RESEARCH

# A High Density Integrated Genetic Linkage Map of Soybean and the Development of a 1536 Universal Soy Linkage Panel for Quantitative Trait Locus Mapping

David L. Hyten,* Ik-Young Choi, Qijian Song, James E. Specht, Thomas E. Carter, Jr., Randy C. Shoemaker, Eun-Young Hwang, Lakshmi K. Matukumalli, and Perry B. Cregan

**ABSTRACT**

Single nucleotide polymorphisms (SNPs) are the marker of choice for many researchers due to their abundance and the high-throughput methods available for their multiplex analysis. Only recently have SNP markers been available to researchers in soybean [*Glycine max* (L.) Merr.] with the release of the third version of the consensus genetic linkage map that added 1141 SNP markers to the map. Our objectives were to add 2500 additional SNP markers to the soybean integrated map and select a set of 1536 SNPs to create a universal linkage panel for high-throughput soybean quantitative trait locus (QTL) mapping. The GoldenGate assay is one high-throughput analysis method capable of genotyping 1536 SNPs in 192 DNA samples over a 3-d period. We designed GoldenGate assays for 3456 SNPs (2956 new plus 500 previously mapped) which were used to screen three recombinant inbred line populations and diverse germplasm. A total of 3000 workable assays were obtained which added about 2500 new SNP markers to create a fourth version of the soybean integrated linkage map. To create a "Universal Soy Linkage Panel" (USLP 1.0) of 1536 SNP loci, SNPs were selected based on even distribution throughout each of the 20 consensus linkage groups and to have a broad range of allele frequencies in diverse germplasm. The 1536 USLP 1.0 will be able to quickly create a comprehensive genetic map in most QTL mapping populations and thus will serve as a useful tool for high-throughput QTL mapping.

D.L. Hyten, I-Y. Choi, Q. Song, E-Y Hwang, and P.B. Cregan, USDA-ARS, Soybean Genomics and Improvement Lab., Beltsville, MD 20705. Q. Song and E-Y. Hwang, Dep. of Plant Science and Landscape Architecture, Univ. of Maryland, College Park, MD 20742. J.E. Specht, Dep. of Agronomy and Horticulture, Univ. of Nebraska, Lincoln, NE 68583. T.E. Carter, USDA-ARS, Soybean and Nitrogen Fixation Research, Raleigh, NC 27607. R.C. Shoemaker, USDA-ARS, Dep. of Agronomy, Iowa State Univ., Ames, IA 50011. L.K. Matukumalli, USDA-ARS, Bovine Functional Genomics Lab., Beltsville, MD 20705. I-Y. Choi, current address: National Instrumentation Center for Environmental Management, Seoul National Univ., Seoul, 151-921, South Korea. Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval of a product to the exclusion of others that may be suitable. Received 30 June 2009. *Corresponding author (David.Hyten@ars.usda.gov).

**Abbreviations:** BARC, Beltsville Agricultural Research Center; BSA, bulked segregant analysis; cM, centimorgan; INDEL, insertion/deletion; LG, linkage group; OPA, oligo pool all; QTL, quantitative trait locus; RIL, recombinant inbred line; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; STS, sequence tagged site; USLP 1.0, Universal Soy Linkage Panel 1.0.

The first report of the polymorphism and heritability of simple sequence repeat (SSR) markers in plants was in soybean (Akkaya et al., 1992). Subsequently, 606 soybean SSR loci were developed and mapped in three mapping populations to create the first version of a soybean integrated genetic linkage map (Cregan et al., 1999a). Five years later, a second version of the integrated genetic linkage map was published which was based on five mapping populations

Published in Crop Sci. 50:960–968 (2010).
doi: 10.2135/cropsci2009.06.0360
Freely available online through the author-supported open-access option.
Published online 9 Mar. 2010.
© Crop Science Society of America | 5585 Guilford Rd., Madison, WI 53711 USA

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

and contained a total of 1015 SSR loci (Song et al., 2004). More recently, single DNA base differences between homologous DNA fragments, plus small insertions and deletions (INDELs), collectively referred to as SNPs, were shown to be the most abundant source of DNA polymorphisms in soybean (Hyten et al., 2006; Zhu et al., 2003).

Despite being the most common molecular marker in soybean, the SNP frequency is relatively low compared to other cultivated crop species (Hyten et al., 2006; Zhu et al., 2003). The relatively low sequence variation in *G. max* can be attributed to the domestication bottleneck which reduced variation by 50% and the low sequence variation in the wild ancestor of soybean, *G. soja* (Seib. et Zucc.) (Hyten et al., 2006). Even so, Choi et al. (2007) was able to successfully discover approximately 5500 SNPs in 2032 gene transcripts, and mapped at least one SNP from 1141 of those transcripts to create what is now called the version three soybean integrated linkage map. This version three map filled 72 of the 112 gaps of 5 to 10 centimorgans (cM) that were present in the version two integrated genetic linkage map. However, despite the current availability of more than 2000 PCR based markers on the version three map, the marker density in any one population is likely to be inadequate to allow a thorough scan of the genome for purposes of QTL discovery, association analysis, map-based cloning, and the integration of the physical map and whole genome sequence with the genetic map.

A unique feature of SNP markers is the modest cost per data point and speed of data acquisition (Fan et al., 2006). One versatile SNP detection system is the GoldenGate assay (Illumina Inc., San Diego, CA). Recently, the GoldenGate assay was tested in soybean for its efficiency of SNP detection. In an initial trial of 384 soybean SNPs using the GoldenGate assay, Hyten et al. (2008) observed that 90% of the SNPs in the array produced high quality data despite the fact that the soybean is a diploidized ancient tetraploid that retains many homeologous and other duplicated genomic segments (Shoemaker et al., 1996).

The high-throughput nature of the GoldenGate assay makes it a good platform for genotyping biparentally-derived populations used in QTL mapping. For the ~2500-cM genetic map of soybean, a QTL detection study would require about 125 to 250 markers uniformly distributed at 10- or 20-cM intervals on each linkage group. Through the use of gel-based SSR genotyping or SNP analysis at low multiplex levels this can be a very labor intensive and time consuming process when dealing with large numbers of progeny. In contrast, the GoldenGate assay can be configured to include 1536 SNP markers that have been preselected to have a high probability of polymorphism in a targeted range of germplasm to produce a sufficient number of polymorphic markers in most QTL mapping populations to obtain the same or better coverage than is currently obtained in most soybean QTL studies.

Our first objective was to use the GoldenGate assay to map an additional 2500 SNPs and thereby augment the third version of the soybean integrated linkage map (Choi et al., 2007). Using a combination of the new and previously developed GoldenGate assays, our second objective was to identify 1536 SNPs that are well distributed throughout the genome, have an optimal minor allele frequency in both exotic and adapted soybean germplasm, and can be scored accurately using automatic calling programs. This set of 1536 SNPs would then be used to construct a Universal Soy Linkage Panel (USLP 1.0) which could adequately genotype most QTL mapping populations for high-throughput QTL discovery.

## MATERIALS AND METHODS

### Plant Materials

Mapping populations: Three mapping populations were used to position newly identified SNPs on the consensus genetic map (Choi et al., 2007). The University of Utah 'Minsoy' × 'Noir 1' (M×N) and Minsoy × 'Archer' (M×A) populations were described by Cregan et al. (1999a), Mansur and Orf (1995), and Mansur et al. (1996) and the University of Minnesota 'Evans' × 'Peking' (ExP) population was described by Cregan et al. (1999b) and Concibido et al. (1997). Genomic DNA was obtained from each of the two parents and each of the 89 recombinant inbred lines (RILs) of the three mapping populations from bulked leaf tissue of 20 to 50 plants of each line using the method described by Keim et al. (1988). The M×N and M×A populations had been used in creating the current consensus genetic linkage map (Choi et al., 2007), while the E×P population was recently integrated into the linkage map (Hyten et al., 2008).

### Diverse Germplasm

A group of 96 diverse landrace accessions were selected that consisted of plant introductions from China, Korea, and Japan (Hyten et al., 2008). The landraces represent a range of geographic origin, morphological descriptors, and maturity classes to maximize the diversity sampled. Pure line seeds of the 96 landrace genotypes were obtained from the USDA Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, Univ. of Illinois, Urbana). In addition, a group of 96 elite cultivars were selected to represent the diversity extant in North American breeding programs. The 96 cultivars were chosen from among the 209 public commodity cultivars released during 1990 thru 2000. All 209 cultivars were first grouped by analysis of coefficient of parentage as described by Gizlice et al. (1996) and then a subset of 96 cultivars was selected to represent the groupings to maximize the diversity sampled (Supplemental Table 1). Pure line seeds of all elite cultivar genotypes were obtained either from the USDA Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, Univ. of Illinois, Urbana), or from the individual breeder as listed in Supplemental Table 1. DNA was extracted from bulked leaf tissue of 20 to 50 plants of each landrace and elite cultivar as described by Keim et al. (1988).

## Single Nucleotide Polymorphism Discovery

A total of 14,527 PCR primer pairs were designed from (i) gene or EST sequences deposited in GenBank along with 3'-end soybean unigene sequences deposited in GenBank by Vodkin et al. (2004), (ii) BAC-end sequences deposited in GenBank as a result of the Williams 82 physical map project (http://www.soybase.org), (iii) BAC-end sequences deposited in GenBank as a result of the Forrest minimum tile path (Saini et al., 2008; Shultz et al., 2007), and (iv) targeted BAC subclone sequences developed as described by Cregan et al. (1999c). These PCR primers were designed using Primer 3 software (Rozen and Skaletsky, 2000) to amplify products in the 300 to 500 bp range for EST sequences, and in the 600 to 800 bp range for genomic sequences. DNA from the cultivar 'Archer' was amplified by all primer pairs to initially screen the primer pairs for a single amplicon, which if generated, was then sequenced using Big-Dye 3.1 (Applied Biosystems, Foster City, CA) on an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, CA) as described by Choi et al. (2007). When high quality sequence was obtained from primer pairs, indicating a robust sequence tagged site (STS), they were used to amplify DNA from each of the parents of the three RIL mapping populations: Archer, Minsoy, Noir 1, Evans, and Peking. The resulting amplicons were sequenced as described above followed by alignment and SNP discovery via analysis using SNP-PHAGE software (Matukumalli et al., 2006). All STS primer sequences and amplification condition information from this study and Choi et al. (2007) were deposited in the NCBI UniSTS database (www.ncbi.nlm.nih.gov) (accessions GF091800-GF097726). Likewise, all SNPs identified from this study and Choi et al. (2007) were deposited in the NCBI database dbSNP (www.ncbi.nlm.nih.gov/SNP/) (accessions ss107912501-ss107931092).

## GoldenGate Single Nucleotide Polymorphism Mapping

A total of 2572 bi-allelic SNPs that were polymorphic between the parents of one or more of the RIL mapping populations were chosen for the design of two 1536 custom GoldenGate assays. To complete the complement of 3072 SNPs needed to design two 1536 custom GoldenGate assays, an additional 500 bi-allelic SNP loci were selected from SNPs that had previously been mapped by Choi et al. (2007).

The SNP-containing sequences were screened with RepeatMasker software (www.repeatmasker.org), using repeat databases obtained from Dr. Randy Shoemaker of the USDA-ARS, Ames, IA; Dr. Scott Jackson of Purdue Univ., West Lafayette, IN (http://www.soymap.org/); and Dr. Gary Stacy of the Univ. of Missouri, Columbia (http://www.soybeangenome.org/). Repeats contained within the SNP-containing sequences were replaced with lowercase letters before submission to Illumina Inc. to undergo a preliminary design phase of the custom oligo pool all (OPA), which contains the allele-specific oligos and the locus-specific oligos for each of the SNPs included in the assay. A designability rank score ranging from 0 to 1.0 was given to each SNP by Illumina using an in-house assay design tool, where a rank score of <0.4 has a low success rate, 0.4 to 0.6 has a moderate success rate, and >0.6 has a high success rate for the conversion of a SNP into a successful GoldenGate assay. A total of two OPA sets (SoyOPA-2 and SoyOPA-3) were selected which consisted of 1536 SNPs each, SoyOPA-2 had a designability rank score of 0.6 or higher for each SNP while SoyOPA-3 had a designability rank score of 0.4 to <0.6 for 290 SNPs with the remaining SNPs having a designability rank score of 0.6 or higher. When multiple SNPs were located in the same STS, only the SNP with the highest designability rank score was selected for inclusion (the others were not used).

The GoldenGate assay was first tested on the M×A, M×N, and E×P RIL mapping populations and then tested on the 96 diverse landraces and 96 elite cultivars as per the manufacturer's protocol and as described in Fan et al. (2003) and Hyten et al. (2008). The Illumina BeadStation 500G (Illumina Inc., San Diego, CA) was used for genotyping the GoldenGate assay. The automatic allele calling for each locus and the calculation of minor allele frequencies were accomplished with the GenCall software (Illumina Inc., San Diego, CA). All GenCall data were visually inspected and rescored if any errors in calling the homozygous or heterozygous clusters were evident.

## Consensus Map Construction

The soybean SNP markers from SoyOPA-1 developed by Hyten et al. (2008), and those from SoyOPA-2 and SoyOPA-3 were first mapped using JoinMap 3.0 software (Van Ooijen and Voorrips, 2001) within each of the three mapping populations (M×A, M×N, and E×P) to assign a soybean consensus linkage group to all markers which corresponds to the linkage groups of the previous versions of the integrated map (Choi et al., 2007). A minimum LOD ≥ 3.0 and a maximum distance ≤50 cM were used to test linkages among markers. JoinMap 3.0 software was then used for map integration, one linkage group at a time, due to computational requirements of the large dataset. The datasets for map integration included all markers mapped in the three mapping populations from this study and previous studies (Choi et al., 2007; Cregan et al., 1999a; Hyten et al., 2008; Song et al., 2004) plus markers previously mapped in two additional RIL populations used by Choi et al. (2007). The complete dataset consisted of 363 markers mapped in Evans × PI 209332 (E×PI), 999 in A81–356022 × PI 468916 (A×PI), 1911 in E×P, 2354 in M×A, and 3005 in M×N (Supplemental Table 2). Genetic distances were calculated using the Kosambi mapping function.

Eleven of the 20 previously identified linkage groups (Cregan et al., 1999a) had been assigned chromosome numbers by Zou et al. (2003). To facilitate the association of (and eventual replacement of) linkage group alphanumeric names with the Arabic numbers of the corresponding chromosomes, the remaining nine linkage groups were associated with newly assigned chromosome numbers. This was accomplished by assigning the longest to shortest linkage groups (in terms of centiMorgan length) to the lowest to highest chromosome numbers which were still available (chromosomes 2, 6, 7, 10, 11, 12, 14, 15, and 16) in an effort to establish an inverse relationship between linkage group length and chromosome number. These nine newly assigned chromosome numbers along with the 11 previously assigned numbers correspond to the U.S. Department of Energy, Joint Genome Institute (DoE, JGI) soybean 8x whole genome sequence (Glyma1.01) (Schmutz et

al., 2010). Details of the linkage group to chromosome assignments can be found at the SoyBase web site (http://soybase.org/LG2Xsome.php).

### Statistical Analysis

The average number and standard deviation of SNPs expected to be polymorphic in the 1536 USLP 1.0 for all biparental matings involving the various combinations of the 96 elite cultivars genotypes and 96 landrace genotypes, either overall, or clustered into specific maturity groups were calculated as

$$\bar{x} = \sum X_{ij};\ \mathrm{sd} = \sqrt{(\sum X_{ij}^2 - (\sum X_{ij})^2 / f))/(\sum f - 1)}$$

where $\bar{x}$ and sd are mean and standard deviation, respectively, $X_{ij}$ is the number of SNPs polymorphic between the *i*th and the *j*th genotypes, *f* is the total number of pairs of genotypes. The confident interval (L1–L2) was estimated as $L1 = \bar{x} - t_{\alpha,\upsilon} \times \mathrm{sd}/\sqrt{f}$ and $L2 = \bar{x} + t_{\alpha,\upsilon} \times \mathrm{sd}/\sqrt{f}$, where the $t_{\alpha,\upsilon}$ is the t value when degree of freedom is equal to $\upsilon$ and the probability (two tails) is $\alpha$.

## RESULTS

A total of 3682 primer pairs of the 14,527 primer pairs tested produced robust STS that contained at least one SNP. The 3682 SNP-containing STS contained a total of 13,042 SNPs for an average of 3.5 SNPs per polymorphic STS. These SNPs were combined with the 5551 SNPs discovered by Choi et al. (2007) to design two Illumina custom 1536 SNP GoldenGate assays designated SoyOPA-2 and SoyOPA-3. These two OPAs consisted of 3072 SNPs each of which was discovered in a separate STS consisting of (i) 1428 random gene sequences of which 518 had previously been mapped by Choi et al. (2007), (ii) 1361 BAC-end sequences, and (iii) 283 targeted BAC subclone sequences. All SNPs chosen for SoyOPA-2 and SoyOPA-3 had been verified through forward and reverse sequencing of PCR amplicons from the five diverse genotypes which are mapping parents for the M×N, M×A, and E×P RIL mapping populations. Of the 3072 SNPs contained on SoyOPA-2 and SoyOPA-3, SNP allelic data were obtained for 2766 of the GoldenGate SNP assays. Of the 290 SNPs that had an Illumina designability score of 0.4 to <0.6, 19% failed to give a good assay while only 9% of the remaining 2782 SNPs which had a SNP designability score of ≥0.6 failed. Since the original design and testing of SoyOPA-2 and SoyOPA-3, the soybean whole genome sequence (Glyma1.01) (Schmutz et al., 2010) has become available for Illumina to incorporate into its assay design tool. Using the updated assay design tool, the 3072 SNPs were given new designability scores. A total of 199 SNPs had a designability score of 0.0 to <0.4 of which 33% failed to produce a good assay. The number of SNPs in the 0.4 to <0.6 increased to 606 SNPs of which 14.5% failed to produce a good assay. Of the remaining 2267 SNPs with a designability score of ≥0.6, only 6.7% failed to produce a good assay. The data from SoyOPA-1 from Hyten et al. (2008) and SoyOPA-2 and SoyOPA3 were combined for all subsequently described analyses.

### Genetic Mapping of Single Nucleotide Polymorphism Loci

In total, the three SoyOPAs included 2651 new SNP markers which were segregating in one or more of the M×N, M×A, or E×P mapping populations. A fourth version of the soybean integrated genetic linkage map (Consensus Map 4.0) was created by compositing the SNP locus data of the E×PI and A×PI mapping populations used by Choi et al. (2007) with the SNP locus data of and the M×N, M×A, and E×P mapping populations to create an integrated genetic linkage map of 5500 markers spanning a genomic map distance of 2296.4 cM (Table 1 and Supplemental Table 2). The mean chromosome length was 114.8 cM, with a mean genetic distance of 0.6 cM between any consecutive pair of mapped SNP markers.

The 3792 mapped SNPs also allow the placement of additional SNPs due to the presence of multiple SNPs that were discovered on the same STS. The mapping of one SNP per STS (which ranged in size from 300 to 800 bp) allowed for an average of 3.3 SNPs to be placed on the Consensus Map 4.0. Thus, a total of 12,586 SNPs of the 18,593 discovered in this project and discovered by Choi et al. (2007) have now been directly or indirectly positioned on the Consensus Map 4.0 (Supplemental Table 3).

There were a total of 18 gaps in the Consensus Map 4.0 that were 5 to 10 cM in length and only one gap of >10 cM. When the RFLP markers are removed from the Consensus Map 4.0 there are a total of 30 gaps between 5 and 10 cM and 3 gaps >10 cM.

### Selection of 1536 Universal Soy Linkage Panel

A set of 1536 SNPs from the 3456 SNPs present in the three SoyOPAs was selected to include a sufficiently large number of polymorphic SNPs to create a comprehensive genetic map that could be used for most QTL mapping applications. The SNPs chosen for this 1536 Universal Soy Linkage Panel (USLP 1.0) had to meet the following criteria. Each SNP was required to have a working GoldenGate assay with a cluster separation >2 as defined by Hyten et al. (2008). All SNPs tested had to be biallelic and the least frequent allele in a given population, termed, minor allele frequency, was calculated for the diverse landraces and the elite cultivars. The SNPs that segregated within 0.2 cM of each other were required to be in a different minor allele frequency class (0.00–0.10, 0.11–0.20, 0.21–0.30, 0.31–0.40, and 0.41–0.50) within the 96 diverse landrace vs. the 96 elite germplasm accessions. If a SNP had a minor allele frequency below 0.1 in both the landraces and the elite cultivars it was not selected unless it was the only SNP present within a 0.5 cM interval.

Table 1. Summary data for Soybean Consensus Map 4.0, including chromosome lengths, numbers of single nucleotide polymorphism (SNP), simple sequence repeat (SSR), restriction fragment length polymorphism (RFLP), and other marker types (which include random amplified polymorphic DNA [RAPD], amplified fragment length polymorphism [AFLP], and classical and isozyme marker loci).

| Chromosome (Linkage Group) | Length (cM) | SNP | SSR | RFLP | Other | Total markers |
|---|---|---|---|---|---|---|
| 1 (D1a) | 98.4 | 134 | 53 | 34 | 5 | 226 |
| 2 (D1b)† | 140.6 | 171 | 58 | 19 | 1 | 249 |
| 3 (N) | 99.5 | 173 | 45 | 28 | 3 | 249 |
| 4 (C1) | 112.3 | 190 | 43 | 21 | 4 | 258 |
| 5 (A1) | 86.7 | 186 | 50 | 33 | 0 | 269 |
| 6 (C2)† | 136.5 | 224 | 53 | 36 | 0 | 313 |
| 7 (M)† | 135.1 | 169 | 51 | 24 | 1 | 245 |
| 8 (A2) | 146.7 | 271 | 64 | 46 | 3 | 384 |
| 9 (K) | 99.6 | 192 | 59 | 23 | 3 | 277 |
| 10 (O)† | 132.9 | 188 | 56 | 31 | 0 | 275 |
| 11 (B1)† | 124.2 | 107 | 35 | 32 | 1 | 175 |
| 12 (H)† | 120.5 | 151 | 38 | 38 | 3 | 230 |
| 13 (F) | 120.0 | 230 | 63 | 35 | 3 | 331 |
| 14 (B2)† | 108.2 | 152 | 35 | 41 | 2 | 230 |
| 15 (E)† | 99.9 | 205 | 42 | 41 | 4 | 292 |
| 16 (J)† | 92.3 | 197 | 47 | 40 | 2 | 286 |
| 17 (D2) | 119.2 | 192 | 60 | 19 | 2 | 273 |
| 18 (G) | 109.9 | 308 | 61 | 49 | 0 | 418 |
| 19 (L) | 101.1 | 149 | 52 | 43 | 1 | 245 |
| 20 (I) | 112.8 | 203 | 41 | 31 | 0 | 275 |
| Total | 2296.4 | 3792 | 1006 | 664 | 38 | 5500 |

†Previously unassigned chromosome number. Longest unassigned linkage group was assigned to the smallest unassigned chromosome number.

To determine which SNPs had good working assays, the data from the three SoyOPAs were pooled to allow composite analysis in GenCall for the three mapping populations (M×N, M×A, and E×P), the 96 diverse landraces accessions, and the 96 diverse elite cultivars. A total of 3049 SNPs produced successful assays in the three mapping populations, the landraces and the elite cultivars. There was an even distribution of the minor allele frequencies in the diverse landrace populations (Fig. 1). However, in the diverse set of 96 elite cultivars, the proportion of SNPs with a minor allele frequency between 0.0 to 0.1 was about two times larger than the other minor allele frequency classes (Fig. 1). A total of 373 SNPs had a cluster separation score below 2 and were eliminated from consideration for the USLP 1.0. Of the remaining 2676 SNPs, an additional 343 were eliminated because their minor allele frequency was <0.1 for both the landraces and the elite cultivars, and there was at least one SNP within 0.4 cM with an allele frequency >0.1 in either the landraces or the elite cultivars. The map positions and minor allele frequency classes of the remaining 2333 SNPs were used to further select SNPs for inclusion in the USLP 1.0. When multiple SNPs were within 0.2 cM of each other and were also in the same minor allele frequency class in both the landraces and elite cultivars, only one SNP was retained for inclusion in the 1536 USLP 1.0. This final step eliminated 797 SNPs, and the remaining 1536 were used to construct the 1536 USLP 1.0 (Supplemental Table 4).

The distribution of the 1536 SNP loci in USLP 1.0 relative to their genetic map positions is shown in Fig. 2. The distribution is not perfectly uniform, but this was not unexpected, given that only 3456 SNPs were available to choose from the three OPAs. There are 16 gaps >10 cM in length in the USLP 1.0 SNP map. Of note is the fact that SSR markers are available in 12 of these 16 gaps (Fig. 2).

To determine how many of the 1536 SNPs might be expected to be polymorphic in an average biparental mapping population, a complete pairwise comparison of the 192 (96 landrace genotypes plus 96 elite cultivars) accessions revealed that, on average, 458 SNPs could be expected to be polymorphic in populations created by mating any two elite cultivars (Table 2). The number of polymorphic markers increases to 544 SNPs between any two mated landraces, and to 590 SNPs polymorphic between any mating involving a landrace to an elite cultivar (Table 2).

## DISCUSSION

The SNP discovery in soybean has been shown to be a difficult and tedious process. Two contributing factors are the low sequence diversity that is present in cultivated soybean, and the complex nature of its genome, wherein the latter can pose difficulties in creating robust STS (Choi et al., 2007; Hyten et al., 2006). Despite these problems, we have created a new integrated Consensus Map 4.0 which

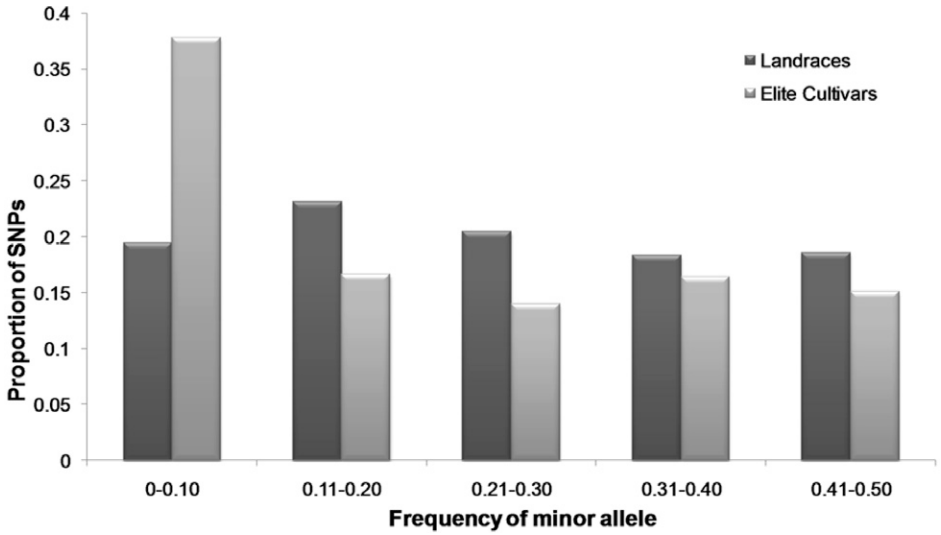


Fig. 1. The bars in this graph display the proportions of the 3049 SNPs in each minor allele frequency class with respect to the 96 diverse landraces and the 96 diverse elite cultivars.

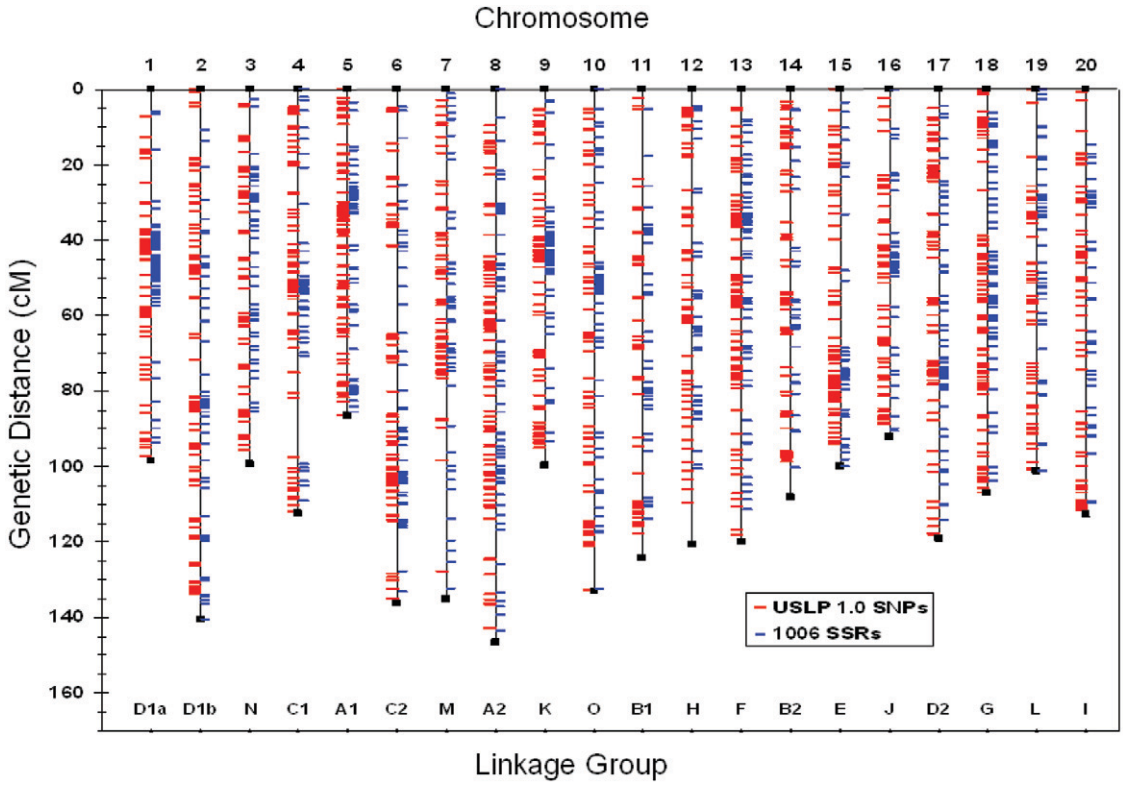


Fig. 2. The vertical lines and tickmarks in this graph represent the distribution within each of the 20 soybean chromosomes (top) and corresponding linkage groups (bottom) of the 1536 SNP markers (leftward ticks) comprising the Universal Soy Linkage Panel 1.0. Also shown is the distribution of the 1006 SSR markers (rightward ticks) mapped in the soybean Consensus Map 4.0. The vertical map distance is scaled in Kosambi centiMorgans.

adds 2651 new SNP markers to create a genetic map with a total of 5500 genetic markers.

The 90% success rate reported in the present study for converting validated SNPs into working GoldenGate assays in SoyOPA-2 and SoyOPA-3 did not differ appreciably from the 89% success rate reported by Hyten et al. (2008) with SoyOPA-1. Since SoyOPA-1 is a 384-plex GoldenGate assay, whereas SoyOPA-2 and SoyOPA-3 are each 1536-plex GoldenGate assays, the similarity in success rates indicated that higher multiplex levels of the GoldenGate assay do not reduce the proportion of successful assays. The availability of a whole genome sequence for Illumina to use in their assay design tool would appear to slightly increase the success rate of converting soybean SNPs into workable GoldenGate assays. The new designability rank score assigned by the current assay design tool

Table 2. Means and the low and high 95% confidence interval limits for the number of single nucleotide polymorphisms (SNPs) in the 1536-SNP Universal Soy Linkage Panel 1.0 that can be expected to be polymorphic for the biparental matings involving various combinations of the 96 elite cultivars genotypes and 96 landrace genotypes, either overall, or clustered into specific maturity groups (MGs).

| Genotypic cluster evaluated | Total number of genotypes in the cluster | Mean number of SNPs polymorphic between all possible genotypic pairs | Low-high 95% confidence interval limits |
|---|---|---|---|
| Elite | 96 | 458 | 324–591 |
| Landrace | 96 | 544 | 355–732 |
| Elite vs. Landrace | 96 vs. 96 | 590 | 454–724 |
| Elite north | 75 | 443 | 320–566 |
| Elite south | 21 | 391 | 262–518 |
| Elite north vs. Elite south | 75 vs. 21 | 493 | 372–612 |
| Elite MG-00 | 10 | 424 | 337–511 |
| Elite MG-0 | 8 | 448 | 343–552 |
| Elite MG-1 | 13 | 418 | 317–518 |
| Elite MG-2 | 18 | 429 | 327–529 |
| Elite MG-3 | 16 | 377 | 259–494 |
| Elite MG-4 | 17 | 417 | 302–531 |
| Elite MG-5 | 6 | 345 | 251–437 |
| Elite MG-7 | 6 | 375 | 289–460 |
| Landrace MG-0 | 4 | 586 | 460–711 |
| Landrace MG-1 | 8 | 625 | 477–773 |
| Landrace MG-2 | 12 | 596 | 444–747 |
| Landrace MG-3 | 13 | 562 | 409–714 |
| Landrace MG-4 | 10 | 573 | 443–701 |
| Landrace MG-5 | 8 | 593 | 395–790 |
| Landrace MG-6 | 12 | 464 | 303–625 |
| Landrace MG-7 | 8 | 464 | 290–637 |
| Landrace MG-8 | 8 | 445 | 182–708 |
| Landrace MG-9 | 8 | 444 | 330–558 |
| Landrace MG-10 | 3 | 442 | 263–620 |

of ≥0.6 would increase assay conversion from 91 to 93.3%. It would also seem to increase the success rate of the SNPs with a designability rank score of 0.4 to <0.6 by 4.5%.

The previous version of the soybean consensus map had a total genetic distance of 2550.3 cM which is 253.9 cM longer than Consensus Map 4.0 total length of 2296.4 cM. One possible reason for this map contraction is the accuracy of SNP genotyping when using the GoldenGate assay, which has been shown to be 99.9% accurate (Fan et al., 2003). Errors in marker scoring can lead to map expansions by creating false recombination events. The contraction of map size could also have been due to increasing the total number of markers from 2989 to 5500 which helped to eliminate large linkage gaps, hence obtaining a more accurate estimation of map distance across these previously large linkage gaps. The previous map had seven gaps of >10 cM in length, and 40 gaps of 5 to 10 cM in total length (Choi et al., 2007). The Consensus Map 4.0 contains only one gap greater than 10 cM and just 18 gaps of 5 to 10 cM in total length.

The mapping of an additional 2651 new STS-based SNP markers was also quite useful in the assembly of the DoE-Joint Genome Institute whole-genome shotgun sequence (Glyma1.01) of soybean into pseudo molecule chromosomes (J. Schmutz, personal communication, 2009). Each of the mapped STS markers had a sequence size ranging from approximately 300 to 800 bp which was used by the DoE-JGI soybean genome sequencing consortium to anchor and orient sequence scaffolds relative to the Consensus Map 4.0 (Schmutz et al., 2010).

The creation of a set of 1536 SNPs selected for an optimum allele frequency in diverse germplasm provides a powerful and versatile tool that can be used by the soybean research community to quickly map genes or QTLs governing important qualitatively or quantitatively inherited traits. For example, the GoldenGate assay was coupled with bulked segregant analysis (BSA) to map the soybean rust resistance gene *Rpp3* using SoyOPA-2 (Hyten et al., 2009). Another application of the USLP 1.0 would be to accelerate the pace of recurrent parent recovery by identifying which of the many backcross two or -three generation progeny should be used as parents in the next backcross mating cycle. The large proportion of the 1536 SNPs in the USLP 1.0 which is expected to be polymorphic between any two parents allows it to be used on a wide array of biparental populations. As suggested by data presented in Fig. 2, many SNPs in the USLP 1.0 are clustered, but the SNPs within clusters have different

allele frequencies, which increases the likelihood of different SNPs being polymorphic in different matings as indicated by the high polymorphism values across all maturity groups reported in Table 2. The high-throughput nature of this assay also allows for significantly larger QTL populations to be genotyped with sufficient marker coverage to increase the power to detect QTL and minimize bias estimates of QTL effects found using QTL populations with a small number of lines (Beavis, 1994).

While the number of polymorphic SNPs in the USLP 1.0 in any given cross may be high there will likely be a few regions of >20 cM without a polymorphic SNP which will reduce the power to detect QTL in these regions. In such cases the region may be identical by descent between the two parents in which case no markers would be available and likely no QTL would be segregating. Another possibility is the regions are not identical by descent and a researcher may test other available markers from the Consensus Map 4.0 in these regions to determine if any of the available markers are polymorphic and use these markers to eliminate gaps not covered by the USLP 1.0.

One main advantage of creating a universal linkage panel that can be widely used for QTL analysis and other genotyping applications is the savings gained in datapoint acquisition time and cost. Significant reductions in cost can be achieved by placing large volume orders of the GoldenGate reagents. For example, recently, a collaborative group of soybean researchers purchased sufficient reagents and OPA of the USLP 1.0 to genotype ~17,000 DNA samples. This large order reduced the reagent cost per datapoint by more than half. The USLP 1.0 has already been used to genotype more than 8800 of the 17,000 DNA samples with one technician being able to genotype 768 DNA samples per week from numerous populations for quantitative traits such as resistance to soybean rust, sudden death syndrome, soybean cyst nematode and iron deficiency chlorosis, as well as the seed quality traits of oleic acid concentration and protein/oil concentration. This community resource is expected to continue to facilitate high-throughput genotyping for QTL discovery in soybean.

## Supplemental Information Available

Supplemental information is available online free of charge at http://crop.scijournals.org.

## Acknowledgments

We thank Alicia Bertles, Tad Sonstegard, and the BFGL-ANRI BARC-East DNA Sequencing Facility for assistance with the genomic STS sequencing. We would also like to thank all the soybean breeders listed in Supplemental Table 1 for providing elite cultivar seed for this project; the Williams 82 physical map consortium (http://www.soybase.org) for providing trace files for the Williams 82 BAC-end sequence; and David Lightfoot (Southern Illinois University) and Chris Town (TIGR) for providing trace files for the Forrest BAC-end sequences. This research effort was partially supported by United Soybean Board funded Projects 7212 and 8212. The support of the United Soybean Board is greatly appreciated.

## References

Akkaya, M.S., A.A. Bhagwat, and P.B. Cregan. 1992. Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131–1139.

Beavis, W.D. 1994. The power and deceit of QTL experiments: Lessons from comparative QTL studies. p. 250–266 *In* D.B. Wilkinson (ed.) Proc. of the 49th Annual Corn and Sorghum Industry Res. Conf., Chicago, IL. 7–8 Dec. 1994. Am. Seed Trade Assoc., Washington, DC.

Choi, I.-Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.-S. Yoon, E.-Y. Hwang, S.-I. Yi, N.D. Young, R.C. Shoemaker, C.P. Van Tassel, J.E. Specht, and P.B. Cregan. 2007. A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696.

Concibido, V.C., D.A. Lange, R.L. Denny, J.H. Orf, and N.D. Young. 1997. Genome mapping of soybean cyst nematode resistance genes in 'Peking', PI 90763 and PI 88788 using DNA markers. Crop Sci. 37:258–264.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999a. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490.

Cregan, P.B., J. Mudge, E.W. Fickus, D. Danesh, R. Denny, and N.D. Young. 1999b. Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the *rhg1* locus. Theor. Appl. Genet. 99:811–818.

Cregan, P.B., J. Mudge, E.W. Fickus, L.F. Marek, D. Danesh, R. Denny, R.C. Shoemaker, B.F. Matthews, T. Jarvik, and N.D. Young. 1999c. Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor. Appl. Genet. 98:919–928.

Fan, J.B., M.S. Chee, and K.L. Gunderson. 2006. Highly parallel genomic assays. Nat. Rev. Genet. 7:632–644.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, and M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. 68:69–78.

Gizlice, Z., T.E. Carter, Jr., T.M. Gerig, and J.W. Burton. 1996. Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. Crop Sci. 36:753–765.

Hyten, D.L., J.R. Smith, R.D. Frederick, M.L. Tucker, Q. Song, and P.B. Cregan. 2009. Bulked segregant analysis using the GoldenGate assay to locate the *Rpp3* locus that confers resistance to soybean rust in soybean. Crop Sci. 49:265–271.

Hyten, D.L., Q. Song, I.Y. Choi, M.S. Yoon, J.E. Specht, L.K. Matukumalli, R.L. Nelson, R.C. Shoemaker, N.D. Young, and P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116:945–952.

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, J.M. Costa, J.E. Specht, R.C. Shoemaker, and P.B. Cregan. 2006.

Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103:16666–16671.

Keim, P., T.C. Olson, and R.C. Shoemaker. 1988. A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. 15:150–152.

Mansur, L.M., and J.H. Orf. 1995. Agronomic performance of soybean recombinant inbreds in northern USA and Chile. Crop Sci. 35:422–425.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan, and K.G. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36:1327–1336.

Matukumalli, L.K., J.J. Grefenstette, D.L. Hyten, I.Y. Choi, P.B. Cregan, and C.P. Van Tassell. 2006. SNP-PHAGE–High throughput SNP discovery pipeline. BMC Bioinformatics 7:468.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132:365–386.

Saini, N., J. Shultz, and D.A. Lightfoot. 2008. Re-annotation of the physical map of *Glycine max* for polyploid-like regions by BAC end sequence driven whole genome shotgun read assembly. BMC Genomics 9:323.

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

Shoemaker, R.C., K. Polzin, J. Labate, J. Specht, E.C. Brummer, T. Olson, N. Young, V. Concibido, J. Wilcox, J.P. Tamulonis, G. Kochert, and H.R. Boerma. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–338.

Shultz, J.L., S. Kazi, R. Bashir, J.A. Afzal, and D.A. Lightfoot. 2007. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor. Appl. Genet. 114:1081–1090.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122–128.

Van Ooijen, J.W., and R.E. Voorrips. 2001. JoinMap 3.0 software for the calculation of genetic linkage maps. Plant Res. Int., Wageningen, the Netherlands.

Vodkin, L.O., A. Khanna, R. Shealy, S.J. Clough, D.O. Gonzalez, R. Philip, G. Zabala, F. Thibaud-Nissen, M. Sidarous, M.V. Stromvik, E. Shoop, C. Schmidt, E. Retzel, J. Erpelding, R.C. Shoemaker, A.M. Rodriguez-Huete, J.C. Polacco, V. Coryell, P. Keim, G. Gong, L. Liu, J. Pardinas, and P. Schweitzer. 2004. Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics 5:73.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134.

Zou, J.J., R.J. Singh, J. Lee, S.J. Xu, P.B. Cregan, and T. Hymowitz. 2003. Assignment of molecular linkage groups to soybean chromosomes by primary trisomics. Theor. Appl. Genet. 107:745–750.