracted from each plant as described previously, assayed for TaqMan assays using an Applied Biosystems PRISM 7000 Sequence Detection System and PRISM software and for GoldenGate assays using an Illumina iScan and genotype calls were made with Illumina BeadStudio. The markers were separated into linkage groups at a log of the likelihood ratio (LOD) 6.0 and maximum recombination level of 0.35, and the most probable marker order was selected after boot strap analysis using AntMap (Iwata and Ninomiya 2006). The positions on the *G. max* genetic map of the *G. latifolia* sequences were extrapolated by comparing the nucleotide sequence positions on *G. max* chromosomes of the aligned *G. latifolia* sequences to nucleotide sequence positions of mapped *G. max* SNPs (Hyten et al. 2010b). Centimorgan (cM) map distances were estimated in *G. latifolia* using the Kosambi mapping function (Kosambi 1944). Goodness of fit of the observed segregation ratios to the expected 1:2:1 ratio was evaluated with χ^2 tests in AntMap. Linkage maps were generated using MapDraw (Liu and Meng 2003).

Results

Phenotypic and sequence analysis

Four weeks after initial inoculations with *S. sclerotiorum*, all plants of PI 559298 (resistant) survived while all plants of PI 559300 (susceptible) died, confirming that the two accessions differed significantly in their response to the pathogen. Reduced representation libraries were prepared from genomic DNAs of PI 559298 and PI 559300, and sequenced on an Illumina sequencer, which produced 3.7 × 10^7 100-nt reads each. Sequencing of randomly sheared DNA from PI 559298 produced 1.54 × 10^8 100-nt reads, for which 3.6 × 10^7 reads were used in comparison with sequences from the restriction enzyme-digested DNAs. A total of 1,214 filtered reads from the genome representations of PI 559298 and PI 559300 that aligned to approximately 110 kb of *G. latifolia* sequence assembled from all of the sequencing reads showed an error rate of 0.14 %. This was very similar to the 0.16 % global error rate for Illumina data reported by Minoche et al. (2011) in their analysis of the genome sequences of *Arabidopsis thaliana* (L.) Heynh. and *Beta vulgaris* (L.). Allowing no mismatches between reads, approximately 26 % of the reads from each accession were singletons (9.0 × 10^6 from PI 559298 and 1.0 × 10^7 from PI 559300), and 50 % of the reads were represented from 2 to 50 times from each parent, suggesting the restriction enzyme digestions had resulted in representations of sufficient depth to detect SNPs between the two *G. latifolia* accessions (Table 1).

Alignment to *G. max* genome and *G. tomentella* sequences

An average of 30 % of the sequence reads from each *G. latifolia* accession aligned to the *G. max* genome sequence with at least 88 % sequence identity allowing insertions and deletions for an overall alignment rate of 1.2 × 10^4 *G. latifolia* reads aligned per 10^6 bp of *G. max* genomic DNA. Of the aligning reads, 48 % aligned at unique locations and 52 % aligned at more than one location. The *G. latifolia* sequence reads were also aligned to 4.5 × 10^6 bp of *G. tomentella* genomic sequence from 30 BACs (Wawrzynski et al. 2008). About 5.6 % of the *G. latifolia* sequence reads aligned to sequences from *G. tomentella* for an alignment rate of 4.6 × 10^5 *G. latifolia* reads aligned per 10^6 bp of *G. tomentella* sequence data, which was much higher than the rate at which *G. latifolia* reads aligned to the *G. max* genome. As with *G. max*, about 50 % of the *G. latifolia* reads aligned to unique locations in the combined *G. tomentella* BAC sequences.

The distribution of the aligned *G. latifolia* sequences on *G. max* chromosomes closely paralleled the densities of

Table 1 Frequency distribution of occurrences of 100-nt reads from restriction enzyme digested DNA of *Glycine latifolia* accessions PI 559298 and PI 559300

No. of occurrences of a 100-nt read	No. of 100-nt reads PI 559298 (%)	No. of 100-nt reads PI 559300 (%)
1	9,954,094 (24.1)	9,026,205 (27.0)
2–10	5,779,470 (18.0)	6,726,420 (15.7)
11–20	4,013,250 (13.4)	5,014,059 (10.9)
21–50	7,179,659 (23.2)	8,706,321 (19.4)
51–100	4,260,433 (9.9)	3,713,213 (11.5)
101–200	1,783,968 (2.8)	1,039,182 (4.8)
201–500	1,065,651 (2.6)	962,012 (2.9)
501–1,000	677,178 (1.6)	596,512 (1.8)
1,001–5,001	994,544 (2.4)	881,218 (2.7)
5,001–10,000	407,785 (1.1)	413,519 (1.1)
>10,001	802,052 (1.0)	371,172 (2.2)
Total	36918084 (100)	37449833 (100)

The percentages of the total reads represented by each class are shown in parentheses

gene sequences on *G. max* chromosomes (Schmutz et al. 2010); few sequences aligned in heterochromatic pericentromeric regions (Singh et al. 1988), while many sequences aligned in distal portions of chromosomes (Fig. 1a). To test the preferential association of *G. latifolia* sequences with *G. max* coding regions, the positions in the *G. max* genome to which *G. latifolia* sequences aligned and a set of randomly generated chromosomal positions were compared with the positions of *G. max* gene models. *Glycine latifolia* sequences aligned on average within 3.1 kb of predicted genes in the *G. max* genome compared with >61.5 kb for randomly selected positions. The possibility that few *G. latifolia* sequence reads aligned in pericentromeric regions of *G. max* chromosomes because of differential sensitivities of euchromatic and heterochromatic DNAs to restriction enzyme digestion was evaluated by comparing the distribution of sequence reads from randomly sheared and enzyme-digested DNAs. The distribution of the sequence reads from restriction enzyme-digested and randomly sheared DNA of PI 559300 was very similar (Fig. 1a), indicating that the distribution of sequences did not result from bias introduced by restriction enzyme digestion, but rather from divergence of nucleotide sequences of centromere-associated elements. When *G. latifolia* chloroplast-derived sequences were included in the analysis, large numbers of chloroplast reads aligned to *G. max* chromosomes, including regions on chromosomes 4, 9,12, and 14 (Fig. 1b), and likely represented integration of chloroplast sequences into *G. max* chromosomes.

The pericentromeric regions of *G. max* chromosomes contain high densities of centromeric repeats and transposable elements (Du et al. 2010b; Schmutz et al. 2010). Even though Class I and Class II transposable elements constitute about 58 % of the *G. max* genome (Schmutz et al. 2010), just 1.7 % of the *G. latifolia* sequence reads aligned with the nucleotide sequences of 32,370 *G. max* transposable elements (Du et al. 2010a), and 0.8 % of the *G. latifolia* sequence reads aligned to the nucleotide sequences of 20 *G. tomentella* transposable elements (Innes et al. 2008). Because nucleotide sequences of transposable elements are poorly conserved in the genus *Glycine* (Lin et al. 2005), the predicted amino acid sequences of *G. latifolia* sequence reads and *G. max* and *G. tomentella* transposable elements were also compared to increase the sensitivity of the analysis. The analysis showed that 6.6 and 13.0 % of the predicted amino acid sequences of *G. latifolia* sequence reads aligned to predicted amino acid sequences of *G. max* and *G. tomentella* transposable elements, respectively. The higher proportion of reads aligning to *G. tomentella* than to *G. max* is not surprising given the closer relationship of *G. latifolia* to *G. tomentella* than to *G. max* (Ratnaparkhe et al. 2011) and the observations of Chesnay et al. (2007) that SIRE-1 retroelements in perennial *Glycine* species were more diverse and were distinct from those of *G. max* and *G. soja*.

Identification and verification of SNPs in *G. latifolia* accessions

Of the 100-nt sequences that were represented 3 to 50 times in each *G. latifolia* accession, over 350,000 were polymorphic at one or more positions between PI 559298 and PI 559300. To differentiate between nucleotide sequence variation in homoeologous genes and alleles of the same gene, the sequence reads containing polymorphisms were aligned to the *G. max* genome allowing at most two mismatches (≥98 % identity) and no insertions and deletions. With those parameters, 2.6 % of the variant *G. latifolia* sequences aligned to unique positions in the *G. max* genome (Table 2), which corresponded to 9,303 SNPs (an average of 456 SNPs per *G. max* chromosome) that might serve to anchor a *G. latifolia* genetic map to the *G. max* genetic map.

To test the quality of the SNPs and their usefulness in comparing the synteny between the *G. latifolia* and *G. max* genomes, a subset of the unique 9,303 SNP-containing sequences were selected for production of allele-specific markers. Initially, nine markers, four on chromosome 4 and five on chromosome 13 were selected for the production of TaqMan PCR markers. All nine of the *G. latifolia* SNPs for which assays were designed segregated in the expected 1:2:1 manner in 92 lines of a *G. latifolia* F_2 population (Table 3), formed two distinct linkage groups, and mapped in similar orders in *G. latifolia* and *G. max* except that markers C13_33.2, C13_34.0, and C13_35.3 on

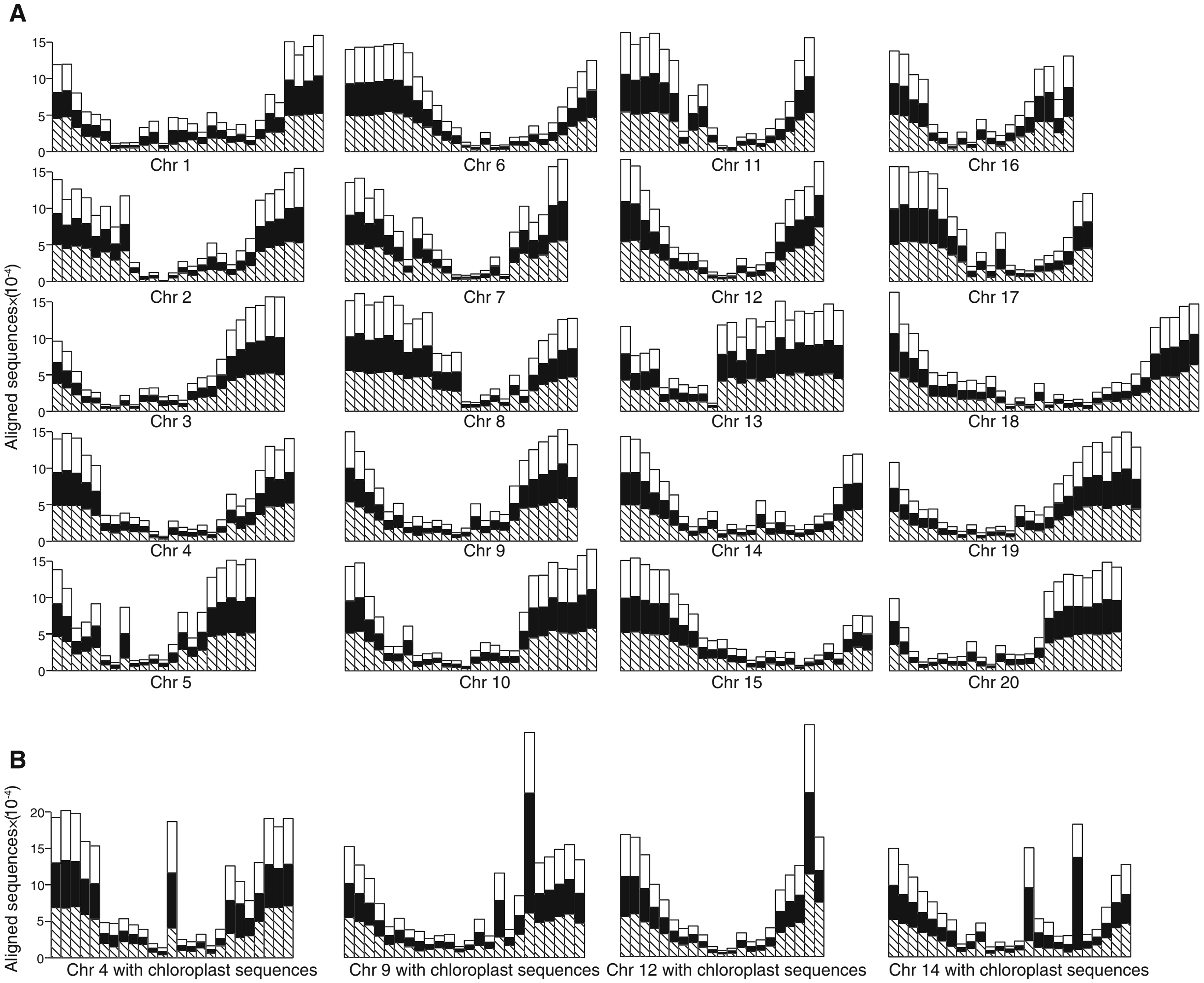

Fig. 1 Densities of alignments to the sequences of the 20 *Glycine max* chromosomes of 100-nt reads from restriction enzyme-digested DNA from *G. latifolia* accessions PI 559298 (black bars) and PI 559300 (*white bars*) and randomly sheared DNA of PI 559298 (*hatched bars*) **a** without chloroplast sequences and **b** including chloroplast sequences. Most of the *G. latifolia* sequences aligned to gene-rich regions on the distal ends of *G. max* chromosomes. Few reads aligned within pericentromeric regions. Sequence reads nearly identical to *G. max* chloroplast genome sequence (GenBank accession DQ317523) aligned in large numbers to multiple *G. max* chromosomes, including chromosomes 4, 9, 12, and 14. The width of each *bar* represents 2 × 10^6 bp

chromosome 13 were inverted in *G. latifolia* relative to *G. max* (Fig. 2). The SNP-containing *G. latifolia* sequences also aligned with homoeologous genes on *G. max* chromosomes 6, 7, and 15 (Table 3). However, the clustering of the markers into two linkage groups indicated that the markers differentiated the homoeologous loci. To further confirm the usefulness of identified SNPs, an additional 384 Illumina GoldenGate SNP assays were evaluated in a second 91-line F_2 population. Of the 384 SNP assays tested, 28 failed, 62 were monomorphic, and three were heterozygous in all lines. A higher percentage of GoldenGate markers were monomorphic than TaqMan markers because the automated selection methods used for GoldenGate assays did not take into account sequence variation within each parent. Most of the monomorphic markers were members of highly repeated, likely multilocus sequences that were disregarded in the manual selection of TaqMan markers. Of the remaining 291 markers, all but 26 segregated in 1:2:1 manner. For 16 of the 26 markers that deviated from 1:2:1, one of the parental lines was heterozygous for the markers. At LOD threshold of 6.0, 16 markers were unlinked, four of which were predicted to be located near the ends of linkage groups where recombination distances may have been too large to be incorporated into one of the linkage groups. Uninterrupted *G. latifolia* linkage maps were assembled that were nearly collinear with *G. max* chromosomes 1, 3, 4, 10, 11, 12, 15, 17, and 18 (Fig. 3a, b). The orders of some closely spaced

Table 2 Numbers of 100-nt reads form *Glycine latifolia* that aligned to a single position on one of the sequences of *G. max* chromosomes

Chromosome	≥88 % identity/Indels/ ≥1 hit		≥98 % identify/ No indels/Unique	
	PI 559298	PI 559300	Total	No. SNPs
1	599,037	602,791	147,240	336
2	555,736	596,865	199,843	530
3	405,784	443,000	163,916	336
4	522,333	538,581	164,113	470
5	480,397	505,656	175,942	314
6	518,784	566,169	200,982	482
7	477,489	498,240	167,620	344
8	638,819	672,305	254,067	464
9	652,933	578,126	176,930	752
10	557,401	592,765	205,402	474
11	478,169	509,346	186,168	336
12	437,672	453,342	153,035	532
13	1,038,953	1,008,404	242,433	672
14	597,554	527,685	129,076	616
15	472,328	505,951	166,248	317
16	343,559	371,813	116,176	264
17	479,794	512,677	194,599	404
18	593,607	587,480	164,670	406
19	490,419	508,367	164,079	392
20	483,587	500,333	153,572	694
Scaffolds	119,415	114,924	96,051	168
Total	10,943,770	11,194,820	3,622,162	9,303

Single-nucleotide polymorphisms (SNPs) were called when at least three 100-nt reads each from plant introductions 559298 (resistant) and 559300 (susceptible) co-varied at the polymorphic positions

markers that mapped to regions that corresponded to pericentromeric regions in the homoeologous *G. max* chromosomes were not resolved correctly (e.g., markers C04_27.9 and C18_33.8 on chromosomes 4 and 18, respectively). Similarly, the orders of *G. max* SNP markers derived from sequences within centromeric regions were difficult to resolve and the genetic maps were not always collinear with the genomic sequence in these regions (Hyten et al. 2010b). *Glycine latifolia* linkage maps that were syntenic with the remaining *G. max* chromosomes were represented by two or more groups of linked markers. The numbers of markers assigned to the *G. latifolia* linkage groups ranged from 8 to 19 for *G. latifolia* linkage groups syntenic with *G. max* chromosomes 19 and 18, respectively, depending on the success rate of marker design. Markers were selected at an average spacing of 2.4×10^6 bp, but as would be expected, map distances in *G. latifolia* linkage groups were much lower for markers that aligned to pericentromeric regions than to the distal arms of *G. max* chromosomes. Total map distance for the combined *G. latifolia* linkage groups was 2,166 cM which is similar to estimates for *G. max* of 2296 to 2,550 cM (Choi et al. 2007; Cregan et al. 1999; Hyten et al. 2010b; Song et al. 2004). The total map distance did not include unmapped terminal markers or gaps between markers, which likely would increase the total size of the linkage map. The results from both TaqMan and GoldenGate markers illustrated the usefulness of a heterologous reference genome sequence to identify unique and informative SNPs.

Discussion

In this study, SNP-containing *G. latifolia* sequences that aligned to unique locations in the *G. max* genome with at least 98 % identity were shown to differentiate homoeologous loci and map in the expected linkage groups. Even though the use of a heterologous reference genome sequence greatly reduced the number of potential SNPs, it permitted the identification of sufficient variant single-copy sequences to allow construction of uninterrupted genetic linkage maps for nine of the 20 *G. latifolia* chromosomes. Although the genetics of resistance to soybean rust has been investigated in segregating populations of *G. argyrea*, *G. canescens* and *G. tomentella* (Burdon 1988; Jarosz and Burdon 1990; Schoen et al. 1992), the genetic linkage maps presented in this report are the first for a perennial *Glycine* species. As in soybean where low levels of genetic recombination are observed in the repeat-rich heterochromatic regions surrounding the centromeres (Schmutz et al. 2010), low levels of recombination were detected in the corresponding regions of *G. latifolia* chromosomes. The conserved marker orders and centromeric positions suggest that at least the nine *G. latifolia* chromosomes for which uninterrupted linkage maps were constructed share high levels of synteny with their *G. max* homoeologs that could be useful for gene identification and assembling complete genome sequences of perennial *Glycine* species. These observations are consistent with previous cytological comparisons of *G. latifolia* and *G. max* chromosomes (Singh et al. 1992).

The observation that few *G. latifolia* sequence reads aligned to pericentromeric regions of *G. max* chromosomes is consistent with prior observations that large-insert clones from pericentromeric regions of *G. max* failed to hybridize with genomic DNA from *Glycine* species other than *G. soja* (Lin et al. 2005). Gill et al. (2009) found that probes to high-copy centromeric satellite repeats, *CentGm-1* and *CentGm-2*, hybridized to chromosomes of *G. soja*, the wild annual progenitor of soybean, but to none of the perennial *Glycine* species analyzed, and concluded that there has been rapid divergence of the centromere-associated DNA sequences within the genus *Glycine*. The lack of sequence conservation in regions proximal to centromeres is common in legume genomes (Cannon et al. 2009) and may at least partially explain the difficulties of producing fertile hybrid plants between *G. max* and perennial *Glycine* species by traditional breeding methods, and the frequent cytogenetic observation of pericentromeric inversions in crosses among A and B genome species of the subgenus *Glycine* (Singh and Hymowitz 1985; Singh et al. 1988). In addition, activation of retrotransposons that occurs when plants harboring different populations of transposable elements are hybridized has also been associated with genome instability that leads to sterility in hybrids of interspecific crosses (Ma et al. 2007; Maheshwari and Barbash 2011).

In contrast to divergent intergenic regions and retrotransposons, gene coding sequences were more highly conserved among *Glycine* species. Ilut et al. (2012) in RNA-Seq analysis of transcripts from *G. dolichocarpa* Tateishi & Ohashi, *G. syndetika*, and *G. tomentella* found that just 10 % of transcript-derived sequences failed to align to *G. max* gene models. In contrast, 70 % of *G. latifolia* genomic DNA sequences failed to align to the *G. max* genome. When *G. latifolia* genomic sequences aligned to the soybean genome, they were positioned closer to highly conserved gene coding regions than would have been expected for randomly selected sequences. *Glycine latifolia* sequences that aligned to the soybean genome were nearly twice as likely to align at more than one location (54 %) than were RNA-Seq sequence reads from the three perennial *Glycine* species (38 %) (Ilut et al. 2012).

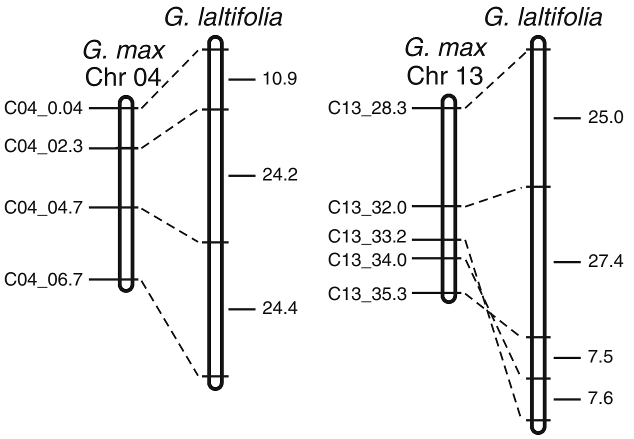

Fig. 2 Relative positions of nine *Glycine latifolia* TaqMan SNP markers that aligned to *G. max* chromosomes 4 or 13 were compared between *G. latifolia* and *G. max*. The positions of the markers in *G. max* were extrapolated from the genomic and genetic positions of previously mapped SNP markers. The genetic distances between markers for *G. latifolia*, indicated in Kosambi cM to the *right* of the *G. latifolia* maps, were estimated using a population of 92 F_2 *lines* from a cross between PI 559298 and PI 559300. As in *G. max*, the markers formed to distinct linkage groups. While the map distances between markers were larger in *G. latifolia* than in *G. max*, the relative orders of markers were similar, except that the order of markers C13_33.2, C13_34.0, and C13_35.3 on chromosome 13 were inverted in *G. latifolia* relative to the positions of the markers in *G. max*. Markers were named for the *G. max* chromosomes (C) and nucleotide positions on the chromosomes ($\times 10^{-6}$) to which the SNP-containing sequences aligned

Glycine latifolia sequences that were nearly identical to chloroplast genomes of *G. max* and other plants were the most highly represented sequences that aligned to the *G. max* genome. While the *G. latifolia* nuclear genome may also contain integrations of chloroplast sequences, the *G. latifolia* sequences were probably derived from plastid DNA, given the higher depth of coverage in regions containing sequences similar to chloroplasts sequences. While plastid-like sequences detected in the soybean genome could be assembly artifacts, integration of chloroplast sequences into nuclear genomes is common in both monocotyledonous and dicotyledonous plants and has been

Table 3 Differentiation of target and homoeologous loci with selected *Glycine latifolia* single-nucleotide polymorphisms (SNPs) in an F_2 population

SNP marker	Observed ratio[a]	Chromosome	Position	Targeted locus	Percent identity[b]	Homoeologous locus	Percent identity
C04_0.04	21:48:21	04	389,640	Glyma04g00680[c]	98	Glyma06g00720	94
C04_02.3	21:49:20	04	2,314,484	Glyma04g03170	98	Glyma06g03220	92
C04_04.7	22:55:14	04	4,677,344	Glyma04g06110	98	Glyma06g06110	94
C04_06.7	22:45:24	04	6,793,290	Glyma04g08670	99	Glyma06g08780	94
C13_28.3	26:47:16	13	28,205,002	Glyma13g24920	99	Glyma07g31520	95
C13_32.0	20:52:19	13	32,070,242	Glyma13g29140	97	Glyma15g09920	95
C13_33.2	22:40:27	13	33,414,540	Glyma13g30920	99	None found	N/A
C13_34.0	24:41:27	13	34,096,111	Glyma13g31700	98	Glyma15g07590	95
C13_35.3	22:46:23	13	35,453,033	Glyma13g33710	98	Glyma15g39070	89

[a] Observed ratios represent the numbers of plants that were homozygous resistant: heterozygous: homozygous susceptible

[b] Percent nucleotide sequence identity between *G. latifolia* accession PI 559298 and the *G. max* genome

[c] Abbreviation indicates plant species (*Glycine max* [Glyma]), two-digit chromosome number, and ordinal gene (g) number (five digits) for each locus

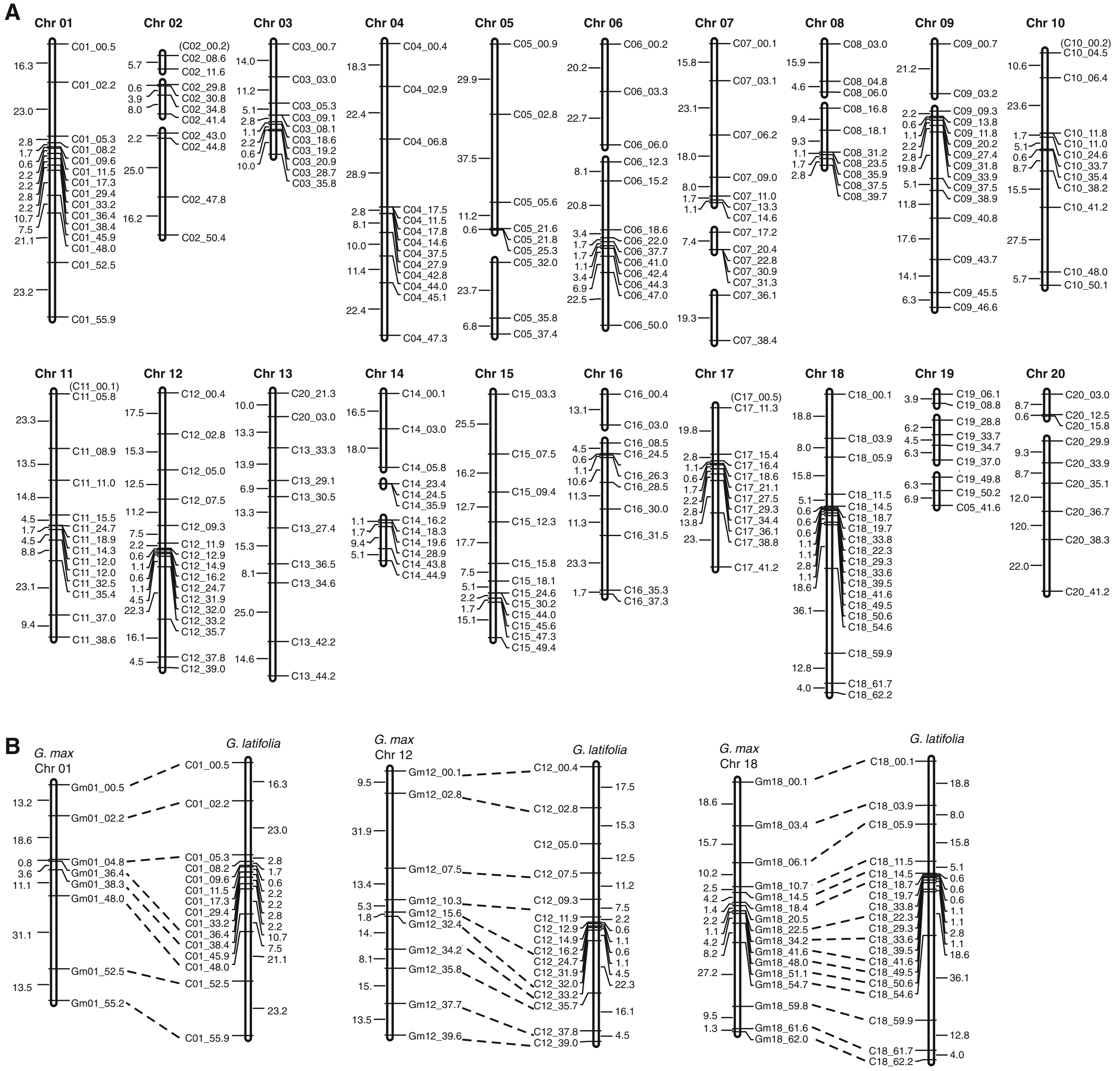

Fig. 3 *Glycine latifolia* linkage groups constructed from GoldenGate SNP markers that aligned to *G. max* chromosomes. The *G. max* chromosome (chr) to which the SNP marker sequences aligned is indicated above each linkage map. The genetic distances between markers for *G. latifolia*, indicated in Kosambi cM to the *left* of the maps, were estimated using a population of 91 F_2 *lines* from a cross between PI 559298 and PI 559300. Markers in parentheses were assigned to the termini of linkage groups, but map distances were not calculated. **a** Uninterrupted *G. latifolia* linkage maps were assembled that were nearly collinear with *G. max* chromosomes 1, 3, 4, 10, 11, 12, 15, 17, and 18. **b** The relative order of markers on three of the nine uninterrupted genetic linkage maps of *G. latifolia* were compared with those of previously mapped *G. max* SNPs at the closest corresponding nucleotide position in the *G. max* genome sequence. Markers were named for the *G. max* chromosome (C) and the nucleotide position ($\times 10^{-6}$) on the chromosome to which the SNP-containing sequences aligned

reported in *A. thaliana*, maize (*Zea mays* L.), rice (*Oryza sativa* L.), and tobacco (*Nicotiana tabacum* L.) (Kleine et al. 2009; Matsuo et al. 2005; Roark et al. 2010; Sheppard and Timmis 2009). For *Arabidopsis*, Huang (2003) estimated that 1 in 16,000 pollen grains contains a chloroplast-to-nuclear-genome transposition. After integration, most plastid sequences are rearranged or lost, but a few genes are stably integrated (Kleine et al. 2009).

Because only low levels of resistance to *S. sclerotiorum* are found in many domesticated crop species, efforts have been made to introgress resistance from secondary and tertiary gene pools. In addition to the wild relatives of soybean, significant levels of resistance to *S. sclerotiorum* have been reported in wild relatives of bean (*Phaseolus vulgaris* L.) (Abawi et al. 1978; Schwartz et al. 2006), canola (*Brassica napus* L.) (Li et al. 2007; Mei et al. 2011), pea (*Pisum sativum* L.) (Porter 2012; Porter et al. 2009), and sunflower (*Helianthus annuus* L.) (Cerboncini et al. 2002; Seiler 1992), which in some cases have been partially transferred to fertile hybrids (Garg et al. 2010; Ronicke et al. 2004). Even so, recovering full levels of resistance is impeded by the often polygenic nature of the phenotype, inefficient integration of alien DNA segments into adapted genomes, and difficulties in phenotypically identifying introgressed genes with small effects (Singh 2001). For example, when crosses were performed between *G. max* and a *G. tomentella* accession resistant to soybean cyst nematode and soybean rust, the resistance was retained in the ampidiploid progeny, but was lost during backcrossing (Bauer et al. 2007; Patzoldt et al. 2007; Riggs et al. 1998). By identifying chromosomal regions associated with resistance to *S. sclerotiorum* in wild relatives before hybridization, it may be possible to more precisely follow and combine alien DNA associated with resistance in progeny from wide crosses than could be accomplished by phenotypic selection alone. With sufficiently high density of molecular markers it may be possible to use map-based cloning to identify genes underlying agronomically important traits and move them by transformation to adapted plant materials without bringing in linked undesirable genes from the wild donor. However, resources for gene mapping are generally lacking in wild relatives of cultivated plants, as is the case with the perennial *Glycine* species. In this study, we identified SNPs in *G. latifolia*, a perennial relative of cultivated soybean by high-throughput sequencing of reduced representations of the genomic DNAs of two *G. latifolia* accessions. These results showed that the SNPs identified from the genome sequences of PI 559298 and PI 555300 will be useful for comparative gene mapping in *G. latifolia* and *G. max* and for moving agronomically valuable genes from a perennial *Glycine* species to cultivated soybean.

Acknowledgments This study was supported by funding from the National Sclerotinia Initiative and the United States Department of Agriculture-Agricultural Research Service. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the USDA or the University of Illinois and does not imply its approval to the exclusion of other products or vendors that may also be suitable.

References

Abawi GS, Provvidenti R, Crosier DC, Hunter JE (1978) Inheritance of resistance to white mold disease in *Phaseolus coccineus*. J Hered 69:200–202

Arahana VS, Graef GL, Specht JE, Steadman JR, Eskridge KM (2001) Identification of QTLs for resistance to *Sclerotinia sclerotiorum* in soybean. Crop Sci 41:180–188

Bauer S, Hymowitz T, Noel GR (2007) Soybean cyst nematode resistance derived from *Glycine tomentella* in amphiploid (*G. max* × *G. tomentella*) hybrid lines. Nematropica 37:277–285

Bronski MJ, Straub SCK, Bogdanowicz SM, Doyle JL, Brown AHD, Doyle JJ (2009) Isolation and characterization of thirteen polymorphic microsatellite loci in the A-genome perennial group of the legume genus *Glycine*. Molec Ecol Res 9: 1548–1550

Burdon JJ (1988) Major gene resistance to *Phakopsora pachyrhizi* in *Glycine canescens*, a wild relative of soybean. Theor Appl Genet 75:923–928

Burdon JJ, Marshall DR (1981) Evaluation of Australian native species of *Glycine* for resistance to soybean rust. Plant Dis 65:44–45

Cannon SB, May GD, Jackson SA (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol 151:970–977

Cerboncini C, Beine G, Binsfeld PC, Dresen B, Peisker H, Zerwas A, Schnabl H (2002) Sources of resistance to *Sclerotinia sclerotiorum* (Lib) de Bary in a natural *Helianthus* gene pool. Helia 25:167–176

Chesnay C, Kumar A, Pearce SR (2007) Genetic diversity of SIRE-1 retroelements in annual and perennial *Glycine* species revealed using SSAP. Cell Mol Biol Lett 12:103–110

Choi IY, Hyten DL, Matukumalli LK, Song QJ, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, Hwang EY, Yi SI, Young ND, Shoemaker RC, van Tassell CP, Specht JE, Cregan PB (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696

Chung G, Singh RJ (2008) Broadening the genetic base of soybean: a multidisciplinary approach. Crit Rev Plant Sci 27:295–341

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung L, Specht JE (1999) An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464–1490

Dinkins RD, Collins GB (2008) *Agrobacterium*-mediated genetic transformation of soybean. In: Kirti PD (ed) Handbook of New Technologies for Genetic Improvement of Legumes. CRC Press, Boca Raton, pp 89–102

Du JC, Grant D, Tian ZX, Nelson RT, Zhu LC, Shoemaker RC, Ma JX (2010a) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11:113

Du JC, Tian ZX, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma JX (2010b) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584–598

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Garg H, Atri C, Sandhu PS, Kaur B, Renton M, Banga SK, Singh H, Singh C, Barbetti MJ, Banga SS (2010) High level of resistance to *Sclerotinia sclerotiorum* in introgression lines derived from hybridization between wild crucifers and the crop *Brassica* species *B. napus* and *B. juncea*. Field Crops Res 117:51–58

Giesler LJ, Ziems AD (2006) Incidence of *Alfalfa mosaic virus*, *Bean pod mottle virus*, and *Soybean mosaic virus* in Nebraska soybean fields. Plant Health Prog. doi:10.1094/PHP-2006-0424-01-HM

Gill N, Findley S, Walling JG, Hans C, Ma JX, Doyle J, Stacey G, Jackson SA (2009) Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol 151:1167–1174

Grant JE (1990) Soybean: wide hybridisation through embryo culture. In: Bajaj YPS (ed) Biotechnology in Agriculture and Forestry 10

Legumes and Oilseed Crops I. Springer-Verlag, Newyork, pp 134–148

Guo XM, Wang DC, Gordon SG, Helliwell E, Smith T, Berry SA, Martin SKS, Dorrance AE (2008) Genetic mapping of QTLs underlying partial resistance to *Sclerotinia sclerotiorum* in soybean PI 391589A and PI391589B. Crop Sci 48:1129–1139

Hammatt N, Lister A, Jones B, Cocking EC, Davey MR (1992) Shoot formation from somatic hybrid callus between soybean and a perennial wild relative. Plant Sci 85:215–222

Hart SE, Glenn S, Kenworthy WW (1991) Tolerance and the basis for selectivity to 2,4-D in perennial *Glycine* species. Weed Sci 39:535–539

Hartman GL, Wang TC, Hymowitz T (1992) Sources of resistance to soybean rust in perennial *Glycine* species. Plant Dis 76:396–399

Hartman GL, Kull L, Huang YH (1998) Occurrence of *Sclerotinia sclerotiorum* in soybean fields in east-central Illinois and enumeration of inocula in soybean seed lots. Plant Dis 82:560–564

Hartman GL, Gardner ME, Hymowitz T, Naidoo GC (2000) Evaluation of perennial *Glycine* species for resistance to soybean fungal pathogens that cause Sclerotinia stem rot and sudden death syndrome. Crop Sci 40:545–549

Hempel K, Peakall R (2003) Cross-species amplification from crop soybean *Glycine max* provides informative microsatellite markers for the study of inbreeding wild relatives. Genome 46: 382–393

Horlock CM, Teakle DS, Jones RM (1997) Natural infection of the native pasture legume, *Glycine latifolia*, by alfalfa mosaic virus in Queensland. Australasian Plant Pathol 26:115–116

Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72–76

Huynh TT, Bastien M, Iquira E, Turcotte P, Belzile F (2010) Identification of QTLs associated with partial resistance to white mold in soybean using field-based inoculation. Crop Sci 50:969–979

Hyten DL, Cannon SB, Song QJ, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010a) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11:38

Hyten DL, Choi IY, Song QJ, Specht JE, Carter TE, Shoemaker RC, Hwang EY, Matukumalli LK, Cregan PB (2010b) A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci 50:960–968

Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A, Doyle JJ (2012) A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-Seq in plant species. Am J Bot 99:383–396

Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NWG, Couloux A, Dalwani A, Denny R, Deshpande S, Egan AN, Glover N, Hans CS, Howell S, Ilut D, Jackson S, Lai HS, Mammadov J, del Campo SM, Metcalf M, Nguyen A, O'Bleness M, Pfeil BE, Podicheti R, Ratnaparkhe MB, Samain S, Sanders I, Segurens B, Sevignac M, Sherman-Broyles S, Thareau V, Tucker DM, Walling J, Wawrzynski A, Yi J, Doyle JJ, Geffroy V, Roe BA, Maroof MAS, Young ND (2008) Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. Plant Physiol 148:1740–1759

Iwata H, Ninomiya S (2006) AntMap: constructing genetic linkage maps using an ant colony optimization algorithm. Breed Sci 56:371–377

Jarosz AM, Burdon JJ (1990) Predominance of a single major gene for resistance to *Phakopsora pachyrhizi* in a population of *Glycine argyrea*. Heredity 64:347–353

Jones RM, Brown AHD, Coote JN (1996) Variation in growth and forage quality of *Glycine latifolia* (Benth.) (Newell and Hymowitz). Genetic Resources Communication No. 26. Division of Tropical Crops and Pastures, CSIRO, Australia

Kim HS, Hartman GL, Manandhar JB, Graef GL, Steadman JR, Diers BW (2000) Reaction of soybean cultivars to sclerotinia stem rot in field, greenhouse, and laboratory evaluations. Crop Sci 40:665–669

Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, Kim DS, Lee YS, Park D, Ma J, Kim WY, Kim BC, Park S, Lee KA, Kim DH, Kim KH, Shin JH, Jang YE, Do Kim K, Liu WX, Chaisan T, Kang YJ, Lee YH, Kim KH, Moon JK, Schmutz J, Jackson SA, Bhak J, Lee SH (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc Natl Acad Sci USA 107: 22032–22037

Kleine T, Maier UG, Leister D (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Ann Rev Plant Biol 60:115–138

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Meth 9:357–359

Li CX, Li H, Siddique AB, Sivasithamparam K, Salisbury P, Banga SS, Banga S, Chattopadhyay C, Kumar A, Singh R, Singh D, Agnihotri A, Liu SY, Li YC, Tu J, Fu TF, Wang YF, Barbetti MJ (2007) The importance of the type and time of inoculation and assessment in the determination of resistance in *Brassica napus* and *B. juncea* to *Sclerotinia sclerotiorum*. Aust J Agric Res 58:1198–1203

Li DM, Sun MM, Han YP, Teng WL, Li WB (2010) Identification of QTL underlying soluble pigment content in soybean stems related to resistance to soybean white mold (*Sclerotinia sclerotiorum*). Euphytica 172:49–57

Lim SM, Hymowitz T (1987) Reactions of perennial wild species of genus *Glycine* to *Septoria glycines*. Plant Dis 71:891–893

Lin JY, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA (2005) Pericentromeric regions of soybean (*Glycine max* L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 170:1221–1230

Liu RH, Meng JL (2003) MapDraw: a Microsoft Excel macro for drawing genetic linkage maps based on given genetic linkage data. Heraditas 25:317–321

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet 23:134–139

Maheshwari S, Barbash DA (2011) The genetics of hybrid incompatibilities. Annu Rev Genet 45:331–355

Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. Plant Cell 17:665–675

Mei J, Qian L, Disi JO, Yang X, Li Q, Li J, Frauen M, Cai D, Qian W (2011) Identification of resistant sources against *Sclerotinia sclerotiorum* in *Brassica* species with emphasis on *B. oleracea*. Euphytica 177:393–399

Minoche A, Dohm J, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol 12:R112

Mueller EE, Grau CR (2007) Seasonal progression, symptom development, and yield effects of *Alfalfa mosaic virus* epidemics on soybean in Wisconsin. Plant Dis 91:266–272

Newell CA, Delannay X, Edge ME (1987) Interspecific hybrids between the soybean and wild perennial relatives. J Hered 78:301–306

Patzoldt ME, Tyagi RK, Hymowitz T, Miles MR, Hartman GL, Frederick RD (2007) Soybean rust resistance derived from *Glycine tomentella* in amphiploid hybrid lines. Crop Sci 47: 158–161

Porter LD (2012) Pea germplasm with partial resistance to *Sclerotinia sclerotiorum* extends the time required by the pathogen to infect host tissue. Crop Sci 52:1044–1050

Porter LD, Hoheisel G, Coffman VA (2009) Resistance of peas to *Sclerotinia sclerotiorum* in the *Pisum* core collection. Plant Pathol 58:52–60

Purdy LH (1979) *Sclerotinia sclerotiorum*: history, diseases and symptomatology, host range, geographic distribution, and impact. Phytopathology 69:875–880

Ratan A, Yu Z, Hayes VM, Schuster SC, Miller W (2010) Calling SNPs without a reference sequence. BMC Bioinfo 11:130

Ratnaparkhe MB, Singh RJ, Doyle JJ (2011) *Glycine*. In: Kole C (ed) Wild crop relatives: genomic and breeding resources: legume crops and forages. Springer, New York, pp 83–116

Rech EL, Vianna GR, Aragao FJL (2008) High-efficiency transformation by biolistics of soybean, common bean and cotton transgenic plants. Nat Protoc 3:410–418

Riggs RD, Wang S, Singh RJ, Hymowitz T (1998) Possible transfer of resistance to *Heterodera glycines* from *Glycine tomentella* to soybean. J Nematol 30:547–552

Roark LM, Hui Y, Donnelly L, Birchler JA, Newton KJ (2010) Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res 129:17–23

Ronicke S, Hahn V, Horn R, Grone I, Brahm L, Schnabl H, Friedt W (2004) Interspecific hybrids of sunflower as a source of *Sclerotinia* resistance. Plant Breeding 123:152–157

Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK (2005) Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. Plant Molec Biol 59:309–322

Schlueter JA, Lin JY, Schlueter SD, Vasylenko-Sanders IF, Deshpande S, Yi J, O'Bleness M, Roe BA, Nelson RT, Scheffler BE, Jackson SA, Shoemaker RC (2007) Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. BMC Genomics 8:330

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183

Schoen DJ, Burdon JJ, Brown AHD (1992) Resistance of *Glycine tomentella* to soybean leaf rust *Phakopsora pachyrhizi* in relation to ploidy level and geographic-distribution. Theor Appl Genet 83:827–832

Schwartz HF, Otto K, Terán H, Lema M, Singh SP (2006) Inheritance of white mold resistance in *Phaseolus vulgaris* × *P. coccineus* crosses. Plant Dis 90:1167–1170

Seiler GJ (1992) Utilization of wild sunflower species for the improvement of cultivated sunflower. Field Crops Res 30: 195–230

Sheppard AE, Timmis JN (2009) Instability of plastid DNA in the nuclear genome. PLoS Genet 5:e1000323

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Singh SP (2001) Broadening the genetic base of common bean cultivars: a review. Crop Sci 41:1659–1675

Singh RJ (2010) Methods for producing fertile crosses between wild and domestic soybean species. Patent No.: US 7,842,850

Singh RJ, Hymowitz T (1985) The genomic relationships among six wild perennial species of the genus *Glycine* subgenus *Glycine* Willd. Theor Appl Genet 71:221–230

Singh RJ, Kollipara KP, Hymowitz T (1988) Further data on the genomic relationships among wild perennial species ($2n = 40$) of the genus *Glycine* Willd. Genome 30:166–176

Singh RJ, Kollipara KP, Ahmad F, Hymowitz T (1992) Putative diploid ancestors of 80-chromosome *Glycine tabacina*. Genome 35:140–146

Singh RJ, Kollipara KP, Hymowitz T (1998) Monosomic alien addition lines derived from *Glycine max* (L) Merr and *G. tomentella* Hayata: production, characterization, and breeding behavior. Crop Sci 38:1483–1489

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. Am J Bot 96: 336–348

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109: 122–128

Vuong TD, Diers BW, Hartman GL (2008) Identification of QTL for resistance to Sclerotinia stem rot in soybean plant introduction 194639. Crop Sci 48:2209–2214

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V, Ameline-Torregrosa C, Cannon E, Chacko B, Couloux A, Dalwani A, Denny R, Deshpande S, Egan AN, Glover N, Howell S, Ilut D, Lai HS, del Campo SM, Metcalf M, O'Bleness M, Pfeil BE, Ratnaparkhe MB, Samain S, Sanders I, Segurens B, Sevignac M, Sherman-Broyles S, Tucker DM, Yi J, Doyle JJ, Geffroy V, Roe BA, Maroof MAS, Young ND, Innes RW (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. Plant Physiol 148:1760–1771

Wu XL, Ren CW, Joshi T, Vuong T, Xu D, Nguyen HT (2010) SNP discovery by high-throughput sequencing in soybean. BMC Genomics 11:469

Zou JJ, Singh RJ, Hymowitz T (2004) SSR marker and ITS cleaved amplified polymorphic sequence analysis of soybean x *Glycine tomentella* intersubgeneric derived lines. Theor Appl Genet 109:769–774