Development of a 10,000 Locus Genetic Map of the Sunflower Genome Based on Multiple Crosses

John E. Bowers,*,† Eleni Bachlava,†,[1] Robert L. Brunick,‡ Loren H. Rieseberg,§, Steven J. Knapp,†,[1] and John M. Burke*,[2]**

*Department of Plant Biology and †Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia 30602, ‡Department of Crop and Soil Science, Oregon State University, Corvallis, Oregon 97331, §Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada, and **Department of Biology, Indiana University, Bloomington, Indiana 47405

ABSTRACT Genetic linkage maps have the potential to facilitate the genetic dissection of complex traits and comparative analyses of genome structure, as well as molecular breeding efforts in species of agronomic importance. Until recently, the majority of such maps was based on relatively low-throughput marker technologies, which limited marker density across the genome. The availability of high-throughput genotyping technologies has, however, made possible the efficient development of high-density genetic maps. Here, we describe the analysis and integration of genotypic data from four sunflower (*Helianthus annuus* L.) mapping populations to produce a consensus linkage map of the sunflower genome. Although the individual maps (which contained 3500–5500 loci each) were highly colinear, we observed localized variation in recombination rates in several genomic regions. We also observed several gaps up to 26 cM in length that completely lacked mappable markers in individual crosses, presumably due to regions of identity by descent in the mapping parents. Because these regions differed by cross, the consensus map of 10,080 loci contained no such gaps, clearly illustrating the value of simultaneously analyzing multiple mapping populations.

Genetic linkage maps can facilitate a wide range of analyses in both plant and animal species. Indeed, such resources play a critical role in the genetic dissection of complex traits (Al-Chaarani *et al.* 2004; Buckler *et al.* 2009; Frary *et al.* 2000; Tanksley and McCouch 1997) and comparative analyses of genome structure (Gale and Devos 1998; Grant *et al.* 2000; Moore *et al.* 1995), as well as in molecular breeding efforts in species of agronomic importance (Brondani *et al.* 2002; Qi *et al.* 2011; Tanksley and Nelson 1996). Until recently, such maps were typically based on relatively low-throughput marker technologies, which tended to limit marker density across the genome. However, with the advent of next-generation sequencing technologies, as well as the development of high-throughput genotyping platforms, it is now possible to simultaneously interrogate thousands of single nucleotide polymorphisms (SNP) from throughout the genome (Gupta *et al.* 2008), making possible the efficient development of increasingly dense genetic maps. In this article, we describe the development and integration of four high-density genetic maps of the sunflower (*Helianthus annuus* L.) genome using a recently developed Illumina Infinium SNP genotyping array (Bachlava *et al.* 2012).

Cultivated sunflower is a globally important oilseed crop that is grown on over 23 million hectares worldwide (http://www.fao.org/). With annual production of 32 million metric tons, sunflower is the world's fourth largest source of vegetable oils (http://www.fas.usda.gov/). Originally domesticated in the east-central United States *ca.* 4000 years ago (Crites 1993; Harter *et al.* 2004; Smith and Yarnell 2009), sunflower was long grown as a source of edible seeds as well as for a variety of nonfood purposes (*e.g.* as a source of dye for textiles and for ceremonial purposes) (Soleri and Cleveland 1993). The use of sunflower as an oil crop is a more recent development, dating back to

doi: 10.1534/g3.112.002659

Manuscript received March 15, 2012; accepted for publication April 23, 2012

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.112.002659/-/DC1

[1] Present address: Monsanto Company, Woodland, CA 95695.

[2] Corresponding author: University of Georgia, Dept. of Plant Biology, Miller Plant Sciences, Athens, GA 30602. E-mail: jmburke@uga.edu

the eighteenth century in Eastern Europe. Over the years, breeding efforts increasingly focused on improving oil yield, and the germplasm that formed the basis of the modern oilseed gene pool was eventually brought back to North America, where commercial production commenced in the mid-twentieth century. In recent years, hybrid oilseed lines have accounted for 80–85% of sunflower production in the United States, with the balance coming from confectionery (*i.e.* nonoilseed) lines.

An extensive collection of genetic linkage maps exists for sunflower developed using various molecular markers, including restriction fragment length polymorphisms (RFLP) (Berry *et al.* 1995; Gentzbittel *et al.* 1995), random amplified polymorphic DNAs (RAPD) (Rieseberg *et al.* 1993), AFLPs (Gedil *et al.* 2001), and simple-sequence repeats (SSR) (Heesacker *et al.* 2008; Tang *et al.* 2002; Yu *et al.* 2003). These maps have been used to identify quantitative trait loci (QTL) underlying numerous traits of agronomic importance (Burke *et al.* 2002; Tang *et al.* 2006; Wills and Burke 2007) and to investigate variation in genome structure between sunflower and other *Helianthus* species (Burke *et al.* 2002; Heesacker *et al.* 2008). However, the relatively low density of these maps, with an average of less than one marker per 2 Mb of DNA in the best cases, has limited their utility for other applications. For example, assembly of the forthcoming sunflower genome (Kane *et al.* 2011) would be greatly assisted by the availability of much denser genetic maps. The initial Arabidopsis genome project relied on a genetic map with an average density of one marker per 151 kb (Kaul *et al.* 2000), and other sequencing projects have depended on similarly dense genetic maps, with densities of one marker per 171 kb in rice (Harushima *et al.* 1998) and one marker per 290 kb in sorghum (Paterson *et al.* 2009). Given that the sunflower genome spans 3.5 Gb (Kane *et al.* 2011), a map containing at least 10,000 loci would be required to approach these marker densities.

As a member of the Asteraceae, which is the largest and one of the most ecologically successful plant families (Magallon *et al.* 1999), sunflower has no close relatives whose genomes have been fully sequenced and published; the only fully-sequenced Asterid genome that has been published is the distantly related potato (Xu *et al.* 2011). Several other Asterid species (including lettuce, which is likewise a member of the Asteraceae, tomato, and *Mimulus*) are, however, currently being sequenced [CoGePedia (genomevolution.org); R. W. Michelmore, personal communication]. Given that the Asteraceae is separated from other Asterids by a polyploidy event near the base of the family (Barker *et al.* 2008) and that the lineage leading to the sunflower genome has experienced an additional ancient polyploidization, the development of a high-density genetic map of the sunflower genome would not only provide a valuable resource to the sunflower research community but also a powerful tool for comparative genomic analyses of the Asterids. Here, we describe the development of such a map based on the analysis and integration of genotypic data from multiple sunflower mapping populations.

MATERIALS AND METHODS

Mapping populations

Four different sunflower mapping populations were used to construct four separate genetic maps in this study. These maps were then combined to produce a consensus map. Each of these populations involved complex, multigenerational pedigrees (supporting information, Figure S1). The populations tested involved parents representing oilseed, confectionary, wild, and landrace types, as well as fertility maintainer and restorer lines that are used in sunflower hybrid breeding programs. The first cross involved HA412-HO, an oilseed maintainer line with high oleic acid content, and RHA415, an oilseed restorer line. The second cross involved HA412-HO and ANN1238, a wild accession of sunflower collected from Keith County, NE. The third cross involved RHA280, an oilseed restorer line, and RHA801, a confectionery restorer line. The final cross involved NMS373, an oilseed restorer line that segregates for nuclear male sterility, and Hopi (PI 369359), a nonoilseed landrace. The HA412-HO × RHA415 and HA412-HO × ANN1238 populations consisted of 94 F_2I_1 individuals, resulting from the intermating of approximately 500 and 100 F_2 plants from a single F_1, respectively, for a single generation. The RHA280 × RHA801 population included 70 recombinant inbred lines (RIL) developed after selfing of F_2 plants derived from a single F_1 via single-seed descent for seven generations, resulting in F_9 RILs. The final population consisted of 153 lines derived from a cross between NMS373 × Hopi involving multiple F_1 individuals that were intermated to produce the F_2 generation. This population was intermated for three additional generations and then self-pollinated and advanced via single-seed descent for seven generations to produce advanced-intercross RILs.

DNA extraction and SNP genotyping

For the first two crosses, DNA was extracted from leaf tissue collected from individual plants. For the latter two crosses, which involved inbred lines, DNA was extracted from pooled leaf tissue that was collected from 4–5 individuals/line. All DNA extractions were done using a modified CTAB protocol (Murray and Thompson 1980), and DNA concentrations were quantified using the Quant-iT PicoGreen dsDNA reagent (Invitrogen, Carlsbad, CA). Genotyping was done using an Illumina Infinium SNP array (Bachlava *et al.* 2012). Genotyping was performed according to the manufacturer's recommendations on the Illumina iScan System (Illumina Inc., San Diego, CA) at the Emory University Biomarker Service Center. Prior to hybridization of the Beadchips, DNA was diluted to 50 ng/μl and quality was assessed via UV spectrophotometry and agarose gel electrophoresis.

All SNP data analyses were performed with the raw intensity data from the Illumina Beadchip using Genome Studio ver. 2011.1 (Illumina). The no-call threshold was set to 0 to force a genotype call on all data, and all loci were visually examined. Loci that exhibited clearly distinct genotypic clusters were scored automatically by the software. Approximately one third of the polymorphic loci required manual scoring to separate genotypes, because the genotypic clusters were too close to one another and the polymorphisms could not be reliably scored via the automated allele calling. A small number of SNP assays showed highly complex patterns that could be used to score two loci (Figure S2). After the first round of scoring and map assembly (details below), loci that appeared to contain double crossovers for individual plants that could not be fixed by reordering loci, as well as those that could not be mapped, were reexamined for scoring errors, and allele calls were manually adjusted as appropriate. To account for differences in genetic background that could alter signal intensities for individual assays, genotype calling was performed separately for each mapping population, along with the respective parental genotypes.

Genetic map construction

Due to the complexity of the dataset, initial map orders were determined manually with spreadsheet software (Microsoft Excel 2010; Microsoft Corp., Redmond, WA). Briefly, this process involved the use of spreadsheets containing raw data with individuals or lines as columns, and loci arranged in rows with the genotypes of each individual plant or line. Adjacent loci were clustered by sequentially sorting genotypes of multiple plants, and clusters of loci with similar segregation patterns were manually joined by moving the rows to minimize the total number of recombination events observed. The map orders derived from the RHA280 × RHA801 and NMS373 × Hopi populations were also independently assembled using MSTMap ver. 2007 (Wu *et al.* 2008) and Mapdisto ver. 1.7.5 (Lorieux *et al.* 2000) to validate the manual approach. Because MSTMap is designed for use with mostly homozygous data, it was not used to produce map orders for the two early-generation mapping populations (*i.e.* HA412-HO × HA415 and HA412-HO × ANN1238). We did, however, cross-validate all maps to ensure that shared markers had a consistent order across populations. The combination of high marker density and exceptionally low error rates in the raw data (Bachlava *et al.* 2012) facilitated the manual map construction. Markers were not tested for segregation distortion prior to map construction because distorted regions may be present in the genetic maps (Xu *et al.* 1997), especially those involving multigenerational populations.

With multigenerational mapping populations, only a fraction of the actual recombination events can be observed, as many recombination events occur in homozygous regions or are lost in subsequent generations. To facilitate distance comparisons among the maps constructed herein as well as with previously published maps, the recombination distances were adjusted to the rate expected for an F_2 population (Winkler *et al.* 2003). These adjustments involved dividing the raw recombination counts for the HA412-HO × RHA415 and HA412 × ANN1238 maps by 1.5, the HA280 × HA801 by 1.996, and the NMS373 × Hopi map by 2.992. Map distances are presented as raw recombination fractions without any mapping function applied. No mapping function was used because the multigenerational nature of the mapping populations resulted in recombination events that occurred in different generations, which violates the mathematical assumptions of both the Kosambi and Haldane map functions regarding recombination interference (Crow 2007). Note that application of the Kosambi mapping function changed the total map length from individual crosses by less than 1%.

A consensus map order was then produced by assigning to all markers an observed or implied map position on all four component maps. Implied map positions were assigned for all loci that did not map in a particular cross based on their positions relative to adjacent loci that were present on multiple maps. If a marker could be assigned to more than one possible map position in a particular component map, its implied map position was computed to assume identical marker order for all component maps where possible. Assays that revealed loci on two or more chromosomes or more than 20 cM apart on the same chromosome when comparing across maps were classified as multilocus assays, and the resulting loci were treated as separate markers.

A total of 1512 PCR-based loci, including indels, SSRs, and SNPs from previous maps derived from the same RHA280 × RHA801 mapping population (Heesacker *et al.* 2008; Lai *et al.* 2005; Tang *et al.* 2002; Yu *et al.* 2003), were then integrated into the consensus map described above. The PCR-based and Illumina SNP mapping populations used overlapping subsets of RILs from a larger RHA280 × RHA801 RIL mapping population with 40 lines in common. The shared RILs allowed for the placement of the PCR-based loci on the RHA280 × RHA801 Illumina SNP map, and their subsequent integration into the consensus map. Chromosome numbers and orientation could then be determined on the basis of the PCR-based marker data of the previous sunflower consensus map.

RESULTS AND DISCUSSION

Genetic map construction

The four mapping populations were used to produce four separate high-density genetic maps containing 3500–5500 loci each (summarized in Table 1; full map and raw data in File S1). For three of the maps (all but the NMS373 × Hopi), all mapped loci could be placed into 17 linkage groups, matching the expected chromosome number. For the RHA280 × RHA801 population, 1512 PCR-based loci from previous mapping efforts (Heesacker *et al.* 2008; Tang *et al.* 2002; Yu *et al.* 2003) using the same cross were integrated into the Illumina SNP map produced herein. Illumina SNP loci on the RHA280 × RHA801 map were used to number and orient the linkage groups on the other three maps. The manual map construction proved to be a practical approach and produced a result nearly identical to that produced by MSTMap and Mapdisto (Figure S3). Due to the complex crossing scheme and because Hopi is an outbred landrace with considerable heterozygosity, the NMS373 × Hopi mapping population produced a complicated mapping result. Because multiple F_1s were intercrossed prior to the production of the RILs, some loci segregated for alleles that were homozygous within each parent, but different between parents, whereas others segregated for alleles that were heterozygous within one parent or the other. The result was a base map of 17 linkage groups containing loci in the former class, along with ancillary linkage groups containing a number of additional linkage groups corresponding to heterozygous regions of the parental genomes. The base map included 3117 of the 5500 total loci segregating in this cross, with all of these loci segregating at an approximately 1:1 ratio.

The ancillary linkage groups from the NMS373 × Hopi population could be placed in three classes. The largest class corresponded to loci that were heterozygous within the Hopi mapping parent. This class included 1879 loci that were segregating 3:1 and that could be assigned to one of 78 linkage groups. Each heterozygous region in the Hopi parent resulted in the production of two genetic maps because, due to the bi-allelic nature of the data, genotypes from either of the 2 Hopi chromosomes could be identical to the NMS373 genotype. This resulted in two different 3:1 segregation patterns for each heterozygous genomic region. Because the genome of the Hopi mapping parent also contained homozygous regions, large chromosomal segments were unmappable in the 3:1 maps, thereby resulting in a larger than expected number of fragmentary linkage groups. The NMS373 parent was also heterozygous in certain (but far more restricted) genomic regions. In total, there were 456 polymorphic loci from NMS373 that were likewise segregating 3:1 and that could be mapped to 12 linkage groups corresponding to portions of 7 presumptive chromosomes. A substantial majority of these loci (304 of 456; 66.7%) mapped to a single region corresponding to a portion of chromosome 10. Finally, there were two small, ancillary linkage groups on chromosomes 10 and 13 that appeared to be the result of loci that were heterozygous in both parents. These two linkage groups contained 48 loci total. Because the majority of heterozygous parental loci that were placed on the submaps were mapped in the other three crosses (1798 of 2383 total; 75.5%), all of the linkage groups derived from the NMS373 × Hopi cross could be assigned to one of the 17 sunflower chromosomes and integrated into the combined map.

Although the NMS373 × Hopi map was far more complex than the other maps, it provided a large amount of additional data. Due to the multiple generations of intermating prior to RIL development, there were many more recombination events, which resulted in considerably higher map resolution. Indeed, this map provided 50% higher resolution compared to one constructed using a conventional

RIL population. Although the heterozygosity of the parental genotypes added complexity to the mapping process, it allowed for the mapping of many more markers and should also provide an opportunity to map QTL arising not only from segregation of the NMS373 *vs.* Hopi alleles, but also due to the segregation of alleles that were heterozygous within the parents.

Data quality and reproducibility

The identification and correction of genotyping errors is an essential aspect of the production of high-density genetic maps. With a map based on thousands of loci, even a low error rate in the raw data can vastly distort the total map length. As expected due to the high level of repeatability of this genotyping array (Bachlava *et al.* 2012), the raw allele calls (after manual editing) proved to be highly robust for genetic mapping. For example, in looking at the HA412-HO × RHA415 and HA412-HO × ANN1238 maps, just 454 and 1184 of the 329,000 and 383,708 individual data points resulted in apparent double-crossovers (Table 2). Due to the very short intermarker distances on these high-density maps, the occurrence of double-crossovers is very unlikely and probably due to genotyping errors. Assuming this to be the case, the estimated raw error rates in these populations were 0.14% and 0.31%, respectively. Interestingly, the apparent genotyping errors for both populations were concentrated in a subset of all plants analyzed, with 155 of 454 and 411 of 1184 instances occurring in just ten plants in each population (note that, even with these errors, the total error rate in these individuals was *ca.* 1%). If we were to instead assume that all of these presumptive errors were true double-crossovers, the total map lengths would have increased by 22% and 70%, respectively. The estimated error rate for these maps was thus far lower than an estimated 2.4% level observed in the 10,000 locus AFLP-based map of potato (van Os *et al.* 2006), demonstrating the high reproducibility of SNP calling with Illumina chips.

Although allele calling in the HA412-HO × RHA415 and HA412-HO × ANN1238 populations proved to be highly accurate for the scoring of both heterozygous and homozygous genotypes, the scoring of heterozygous loci was more problematic for the RIL populations. In these populations (including both RHA280 × RHA801 and NMS373 × Hopi), DNA was pooled from four individuals per line, as has been recommended to aid in the detection of residual heterozygosity (Burr and Burr 1991). Due to the continuous nature of the Infinium data, in which genotypic scores are expressed as a ratio of signal intensity between the two alleles, the pooling of multiple individuals per RIL created some challenges when scoring loci in regions harboring residual heterozygosity. More specifically, instead of exhibiting a common ratio of signal intensities for all heterozygous individuals, the raw data reflected the relative frequency of the two alleles in the pool of four individuals, ranging from 0:8 to 8:0. As such, the cluster of heterozygous genotypes produced by Genome Studio could be somewhat diffuse, and in the extreme, lines that were heterozygous at a particular locus could have been misclassified as homozygous. Because loci exhibiting residual heterozygosity were generally rare (estimated to be 0.19% and 0.39% of all individual/locus combinations in the two RIL mapping populations analyzed herein) and tended to occur in blocks, these errors were relatively easy to detect. We recommend that researchers using a marker technology that relies on relative signal intensity (*vs.* simple presence/absence or fragment length differences, such as with PCR- or RFLP-based approaches) rely on DNA samples prepared from single individuals as opposed to pooling across individuals within lines. If using similar SNP assays for studies that involve QTL mapping of phenotypic traits in RILs which must be

Table 1 Summary of the number of loci mapped in each population and overall

Chromosome	HA412-HO × HA415	HA412-HO × ANN1238	RHA280 × RHA801	NMS373 × Hopi (All)	NMS373 × Hopi 1:1	3:1 Hopi Map	3:1 NMS373 Map	NMS373 × Hopi 1:1 (1:1) NMS373 × (1:1) Hopi	RHA280 × RHA801 PCR (Heesacker *et al.* 2008)	Combined
1	190	297	173	362	241	121			85	642
2	180	201	124	253	125	128			64	456
3	240	250	245	294	229	65			89	615
4	230	244	241	328	265	58	5		99	618
5	321	270	337	362	160	169	33		113	740
6	148	163	126	266	185	81			35	417
7	166	144	140	183	118	65			45	362
8	353	272	278	412	205	158	49		145	753
9	180	266	309	402	232	170			120	707
10	339	342	469	567	179	46	304	38	169	950
11	119	200	144	241	116	122	3		63	454
12	136	245	208	298	174	124			74	558
13	169	196	196	299	163	72	54	10	99	541
14	245	272	211	276	220	56			54	554
15	171	219	171	257	223	34			78	513
16	242	259	183	353	195	158			78	592
17	71	242	230	347	87	252	8		102	611
Total	3500	4082	3785	5500	3117	1879	456	48	1512	10,080

The numbers for RHA280 × RHA801 PCR loci include previously published SSR, SNP, and indel markers.

Table 2 Summary of crossovers observed in each mapping population and overall

Chromosome	Crossover					Centimorgan				
	HA412-HO × HA415	HA412-HO × ANN1238	RHA280 × RHA801	NMS373 × Hopi (All)	Combined	HA412-HO × HA415	HA412-HO × ANN1238	RHA280 × RHA801	NMS373 × Hopi (All)	Combined
1	177	192	190	678	1237	62.8	68.1	68.0	74.1	70.0
2	197	165	183	622	1167	69.9	58.5	65.5	67.9	65.4
3	200	195	198	697	1290	70.9	69.2	70.9	76.1	71.8
4	333	243	253	953	1782	118.1	86.2	90.5	104.1	99.7
5	306	214	250	848	1618	108.5	75.9	89.5	92.6	91.6
6	129	130	166	640	1065	45.7	46.1	59.4	69.9	55.3
7	172	200	170	659	1201	61.0	70.9	60.8	72.0	66.2
8	180	138	205	628	1151	63.8	48.9	73.4	68.6	63.6
9	261	270	253	828	1612	92.6	95.7	90.5	90.4	92.3
10	285	202	228	789	1504	101.1	71.6	81.6	86.2	85.1
11	279	226	225	907	1637	98.9	80.1	80.5	99.1	89.4
12	202	209	179	581	1171	71.6	74.1	64.1	63.5	68.3
13	169	148	219	877	1413	59.9	52.5	78.4	95.8	71.6
14	206	153	218	700	1277	73.1	54.3	78.0	76.5	70.4
15	247	185	204	796	1432	87.6	65.6	73.0	86.9	78.3
16	277	251	254	944	1726	98.2	89.0	90.9	103.1	95.1
17	198	172	252	734	1356	70.2	61.0	90.2	80.2	75.4
Total	3818	3293	3647	12,881	23,639	1353.9	1167.7	1305.1	1406.9	1309.6
Bins represented	855	1008	909	1554						
Bin size	1	1	1	1		0.355	0.355	0.358	0.109	
Progeny	94	94	70	153	411					
Recombination factor	1.5	1.5	1.996	2.992						
Errors[a]	454	1184	402	9332	11,372					
Crossovers caused by errors	878	2324	804	18,664	22,670					
Error distance						311.3	824.1	287.7	2038.5	
Loci	3500	4082	3785	5500	10,080					
Data points	329,000	383,708	264,950	841,500	1,819,158					
Error rate	0.14%	0.31%	0.15%	1.11%	0.63%					

[a] Scores at a single locus that would add a two crossover events to the map.

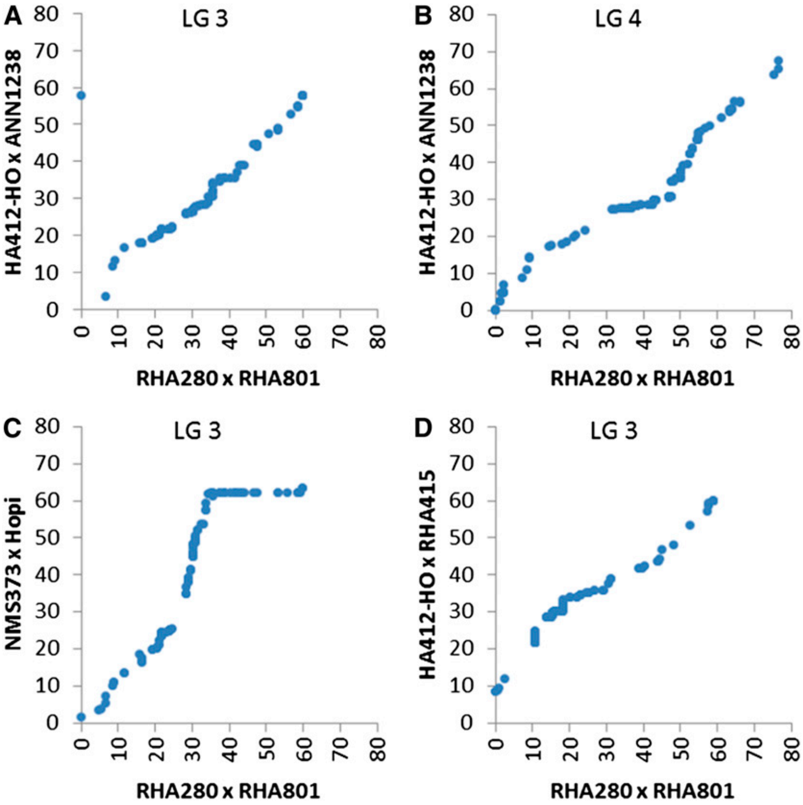

Figure 1 Comparison of relative recombination rates for selected chromosome/populations. (A) LG 3 for HA412-HO × ANN1238. (B) LG4 for HA412-HO × ANN1238. (C) LG3 for NMS373 × Hopi. (D) HA412-HO × RHA415.

scored in replicate, we recommend genotyping the parental individual from the previous RIL generation to reliably identify segregating loci.

Relative recombination rates, map density, and development of a consensus map

The use of a common genotyping array across multiple mapping populations allowed us to investigate variation in local recombination rates among mapping populations (Figure 1 and Figure S4). In most cases, the different maps were largely colinear with an average of 88.7% of all shared loci being syntenic in pairwise comparisons. Although the markers on the same chromosome in different crosses generally showed similar orders, in several cases, large chromosomal regions in one or more crosses appeared to have suppressed recombination. This was most evident in the HA412-HO × ANN1238 cross on chromosomes 6, 13, and 15, and in the RHA280 × RHA801 and NMS373 × Hopi crosses for chromosome 3. This could be the result of localized chromosomal rearrangements between the parents that inhibit recombination. Not surprisingly, the cross that showed the most and largest regions of reduced recombination was the only cross that involved wild sunflower (ANN1238). Similar cross-dependent recombination suppression has been previously observed in maize (McMullen *et al.* 2009).

The use of multiple mapping populations also allowed examination of relative marker densities. The local marker density in specific regions often varied across maps, whereas the overall marker density on the individual maps was roughly similar (Figure 2, A and B). The most notable examples of reduced marker density were on the HA412-HO × RHA415 map. This map contained several regions spanning as much as 26 cM that completely lacked mapped loci, even though the other maps contained many mapped loci in these same regions. The most striking example was on LG 17 from 37–63 cM, which was devoid of polymorphism in the HA412-HO × RHA415 map even though the RHA280 × RHA801 and HA412-HO × ANN1238 maps contained 155 and 162 loci, respectively, in that same region. The most likely explanation for this pattern is recent common ancestry, such that the mapping parents sometimes shared genomic regions that were identical by descent. Notably, HA412-HO and RHA415 are both oilseed lines and might thus be expected to exhibit a higher level of overall genetic similarity. The RHA280 × RHA801 (oilseed × confectionery) and HA412-HO × ANN1238 (oilseed × wild) maps exhibited fewer and shorter marker-poor regions. These marker-poor regions are unlikely due to ascertainment bias, as the SNPs on the chips were designed based on deep EST sequencing of all of the parents of these three crosses (Bachlava *et al.* 2012). Spikes in marker density could potentially be attributed to centromeric regions (Xu *et al.* 1997); however, as several chromosomes contain multiple regions with high marker density, other factors must be involved.

The NMS373 × Hopi map was also variable with respect to marker density, as numerous regions stretching several centimorgans in length in the base (1:1) map completely lacked mappable loci despite these same regions being marker dense on maps derived from the other crosses. However, most of these same regions were marker dense on the ancillary (3:1) maps constructed from the same cross. One potential cause of this pattern would be limited introgression from (or to) modern oilseed cultivars related to NMS373 into the Hopi landrace. These putatively introgressed regions were heterozygous in the Hopi mapping parent, with one allele being identical to the NMS373 allele, resulting in markers that segregated 3:1 and were thus placed on the ancillary linkage groups. As noted above, however, this more complex crossing design allowed for a much larger number of loci to be mapped than otherwise would have been possible. Indeed, had this population been derived from a single F_1, half of these regions would have gone unmapped because of the apparent identity by descent of these regions between NMS373 and one of the Hopi parental chromosomes.

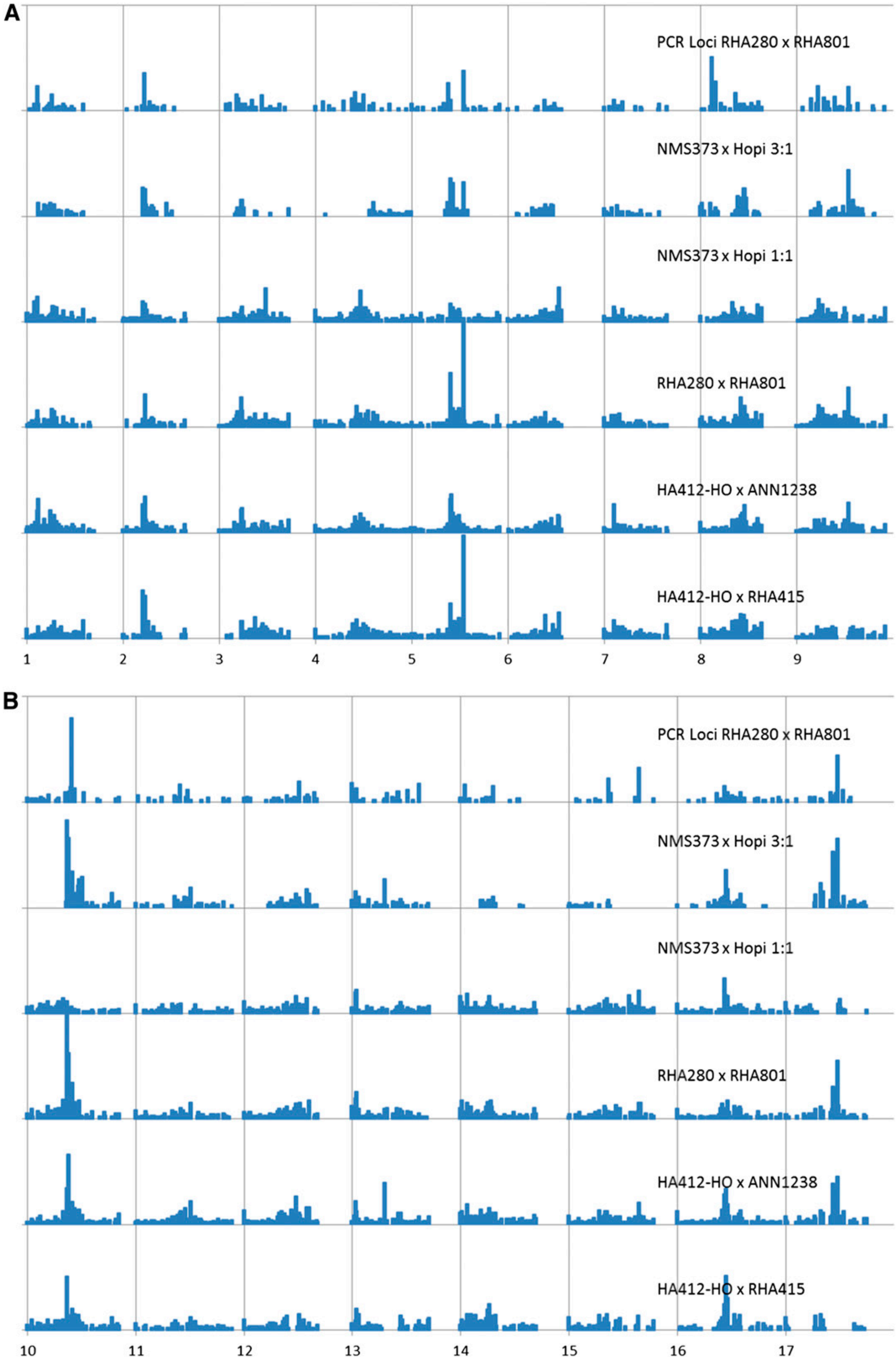

Figure 2 Graphical depiction of marker density across chromosomes expressed as markers per centimorgan (maximum of 102). (A) Chromosomes 1–9. (B) Chromosomes 10–17.

For a subset of assays, we mapped loci to different locations in different crosses. More specifically, of the 5694 SNPs tested that could be mapped in more than one population, 762 mapped to two different chromosomal locations, and 21 mapped to three different locations. The most likely explanation for this phenomenon is that these assays were derived from multicopy genes and we mapped different paralogs in different crosses. Assuming this to be true, it appears that slightly less than 14% of the SNPs tested in sunflower are derived from multicopy genes, with the two copies being similar enough to both hybridize to the assay sequence. Some of these multicopy sequences may represent instances of copy number variation, with an extra copy being present in just a subset of the parental lines, as many could be mapped due to the presence of a segregating null allele.

By using multiple crosses, we were also able to produce a consensus map containing far more loci than were present on any of the component maps and included loci that were mapped at a 3:1 ratio in the NMS373 × Hopi cross. Whereas the individual maps contained 3500–5500 SNP loci and spanned 1168–1407 cM, the combined map contained 10,080 loci, including 1512 from previous work, and spanned 1310 cM (Tables 1 and 2 and Figure S5). This combined map offers much more complete genome coverage than would be possible with any individual mapping population, even if they were genotyped with a much denser array, due to the occurrence of regions exhibiting suppressed recombination and/or identity by descent in any particular cross.

Conclusions and future directions

The availability of high-throughput SNP genotyping platforms has made possible the generation of vast amounts of largely error-free data. In the present case, we were able to generate over 5 million molecular data points from four mapping populations, a feat that would have been prohibitively expensive using traditional genotyping approaches. This study demonstrates the value of simultaneously analyzing multiple mapping populations, which allowed for the mapping of many more markers than would have otherwise been possible, including the filling of gaps that appears to have arisen due to identity by descent, and provided insight into variation in local recombination rates. The result is a high-density genetic map that has the potential to facilitate the assembly of the forthcoming sunflower genome (Kane *et al.* 2011) and to empower analyses that go far beyond what has been possible with the much lower-density genetic maps that have become commonplace, including detailed comparative genomic analyses. The availability of a large suite of genetically mapped SNPs also makes possible the efficient and detailed genotypic characterization of germplasm collections, thereby providing a means for assessing genome-wide patterns of SNP diversity in species of interest. Moreover, with a 3- to 5-fold increase in assay numbers, it may even be practical to assay variation at nearly every gene present in the *Helianthus annuus* genome.

ACKNOWLEDGMENTS

We thank Weining Tang and Mark Bouzyk at Emory University for assistance with the SNP genotyping; Jennifer Mandel, Jessica Barb, and Savithri Nambeesan for editorial and technical comments; and Jason Prothro and Jennifer Wood for assistance in the field maintaining populations. This work was supported by funding from the NSF Plant Genome Research Program (DBI-0820451), Genome Canada, Advanta Semillas, Dow Agrosciences, Pioneer Hi-Bred, Syngenta, and the Georgia Research Alliance.

LITERATURE CITED

Al-Chaarani, G. R., L. Gentzbittel, X. Q. Huang, and A. Sarrafi, 2004 Genotypic variation and identification of QTLs for agronomic traits, using AFLP and SSR markers in RILs of sunflower (Helianthus annuus L.). Theor. Appl. Genet. 109: 1353–1360.

Bachlava, E., C. A. Taylor, S. Tang, J. E. Bowers, J. R. Mandel *et al.*, 2012 SNP discovery and development of a high-density genotyping array for sunflower. PLoS ONE 7: e29814.

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, W. Michelmore *et al.*, 2008 Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25: 2445–2455.

Berry, S. T., A. J. Leon, C. C. Hanfrey, P. Challis, A. Burkholz *et al.*, 1995 Molecular marker analysis of Helianthus-annuus L. 2. Construction of an RFLP linkage map for cultivated sunflower. Theor. Appl. Genet. 91: 195–199.

Brondani, C., P. H. N. Rangel, R. P. V. Brondani, and M. E. Ferreira, 2002 QTL mapping and introgression of yield-related traits from Oryza glumaepatula to cultivated rice (Oryza sativa) using microsatellite markers. Theor. Appl. Genet. 104: 1192–1203.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718.

Burke, J. M., S. Tang, S. J. Knapp, and L. H. Rieseberg, 2002 Genetic analysis of sunflower domestication. Genetics 161: 1257–1267.

Burr, B., and F. A. Burr, 1991 Recombinant inbreds for molecular mapping in maize - theoretical and practical considerations. Trends Genet. 7: 55–60.

Crites, G. D., 1993 Domesticated sunflower in 5th millennium BP temporal context - new evidence from middle Tennessee. Am. Antiq. 58: 146–148.

Crow, J. F., 2007 Haldane, Bailey, Taylor and recombinant-inbred lines. Genetics 176: 729–732.

Frary, A., T. C. Nesbitt, S. Grandillo, E. van der Knaap, B. Cong *et al.*, 2000 fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85–88.

Gale, M. D., and K. M. Devos, 1998 Plant comparative genetics after 10 years. Science 282: 656–659.

Gedil, M. A., C. Wye, S. Berry, B. Segers, J. Peleman *et al.*, 2001 An integrated restriction fragment length polymorphism - amplified fragment length polymorphism linkage map for cultivated sunflower. Genome 44: 213–221.

Gentzbittel, L., F. Vear, Y. X. Zhang, A. Berville, and P. Nicolas, 1995 Development of a consensus linkage RFLP map of cultivated sunflower (Helianthus-annuus L). Theor. Appl. Genet. 90: 1079–1086.

Grant, D., P. Cregan, and R. C. Shoemaker, 2000 Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. USA 97: 4168–4173.

Gupta, P. K., S. Rustgi, and R. R. Mir, 2008 Array-based high-throughput DNA markers for crop improvement. Heredity 101: 5–18.

Harter, A. V., K. A. Gardner, D. Falush, D. L. Lentz, R. A. Bye *et al.*, 2004 Origin of extant domesticated sunflowers in eastern North America. Nat. 430: 201–205.

Harushima, Y., M. Yano, P. Shomura, M. Sato, T. Shimano *et al.*, 1998 A high-density rice genetic linkage map with 2275 markers using a single F-2 population. Genetics 148: 479–494.

Heesacker, A., V. K. Kishore, W. Gao, S. Tang, J. M. Kolkman *et al.*, 2008 SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. Theor. Appl. Genet. 117: 1021–1029.

Kane, N. C., N. Gill, M. G. King, J. E. Bowers, H. Berges *et al.*, 2011 Progress towards a reference genome for sunflower. Botany-Botanique 89: 429–437.

Kaul, S., H. L. Koo, J. Jenkins, M. Rizzo, T. Rooney *et al.*, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Lai, Z., K. Livingstone, Y. Zou, S. A. Church, S. J. Knapp *et al.*, 2005 Identification and mapping of SNPs from ESTs in sunflower. Theor. Appl. Genet. 111: 1532–1544.

Lorieux, M., M. N. Ndjiondjop, and A. Ghesquiere, 2000 A first interspecific Oryza sativa × Oryza glaberrima microsatellite-based genetic linkage map. Theor. Appl. Genet. 100: 593–601.

Magallon, S., P. R. Crane, and P. S. Herendeen, 1999 Phylogenetic pattern, diversity, and diversification of eudicots. Ann. Mo. Bot. Gard. 86: 297–372.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. H. Li *et al.*, 2009 Genetic properties of the maize nested association mapping population. Science 325: 737–740.

Moore, G., K. M. Devos, Z. Wang, and M. D. Gale, 1995 Cereal genome evolution-grasses, line up and form a circle. Curr. Biol. 5: 737–739.

Murray, M. G., and W. F. Thompson, 1980 Rapid isolation of high molecular-weight plant DNA. Nucleic Acids Res. 8: 4321–4325.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551–556.

Qi, L. L., B. S. Hulke, B. A. Vick, and T. J. Gulya, 2011 Molecular mapping of the rust resistance gene R(4) to a large NBS-LRR cluster on linkage group 13 of sunflower. Theor. Appl. Genet. 123: 351–358.

Rieseberg, L. H., H. C. Choi, R. Chan, and C. Spore, 1993 Genomic map of a diploid hybrid species. Heredity 70: 285–293.

Smith, B. D., and R. A. Yarnell, 2009 Initial formation of an indigenous crop complex in eastern North America at 3800 BP. Proc. Natl. Acad. Sci. USA 106: 6561–6566.

Soleri, D., and D. A. Cleveland, 1993 Hopi crop diversity and change. Journal of Ethnobiology 13: 203–231.

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.

Tang, S. X., A. Leon, W. C. Bridges, and S. J. Knapp, 2006 Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower. Crop Sci. 46: 721–734.

Tanksley, S. D., and J. C. Nelson, 1996 Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor. Appl. Genet. 92**:** 191–203.

Tanksley, S. D., and S. R. McCouch, 1997 Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277: 1063–1066.

van Os, H., S. Andrzejewski, E. Bakker, I. Barrena, G. J. Bryan *et al.*, 2006 Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173: 1075–1087.

Wills, D. M., and J. M. Burke, 2007 Quantitative trait locus analysis of the early domestication of sunflower. Genetics 176: 2589–2599.

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.

Wu, Y., P. R. Bhat, T. J. Close, and S. Lonardi, 2008 Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet. 4: e1000212.

Xu, X., S. K. Pan, S. F. Cheng, B. Zhang, D. S. Mu *et al.*, 2011 Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.

Xu, Y., L. Zhu, J. Xiao, N. Huang, and S. R. McCouch, 1997 Chromosomal regions associated with segregation distortion of molecular markers in F-2, backcross, doubled haploid, and recombinant inbred populations in rice (Oryza sativa L). Mol. Gen. Genet. 253: 535–545.

Yu, J. K., S. Tang, M. B. Slabaugh, A. Heesacker, G. Cole *et al.*, 2003 Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Sci. 43: 367–387.

Communicating editor: J. Wendel