Published August 10, 2015

NOTE

A Unified Single Nucleotide Polymorphism Map of Sunflower (*Helianthus annuus* L.) Derived from Current Genomic Resources

Brent S. Hulke,* Christopher J. Grassa, John E. Bowers, John M. Burke,
Lili Qi, Zahirul I. Talukder, and Loren H. Rieseberg

ABSTRACT

Dense genetic maps are critical tools for plant breeders and geneticists. While many maps have been developed for sunflower in the last few decades, most have been based on low-throughput technologies and include marker numbers in the hundreds. However, two maps with reasonably dense coverage of about 5000 and 9000 single nucleotide polymorphism (SNP) loci each have recently been produced using high-throughput genotyping methods. Unfortunately, no mapping population is common between the two maps, making the development of a joint map a challenge. With genome sequencing and resequencing of mapping populations currently in progress, there will be opportunities in the near future to develop much more informative resources. In the meantime, there is much demand from the sunflower community, particularly plant breeders, to combine these two maps to develop a denser map for immediate needs. In this paper, we used an *in silico* approach to join the two SNP maps by placing our existing marker sequences on draft genome scaffolds. Genetic map positions of the markers were determined from a resequenced mapping population aligned to the same draft genome scaffolds. In this way, we were able to directly place 10,247 SNP and insertion-deletion (Indel) loci on a common linkage map, and also provide the ability to infer genetic position of a further 6724 SNP loci from both previously published maps. These results will allow researchers to compare previous genetics research conducted on the separate maps, and facilitate collaborative work on marker-assisted breeding approaches in sunflower.

B.S. Hulke and L.L. Qi, Sunflower and Plant Biology Research Unit, Northern Crop Science Lab., 1605 Albrecht Blvd. N., Fargo, ND 58102-2765; C.J. Grassa and L.H. Rieseberg, Univ. of British Columbia, 6270 University Blvd., Vancouver, BC V6T 1Z4; J.E. Bowers and J.M. Burke, Univ. of Georgia, Dep. of Plant Biology, Miller Plant Sciences Bldg., Athens, GA 30602; and Z.I. Talukder, Dep. of Plant Sciences, North Dakota State Univ., P.O. Box 6050, Fargo, ND 58108-6050. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. Received 5 Nov. 2014. Accepted 18 Mar. 2015. *Corresponding author (brent.hulke@ars.usda.gov).

Abbreviations: cM, centiMorgan; ESTs, expressed sequence tags; Indel, insertion-deletion; LG, linkage group; RAD, restriction site associated DNA; RIL, recombinant inbred line; SSRs, simple sequence repeats; SNP, single nucleotide polymorphism; WGS, whole-genome shotgun.

A prerequisite for any modern genetic study is the development of large numbers of highly repeatable, high-throughput, and low cost genetic markers. In sunflower, some work on development of single nucleotide polymorphism (SNP) and insertion-deletion (Indel) markers that are compatible with most high-throughput marker scoring platforms has been done, with two SNP marker consortia recently publishing their results. Both marker consortia were the result of separate public-private partnerships with recently lapsed data embargoes. The SNP marker map described by Bowers et al. (2012; see also Bachlava et al. (2012), for marker development data), which we will refer to as the "Bowers map," consists of over 10,000 mapped loci (8571 SNPs + 1512 simple sequence repeats or SSRs) aligned to the linkage

Published in Crop Sci. 55:1696–1702 (2015).
doi: 10.2135/cropsci2014.11.0752

group (LG) designations of the SSR map of Tang et al. (2002), which is the conventional system of the worldwide sunflower community. A second map described by Talukder et al. (2014; see also Pegadaraju et al. (2013), for marker development data), which we will refer to as the "Talukder map," describes the development of an additional 10,000 SNP and Indel loci for high-throughput analysis, of which only 5019 loci have been genetically mapped. This map also followed the Tang et al. (2002) LG nomenclature. Both sets of SNP markers had favorable quality scores for use with the Illumina Infinium Genotyping Technology (Illumina, San Diego, CA, USA), and represent a potentially powerful tool for the sunflower breeding community.

The Bowers markers were developed from long-read ESTs derived from the Compositae Genome Project (CGP, 2014), as well as short-read transcriptome sequences derived from additional genotypes (Bachlava et al., 2012), resulting in SNPs exclusively from expressed regions of the genome. However, the Talukder markers were developed using Restriction site Associated DNA (RAD) sequence technology, which can capture polymorphism in both gene coding and non-coding regions (Pegadaraju et al., 2013). For this reason, we anticipate that a combination of the two maps may allow some areas of low marker density in the genetic map to be improved.

Since separate consortia have developed these maps and the maps are derived from different mapping populations, there is considerable interest, particularly in the private sector, to consolidate the maps. With the sequencing of the sunflower genome well underway, the availability of dense genetic mapping data derived from assembled whole-genome shotgun (WGS) sequences provides a means to determine locations of some of the markers on a common recombinant inbred line (RIL) map (Kane et al., 2011; Gill et al., 2014; Renaut et al., 2013). Such an analysis would first require sequence similarity of the published SNP and Indel sequences to the scaffolds of the unpublished draft sunflower genome, which is about 3.6 Gbp in length (Gill et al., 2014). It would further require that the scaffolds be polymorphic in a sequenced mapping population, allowing assignment of centiMorgan (cM) positions for each marker. Thus, our objectives were to (1) develop a common map for a large number of markers from the Bowers and Talukder maps using sunflower sequence data, and (2) determine the number of unique markers between the Bowers and Talukder maps, using sequence parsing tools.

MATERIALS AND METHODS

Our first objective requires the development of an ultra-high density, scaffold-based genetic map. Such a map was developed from the mapping population 'RHA 280/RHA 801' (Tang et al., 2002) using Illumina HiSeq reads as described by Renaut et al. (2013). Briefly, whole genome shotgun sequencing was conducted on the two parent lines 'RHA 280' and 'RHA 801' (Roath et al., 1981) to a 10x depth, and ninety-three RIL progeny of this population to a 1x depth. The whole genome shotgun sequences were aligned to scaffolds of two reference genomes, both of inbred line HA 412HO: one assembled from Illumina reads using Allpaths-LG, and one assembled from 454 reads using CABOG (data available by request from the authors). The methods for mapping reference genome scaffolds were mutually consistent for both assemblies, and are as follows. First, all fixed, polymorphic SNPs evident in the RHA 280 and RHA 801 sequences were used as genetic map markers. RIL reads, after alignment to the HA 412HO reference, were genotyped for parental origin at each SNP. Parental origin of each RIL scaffold was determined if at least one polymorphic locus was available, and recombination frequency between adjacent scaffolds calculated. The scaffolds were then assigned cM positions in the same manner as the Bowers map (Bowers et al., 2012), which involved dividing the raw recombination counts by (1.996*2*93) to adjust for expected double recombinants in a population of 93 lines (Winkler et al., 2003). Where there were multiple polymorphic sites on a scaffold with different genetic map locations, the average of the distances was used.

SNP probe sequence from both the Bowers and Talukder maps, as well as SSR primers, Indel primers, and SNP sequence from earlier maps and other data (*i.e.* Supplemental Table S1) were queried against the mapped reference genome scaffolds, described above, using BLAT (Kent, 2002) with default parameters. The Bowers marker sequences are 120 bp or less in length and the Talukder markers are at least 201 bp in length, with many greater than 400 bp. In the case that primers were available instead of sequence, the reverse complement of the reverse primer was formed using BioEdit (Hall, 1999), and appended to the end of the forward primer. The BLAT output was filtered to provide only the best hits of query marker sequence to scaffolds using pslReps software (Kent, 2002) and the resulting output imported into spreadsheet software. Quality of the alignments were determined by scoring each on a $-\infty$ to 1 scale, which was calculated as the fraction of the query sequence successfully aligned, penalized by -1 for each bp of gap length added to the alignment (with the gap between primers excluded in the case of SSRs). In this way, complete sequence matches with no gaps received a score of 1. Mismatched single nucleotides (either SNPs or sequencing errors) were not penalized. Only those sequence matches with a score of 0.98 and greater were retained for further analysis. This process was completed separately for the two reference genomes and the results compared.

After assigning the markers to scaffolds, the filtered output was combined with map position data from the Bowers and Talukder maps, and the scaffolds. Quality control of the resulting scaffold map was conducted as follows. If there was one high scoring scaffold-marker match, and the scaffold map location was not in syntenic agreement with the position of the marker on the original map, the marker was removed from the dataset. Some of the markers had more than one high scoring match to the scaffolds, and in some cases these scaffolds were on different LGs. In these cases, the map locations were compared to the original genetic maps and the matching map placement retained. This was seen as a reasonable way to handle this issue because the existence

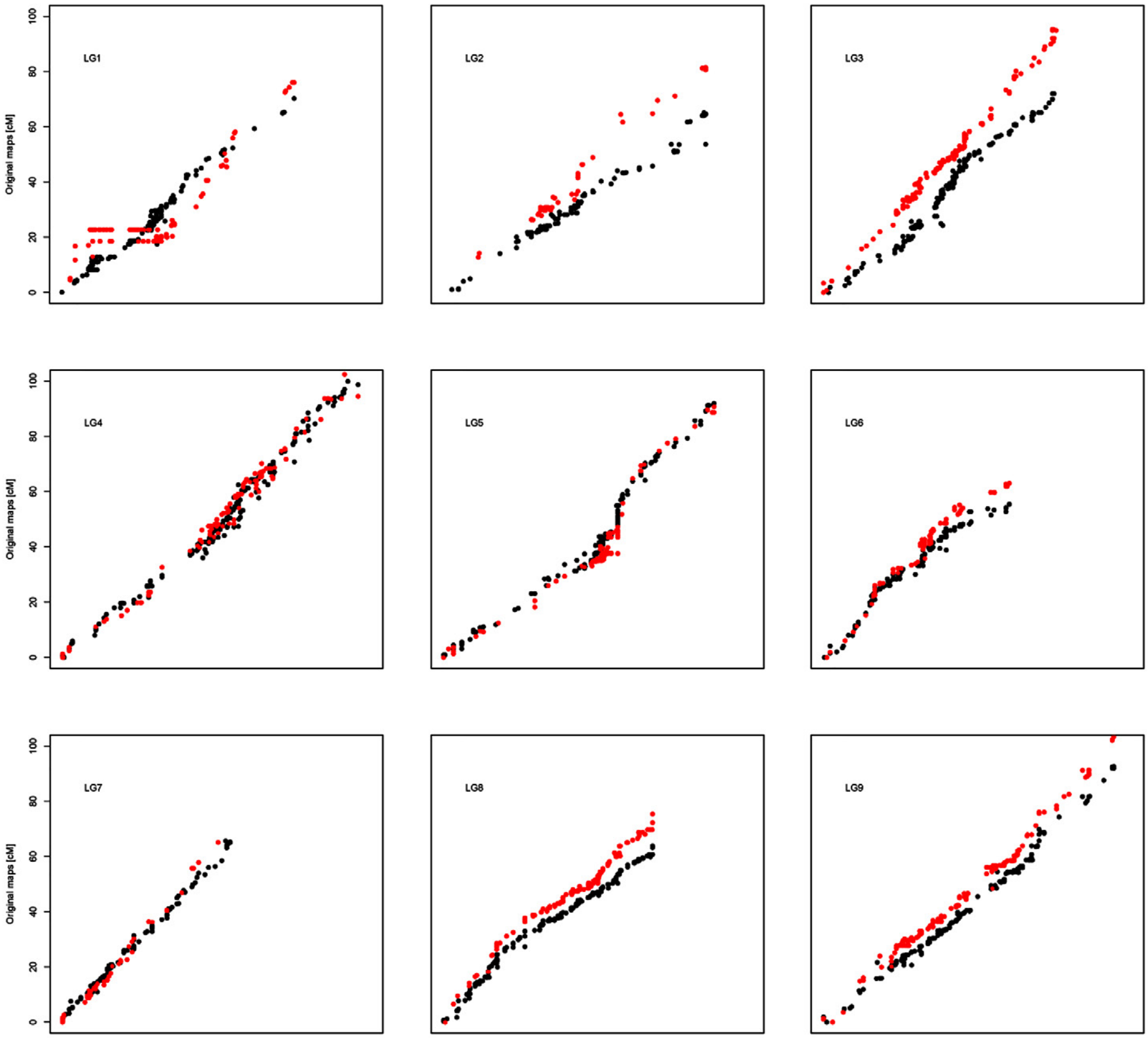

Figure 1. X-Y scatterplots of markers common between the scaffold map (*x* axis) and one or both of the Bowers and Talukder maps (*y* axis).

of previous map positions should mean that the SNPs segregated according to Mendelian norms. If there was no published map position available for a marker, then multiple, conflicting scaffold positions were handled by removing the ambiguous marker completely from the data set. This was completed for the two scaffold maps generated by 454 and Illumina reads.

The 454 and Illumina read maps were then used to validate placement of the markers on our final scaffold map. Since the 454 and Illumina maps were built on the same mapping population and map distance calculation, they are directly comparable. If the two scaffold maps disagreed for a particular locus, the position that was syntenic with the previously published position was used. If the 454 and Illumina maps disagreed and the locus was not previously mapped, the marker was discarded from the results.

Markers with both a published and a scaffold map position were then plotted on an x-y scatter diagram. Any outliers, suggesting placement incongruent with previous knowledge, were removed from the data set. An x-y scatter diagram was developed for the final data set (Fig. 1).

To fulfill our second objective of finding unique markers between the Bowers and Talukder maps, duplicate loci between the Talukder and Bowers maps were discovered using BLAT, with the longer Talukder SNP sequences as the template set and the shorter Bowers SNP sequences as the query set. The results were filtered with pslReps and the resulting output converted using pslPretty (Kent, 2002) to a traditional text alignment format. Matches that appeared identical were then reinspected using BioEdit software, which unambiguously showed the SNP sites in the sequence and allowed identification of matches between

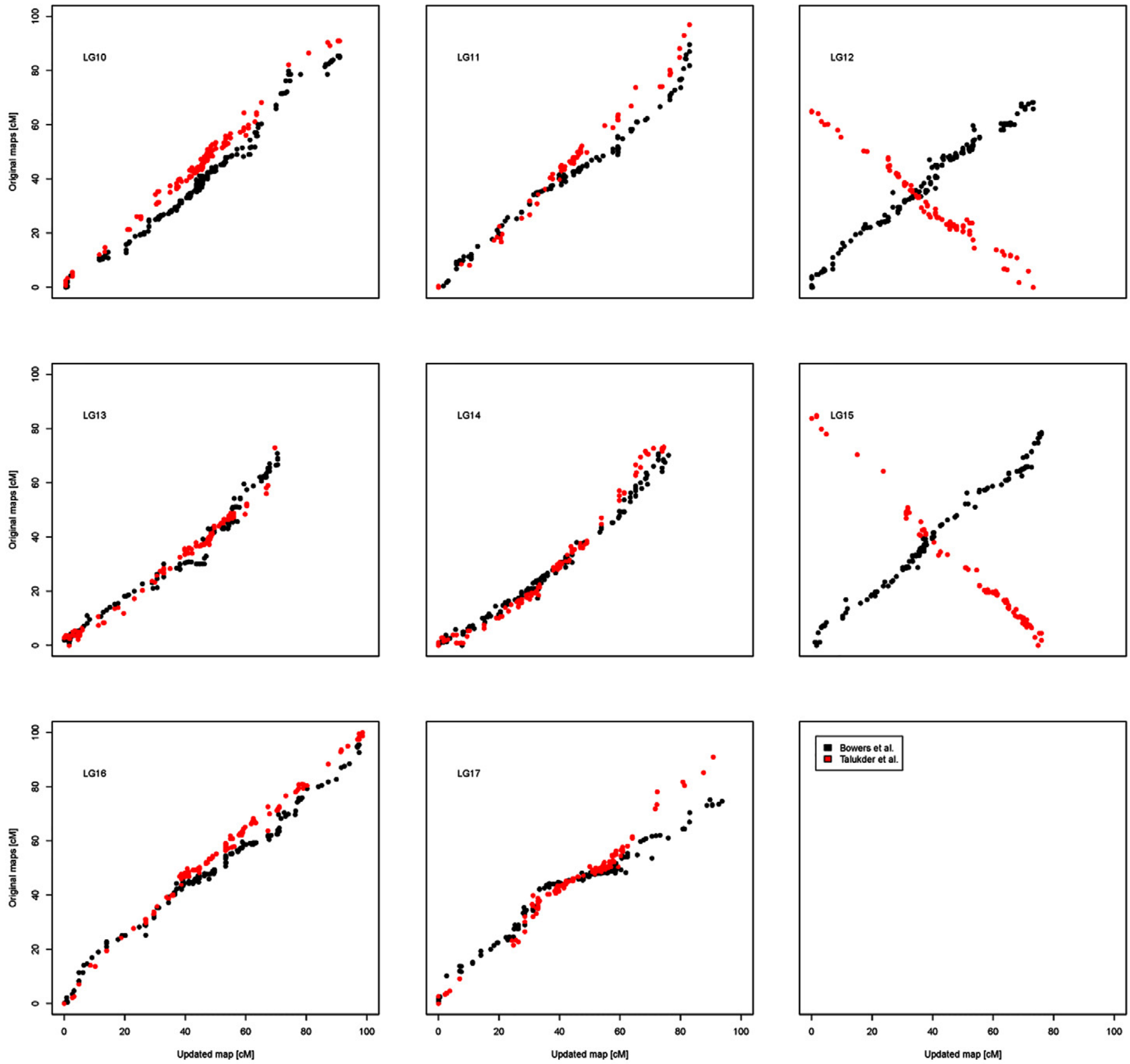

Figure 1. Continued.

the Bowers and Talukder markers. Any matches found in this manner indicate two markers of different name marking the same locus, and were counted as a single, common locus between maps (union region between Talukder and Bowers in Fig. 2).

RESULTS AND DISCUSSION

The sunflower genome is the product of ancient genome duplication events, as well as more recent duplications, inversions, and transposable element activity (Kane et al., 2011; Gill et al., 2014). It has been estimated that approximately 80% of the sunflower genome consists of repetitive elements, and that transposition is not only active but may be required for normal plant development (Staton et al., 2012; Gill et al., 2014). Presumably recent changes in genomic organization have been found in the resequenced genomes of several recently developed public inbred lines of sunflower. All of these factors complicate the identification of single-locus SNP sequences that segregate in a Mendelian manner in sunflower.

The results of our *in silico* approach for combining maps are also affected by these events, but in general, most markers were unambiguous in their placement and their positions were consistent with previous maps. A total of 15.4% of the best scaffold matches were not placed at the previously published position among the Bowers map markers. This compares to the previously reported rate of 14% for multi-locus

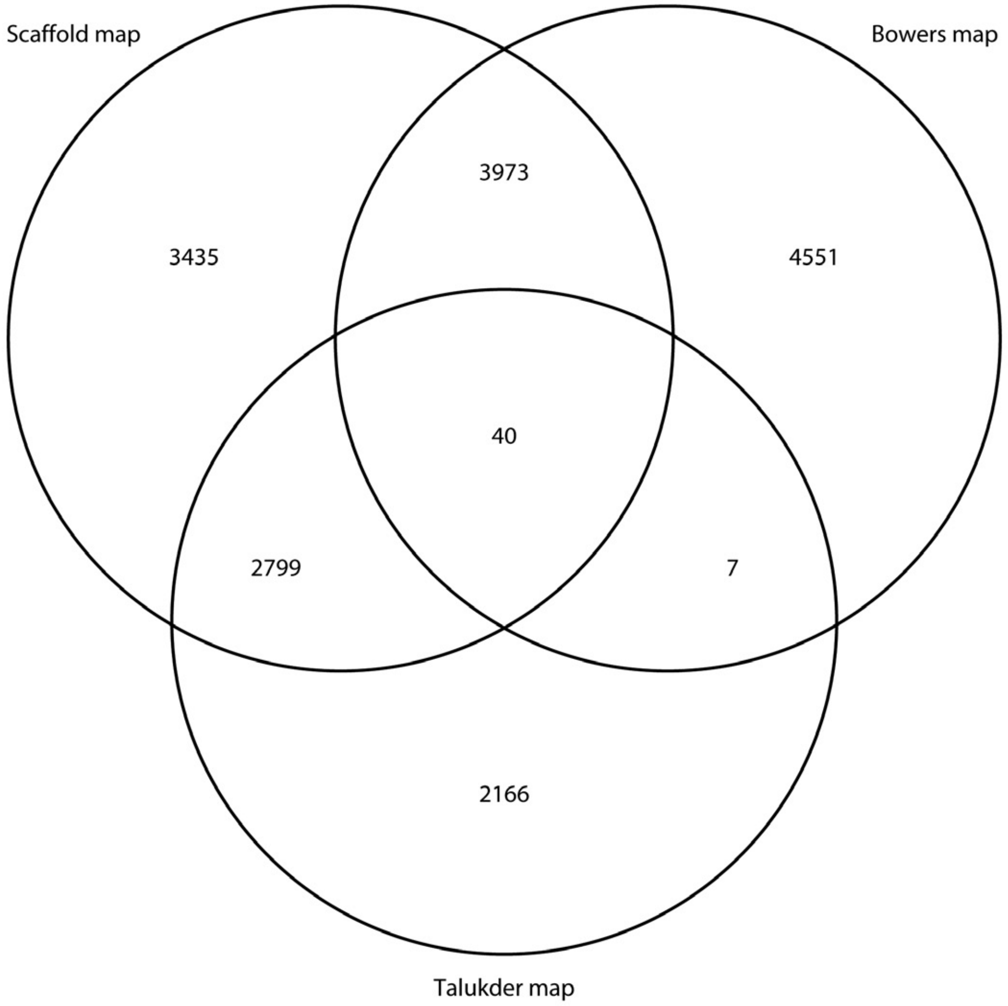

Figure 2. Venn diagram describing the distribution of SNP marker loci of the Bowers and Talukder sets on the Bowers, Talukder, and Scaffold genetic maps.

placement of the Bowers markers among individual linkage maps (Bowers et al., 2012). For the Talukder markers, the rate of mismatch between the scaffold and published map was less at 10.5%. This, however, is much higher than the < 1% of markers that were deemed multi-locus in the original publication (Talukder et al., 2014). Unique segmental duplications have been observed among resequenced inbred lines of recent breeding, indicating differences in segment copy number between populations is a common issue in genetic mapping (unpublished data, 2014). The Bowers and Talukder sets were mapped with multiple biparental populations derived from seven and five sunflower genotypes, respectively. Any copy number differences between the genomes of the original Bowers and Talukder mapping populations and the RHA 801/RHA 280 mapping population of our scaffold map could be observed as multi-locus markers. Base pair length of the SNP probe sequence may matter in distinguishing these copies, with the longer sequences providing more certainty in obtaining one highly similar match. The existence of recent duplication highlights the uphill battle of SNP design without a complete reference genome.

Two LGs exhibited anomalies when the nearly complete map was visualized on scatter diagrams. LG1 showed non-collinearity between the Talukder map and the scaffold map between the 4 and 40 cM positions on the scaffold map. This was not observed with the Bowers map. Figure 3 of Talukder et al. (2014) shows a similar anomaly in populations 1 and 2 of their consensus map. We attribute this to a sequence inversion that was found among resequenced lines (unpublished data, 2014), and retained these markers in the final map. On LG13, markers at the Bowers 2 cM and 30 cM positions appeared scattered throughout the LG on our scaffold map (data not shown). This was not seen on the Talukder map. To study this further, we BLAST searched both the problem markers and their associated scaffolds against the current genome assembly. While the markers themselves showed one sufficiently similar match (as seen previously from the BLAT results), the scaffolds for each marker showed high similarity to scaffolds on LG13 and other LGs (e-values < $< 10^{-100}$), suggesting that these sequences are frequently repeated. Besides providing us with difficulty in placing the markers, this suggests that there may also be technical concerns about these particular markers. For that reason, they were removed from our dataset and our final scatter diagram (Fig. 1). While the Talukder markers did not appear to have these difficulties, the Talukder map also appears to have a gap at those map positions, suggesting markers in these regions may have been eliminated previously for quality concerns or never discovered in their work. Another possibility is that the longer SNP sequence in the Talukder markers made for more accurate placement, eliminating the issue.

In the Talukder map, LG12 and 15 were in reverse order compared to the Tang et al. (2002) reference map (Fig. 1). This is due to the lack of framework SSR markers on LG12 and 15 in the original work to guide alignment in the Talukder map. The Bowers map was organized in exactly the same fashion as the Tang et al. (2002) map on all LG.

A single, biparental mapping population, even one as polymorphic as RHA 280/RHA 801, is expected to be monomorphic at many loci. This limited our ability to place on the scaffold map a large number of the SNPs that were successfully mapped previously. Our method was aided by the fact that only a single polymorphic site on a scaffold was required to place it on the genetic map. Even so, a larger dataset with more mapping populations would certainly help place more of these markers on a common map. Genomic resources such as this are in development. A total of 6812 previously published SNP loci were positioned both on a published map as well as a scaffold map; however, 6724 previously mapped SNP loci were not positioned on our new map because they did not align with a polymorphic scaffold in the RHA 280/RHA 801 population under our quality control criteria (Fig. 2).

We were successful in finding map positions for some of the previously unmapped loci from Bowers and Talukder. Among all the unmapped markers, 3.3% were removed because the scaffold-based position was multilocus within an assembly and a further 4.0% were removed because the position based on the two genome assemblies did not agree, for a total of 7.2% removed. Again, duplications within the genome are the likely cause of ambiguity. A total of 2720 unmapped SNP loci of Talukder et al. (2014) and 715 unmapped loci of Bowers et al. (2012) passed these criteria and were mapped to a polymorphic scaffold in our work. These are the SNPs that are unique to the scaffold map in the Venn diagram (Fig. 2). Since these markers have not been previously mapped using conventional techniques, we caution the reader that these markers may have technical or biological issues that may make them unreliable. For this reason, these markers are italicized in the Supplementary Tables, to set them apart from those with higher quality data.

In comparing the two previously published marker sets, we expected that there would be a small number of loci in common, despite being derived from datasets from two different sequencing strategies. Each of 40 marker pairs had identical sequence, syntenic map positions in the Bowers and Talukder maps, and were placed on the scaffolds, as seen in the center of the Venn diagram (Fig. 2). Another 7 pairs were common between the Talukder and Bowers maps, but were unmapped in the scaffold map due to lack of polymorphism. A further 18 marker pairs were common between the scaffold map and one published map, again due to lack of polymorphism in one of the mapping projects. Given that the Bowers assay has the shorter sequence, the Bowers assay will generate a marker equivalent to the Talukder assay for each of these pairs. In Supplementary Tables S2, S3, and S4, the second marker of an identical pair is listed in the 'alias' column.

The scaffold-based map resulting from our work includes 10,247 SNP loci from the Bowers and Talukder sets, with 116 SSRs, Indels, and gene sequences from earlier maps and other data included in Supplemental Table 1. Of the 10,247, about 3500 placed SNPs were previously unmapped. In the converse case, where we were unable to place on the scaffolds SNPs that were previously mapped, the positions can be inferred from both previously published maps, for a total of 16,971 SNP loci mapped from the Bowers and Talukder sets (Fig. 2). To assist the reader, we supplied Supplemental Tables 2 and 3, which includes the entire map of Talukder and Bowers, respectively, with additional information provided by our work. This is especially helpful to those scientists regularly using one of those two maps. Supplementary Tables S2, S3, and S4 mirror the composition of the scaffold, Talukder, and Bowers circles in the Venn diagram (Fig. 2).

Acknowledgments

The authors wish to thank the National Sunflower Association, which provided funding for the Talukder map and encouraged the authors to conduct this work, as well as Genome Canada and Genome British Columbia, who provided funding for the sunflower scaffold map. The data from the Bowers map were produced with the support of the NSF Plant Genome Research Program (DBI-0820451), Advanta Semillas, Dow Agrosciences, Pioneer Hi-Bred, Syngenta, and the Georgia Research Alliance. We also wish to thank David Horvath of USDA-Agricultural Research Service, Fargo, ND, for insightful conversations that were helpful at the design phase of this work.

References

Bachlava, E., C.A. Taylor, S. Tang, J.E. Bowers, J.R. Mandel, J.M. Burke, and S.J. Knapp. 2012. SNP Discovery and Development of a High-Density Genotyping Array for Sunflower. PLoS ONE 7(1):E29814. doi:10.1371/journal.pone.0029814

Bowers, J.E., E. Bachlava, R.L. Brunick, L.H. Rieseberg, S.J. Knapp, and J.M. Burke. 2012. Development of a 10,000 Locus Genetic Map of the Sunflower Genome Based on Multiple Crosses. G3 2(7): 721–729. doi:10.1534/g3.112.002659

CGP. 2014. The Compositae genome project [Online]. Available at http://compgenomics.ucdavis.edu/ (verified 2 Oct. 2014).

Gill, N., M. Buti, N. Kane, A. Bellec, N. Helmstetter, H. Berges, and L. Rieseberg. 2014. Sequence-based analysis of structural organization and composition of the cultivated sunflower (Helianthus annuus L.) genome. Biology 3:295–319. doi:10.3390/biology3020295

Hall, T.A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.

Kane, N.C., N. Gill, M.G. King, J.E. Bowers, H. Berges, J. Gouzy, E. Bachlava, N.B. Langlade, Z. Lai, M. Stewart, J.M. Burke, P. Vincourt, S.J. Knapp, and L.H. Rieseberg. 2011. Progress towards a reference genome for sunflower. Botany 89:429–437. doi:10.1139/b11-032

Kent, W.J. 2002. BLAT–The BLAST-Like Alignment Tool. Genome Res. 12:656–664. doi:10.1101/gr.229202. Article published online before March 2002

Pegadaraju, V., R. Nipper, B. Hulke, L. Qi, and Q. Schultz. 2013. De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. BMC Genomics 14:556. doi:10.1186/1471-2164-14-556

Renaut, S., C.J. Grassa, S. Yeaman, B.T. Moyers, Z. Lai, N.C. Kane, J.E. Bowers, J.M. Burke, and L.H. Rieseberg. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat. Commun. 4:1827. doi:10.1038/ncomms2833

Roath, W.W., J.F. Miller, and T.J. Gulya. 1981. Registration of RHA 801 sunflower germplasm. Crop Sci. 21:479. doi:10.2135/cropsci1981.0011183X002100030041x

Staton, S.E., B. Hartman-Bakken, B.K. Blackman, M.A. Chapman, N.C. Kane, S. Tang, M.C. Ungerer, S.J. Knapp, L.H. Rieseberg, and J.M. Burke. 2012. The sunflower (Helianthus annuus L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72:142–153. doi:10.1111/j.1365-313X.2012.05072.x

Talukder, Z.I., L. Gong, B.S. Hulke, V. Pegadaraju, Q. Song, Q. Schultz, and L. Qi. 2014. A High-Density SNP Map of Sunflower Derived from RAD-Sequencing Facilitating Fine-Mapping of the Rust Resistance Gene R12. PLoS ONE 9(7):E98628. doi:10.1371/journal.pone.0098628

Tang, S., J.-K. Yu, M.B. Slabaugh, D.K. Shintani, and S.J. Knapp. 2002. Simple sequence repeat map of the sunflower genome. Theor. Appl. Genet. 105:1124–1136. doi:10.1007/s00122-002-0989-y

Winkler, C.R., N.M. Jensen, M. Cooper, D.W. Podlich, and O.S. Smith. 2003. On the determination of recombination rates in intermated recombinant inbred populations. Genetics 164:741–745.