Progress towards a reference genome for sunflower

N.C. Kane, 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

Abstract: The Compositae is one of the largest and most economically important families of flowering plants and includes a diverse array of food crops, horticultural crops, medicinals, and noxious weeds. Despite its size and economic importance, there is no reference genome sequence for the Compositae, which impedes research and improvement efforts. We report on progress toward sequencing the 3.5 Gb genome of cultivated sunflower (*Helianthus annuus*), the most important crop in the family. Our sequencing strategy combines whole-genome shotgun sequencing using the Solexa and 454 platforms with the generation of high-density genetic and physical maps that serve as scaffolds for the linear assembly of whole-genome shotgun sequences. The performance of this approach is enhanced by the construction of a sequence-based physical map, which provides unique sequence-based tags every 5–6 kb across the genome. Thus far, our physical map covers ~85% of the sunflower genome, and we have generated ~80× genome coverage with Solexa reads and 15.5× with 454 reads. Preliminary analyses indicated that ~78% of the sunflower genome consists of repetitive sequences. Nonetheless, ~76% of contigs >5 kb in size can be assigned to either the physical or genetic map or to both, suggesting that our approach is likely to deliver a highly accurate and contiguous reference genome for sunflower.

Key words: Compositae, *Helianthus annuus*, next generation sequencing, physical map, reference genome, sunflower.

Résumé : Les Compositae constituent une des familles de plantes à fleurs les plus grandes et une des plus importantes économiquement. Elle inclut un ensemble diversifié d'aliments cultivés, de plantes médicinales ainsi que de plantes adventices nuisibles. En dépit de sa dimension et de son importance économique, il n'existe pas de séquence de référence du génome pour les Compositae, ce qui nuit à la recherche et aux efforts d'amélioration. Les auteurs font état de progrès en vue du séquençage du génome de 3.5 Gb du tournesol cultivé (*Helianthus annuus*), la culture la plus importante de la famille. Leur stratégie de séquençage combine le séquençage « whole-genome-shotgun » (WGS) en utilisant les plateformes Solexa et 454 avec la génération de carte génétique et physique à haute densité servant d'échafaudages pour l'assemblage linéaire des séquences WGS. La performance de cette approche se voit renforcée par la construction d'une carte physique basée sur les séquences, laquelle fournit des étiquettes uniques basées sur des séquences à toutes les 5–6 kb sur l'ensemble du génome. Jusqu'ici, la carte physique couvre ~85% du génome du tournesol, et les auteurs ont généré une couverture ~80× du génome avec les lectures du Solexa et 15.5× avec les lectures du 454. Des analyses préliminaires indiquent que ~78% du génome du tournesol est constitué de séquences répétitives. Cependant, on peut assigner environ 76% des séquences contiguës >5 kb à une carte physique ou génétique, ou aux deux, ce qui suggère que l'approche des auteurs est susceptible de conduire à un génome de référence très précis et contiguë pour le tournesol.

Mots-clés : Compositae, *Helianthus annuus*, séquençage de la prochaine génération, carte physique, génome de référence, tournesol.

[Traduit par la Rédaction]

Received 18 February 2011. Accepted 22 April 2011. Published at www.nrcresearchpress.com/cjb on 3 August 2011.

N.C. Kane, N. Gill, M.G. King, and M. Stewart. Department of Botany, The University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada.
J.E. Bowers and J.M. Burke. Department of Plant Biology, University of Georgia, 2502 Miller Plant Sciences, Athens, GA 30602, USA.
H. Berges. Institut national de la recherche agronomique – Centre national de ressources génomique végétales, chemin de Borde Rouge, B.P. 52627, 31326 Castanet Tolosan, France.
J. Gouzy, N.B. Langlade, and P. Vincourt. Institut national de la recherche agronomique – Centre national de la recherche scientifique, Laboratoire des interactions plantes micro-organismes, chemin de Borde Rouge, B.P. 52627, 31326 Castanet Tolosan, France.
E. Bachlava and S.J. Knapp. Monsanto Vegetable Seeds, 37437 State Highway 16, Woodland, CA 95695, USA.
Z. Lai. Center for Genomics and Bioinformatics, Indiana University, 915 East Third Street, Bloomington, IN 47405, USA.
L.H. Rieseberg. Department of Botany, The University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada, and Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA.

Corresponding author: L.H. Rieseberg (e-mail: loren.rieseberg@botany.ubc.ca).

Introduction

The Compositae, or sunflower family (Asteraceae), is the largest plant family on earth, with over 24 000 described species, representing roughly 10% of all flowering plant species (Stevens 2010). Compositae species live on every continent except Antarctica and are found in a huge diversity of habitats, including forests, grasslands, deserts, wetlands, mountaintops, salt marshes, lawns, and agricultural fields (Funk et al. 2009). They can grow as herbs, shrubs, trees, or even vines. They include economically important crops, rare and beautiful wildflowers, common allergens, valuable medicinals, and costly invasive plants and rangeland weeds (Dempewolf et al. 2008). They are referred to as Composites because what looks like a single large flower is actually a composite of many tiny flowers, sometimes thousands. Some of the more well-known Composites include sunflowers, lettuce, artichokes, dandelions, thistles, daisies, ragweed, goldenrod, and chicory.

Despite the wide diversity and economic importance of plants in this family, there is no genome sequence for any of these species or for any plants from closely related families. This has slowed genetic research and crop breeding and made many experiments more difficult or impossible. Because the genomes of Compositae crops are quite large, it has until recently been considered overly costly and impractical to sequence them. For instance, the sunflower genome is ~3.5 billion bases long (Baack et al. 2005), slightly longer than the human genome, which cost roughly US$3 billion to sequence the first time. However, DNA sequencing technology has advanced dramatically in the past decade, making it more practical and much less expensive to sequence and assemble a new genome (Mardis 2008; Schatz et al. 2010).

Here we report on our progress towards developing a high-quality "reference" genome for sunflower, *Helianthus annuus* L., which ranked eleventh in 2008 among the world's food crops in terms of area harvested (http://www.fao.org/). Sunflower's economic importance, excellent germplasm resources, and highly developed genetic toolkit make it a logical candidate for large-scale sequencing within the Compositae. Sunflower is also the only major crop to have been domesticated in North America (Harter et al. 2004). A whole-genome sequence for sunflower will provide a reference genome for resequencing studies and will increase the utility of the extensive sunflower expressed sequence tag resources (Lai et al. 2005*a*; Church et al. 2007; Barker et al. 2008; Heesacker et al. 2008), revealing gene locations, gene family size, promoter and intron sequences, relationships between paralogs, and alternative splicing. A sunflower genome will also serve as a useful reference for studies of genome evolution in *Helianthus* (Rieseberg et al. 1993, 1995, 2003; Burke et al. 2004; Baack et al. 2005; Lai et al. 2005*b*, 2006; Ungerer et al. 2006; Heesacker et al. 2009), as well as between *Helianthus* and other species in the Compositae (Natali et al. 2006; Timms et al. 2006).

Overall strategy

Several considerations influenced our strategy for sequencing the sunflower genome. First, we were aiming for a gold standard reference sequence in terms of both sequence accuracy and contiguity. The latter may not be critical for many sequencing projects, but it is important for crop plants where many applications require knowledge of the precise chromosomal location of the sequence of interest (e.g., Blackman et al. 2010, 2011; Wieckhorst et al. 2010). Second, the sunflower genome is very large with abundant repetitive sequences (Cavallini et al. 2009), so an approach based exclusively on whole-genome shotgun (WGS) of second-generation sequence data is unlikely to reliably link, order, and orient sequenced contigs (Rounsley et al. 2009; Schatz et al. 2010). Third, funds available for sequencing the sunflower genome are limited, so we could not afford clone-by-clone bacterial-artificial-chromosome-based (BAC-based) sequencing, such as that employed for the *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), *Oryza sativa* subsp. *japonica* (International Rice Genome Sequencing Project 2005), and *Zea mays* (Schnable et al. 2009) genome projects. Even the relatively low coverage of 6× WGS obtained with Sanger sequencing is not affordable for sunflower because of the large size of its genome.

Given these considerations, we have developed a hybrid approach that combines WGS sequencing using the Solexa platform (acquired by Illumina) and 454 Life Sciences platform (acquired by Roche) with the generation of high-density genetic and physical maps that can serve as scaffolds for the linear assembly of WGS sequences. The performance of this approach is enhanced by the construction of a sequence-based physical map, which will provide unique sequence-based markers every 5–6 kb across the genome. The combination of Solexa and 454 reads is effective because the two sequencing platforms bring different strengths: Solexa brings great depth, but cannot bridge regions of low complexity, whereas 454 reads can span longer repeats, but their higher cost makes deep coverage expensive (Dalloul et al. 2010; Schatz et al. 2010). A final finishing step involves targeted sequencing of individual BACs or BAC pools from incomplete or poorly assembled genomic regions. This strategy should provide a reference genome that is equivalent in accuracy and connectivity to one generated by high-coverage Sanger sequencing, but at a small fraction of the cost. Our approach thus takes advantage of second-generation sequencing platforms while avoiding the pitfalls associated with the de novo assembly of large and highly repetitive plant genomes from the relatively short reads generated by these platforms.

A similar approach has been used to successfully sequence the 1.1 Gb turkey genome (Dalloul et al. 2010), except that the turkey genome project employed traditional fingerprinting and BAC end sequencing for physical map construction. Nonetheless, they were able to assign 917 Mb of sequence (83% of the genome) to specific turkey chromosomes.

Genetic mapping

A dense genetic map for sunflower is expected to greatly facilitate the placement and orientation of sequence contigs. Prior to this project, approximately 2000 sequence-based markers had been mapped to 17 linkage groups in sunflower, a number that corresponds to the haploid chromosome number of the species (http://www.sunflower.uga.edu/cmap/; Berry et al. 1995; Gentzbittel et al. 1995, 1999; Jan et al. 1998; Gedil et al. 2001; Burke et al. 2002, 2004; Tang et al. 2002; Yu et al. 2003; Lai et al. 2005*a*; Chapman et al. 2008; Heesacker et al. 2009). Mapping efforts currently underway

should provide close to a 10-fold increase in the density of the genetic map.

Two gene chips have been constructed for genetic mapping: (*i*) a 10 640 single-nucleotide repeat (SNP) Infinium array developed by researchers at the University of Georgia in collaboration with Advanta Seeds, Dow Agrosciences, Syngenta, and Pioneer Hi-Bred; and (*ii*) a 2.56 million-feature Affymetrix chip developed by a consortium of researchers at the French National Institute for Agricultural Research (INRA), The University of British Columbia, the University of Georgia, Syngenta AG, and Biogemma.

To date, four populations of *H. annuus* have been genotyped with the Infinium array (J. Bowers et al., unpublished data). Two of these are intermated F2 populations that involve the cultivar we are sequencing, HA412-HO, a high-oleic oilseed line (Miller et al. 2006). These populations will be especially useful for scaffolding, because intermating permits very high resolution genetic mapping. Between 3500 and 4100 markers have been placed on each map, and >70% of the 10 640 SNPs have been placed on at least one of the four maps. An integrated map created from the four separate SNP maps plus simple sequence repeats mapped previously contained over 10 000 loci, including ~1500 simple sequence repeats and ~8500 SNP loci. Because the SNP genotyping error rates were low (<1%), the resulting genetic map is of high quality and will allow for high-confidence placement of sequences on linkage groups.

The Affymetrix array is based on 284 251 expressed sequence tags from seven *Helianthus* species that were available at the National Center for Biotechnology Information in September 2007. Most of the sequences (~93%) were generated by the Compositae Genome Project (Barker et al. 2008; Heesacker et al. 2008), with the remainder coming from other sources, including Genoplante (http://www.genoplante.com/). The sequences were assembled into 87 237 unigenes, of which 8378 have ambiguous orientation. Both orientations of the latter were included for array development, so probes were developed for a total of 95 589 sequences, with an average of 27 probes per sequence. Genotyping of single-feature polymorphisms (SFPs) is underway in two recombinant inbred line populations, one at INRA (N. Langlade et al., unpublished data) and one at the University of Georgia (J. Bowers et al., unpublished data). Preliminary analyses of SFPs in the INRA recombinant inbred lines indicate that it should be possible to map at least 10 000 SFPs in this population.

Physical mapping

To develop a robust physical map for sunflower, BAC libraries were constructed for HA412-HO by the French Plant Genome Resource Center (http://cnrgv.toulouse.inra.fr/en/library/sunflower). Three different restriction enzymes (*Hin*dIII, *Bam*HI, *Eco*RI) were used for library development to provide more complete genomic coverage (Wu et al. 2004). The *Hin*dIII library includes 238 080 clones with an average insert size of 132 kb (~9× coverage); the *Bam*HI library consists of 86 400 clones with inserts averaging 114 kb (~2.6× coverage); and the *Eco*RI library contains 81 792 clones averaging 93 kb (~2.2×). Total BAC library coverage is 13.8×.

For physical mapping, we employed the sequence-based mapping approach developed by Keygene N.V., which takes advantage of the very low cost of Solexa sequencing to generate 20–30 unique sequences for each BAC clone. Briefly, sequence tags are produced from the terminal ends of restriction fragments from two-dimensional pooled BAC clones using the Illumina Genome Analyzer platform. BAC pools are tagged individually to allow assignment of sequences to individual BAC clones using the coordinates in the two-dimensional pool screening. The BAC clones can be ordered into contigs on the basis of shared sequence tags using a modified version of FPC (Soderlund et al. 2000; Engler et al. 2003), the same software used for physical mapping based on restriction profiles.

Thus far, 191 232 BAC clones from the three different libraries have been employed to develop a preliminary 6.2× physical map. The results are encouraging. Eighty-six percent (165 313) of the BACs contain unique sequence tags, with an average of 20 unique tags per BAC. A high-stringency BAC assembly (1.0×10^{-30}) includes 15 702 contigs that cover 3065 Mb, or approximately 85% of the genome. However, contig sizes are small, averaging 5.8 BACs per contig, with an N50 contig size of 8 BACs per contig. In a less stringent assembly (1.0×10^{-15}), contigs are larger: 10.9 BACs per contig on average with an N50 contig size of 15. We expect contig size to expand considerably in the final 12× map.

Sequencing

In general, the quality of an assembly will increase with longer reads and greater sequencing depth (Schatz et al. 2010). Longer reads can span more repeats, thereby increasing sequence connectivity, whereas greater sequencing depth increases the accuracy of base calls. Sequence depth can increase the connectivity of sequences as well, but it cannot compensate for situations in which repeats exceed read lengths, leading to gaps in the assembly. These gaps can be spanned by paired-end reads (reads from both ends of a single DNA fragment), as long as the distance between the paired reads is greater than the length of the gap in the assembly. Unfortunately, paired-end sequencing on second-generation platforms is technically challenging, especially for long paired-end (or mate pair) libraries, which typically are highly redundant and prone to chimera formation. In addition, it can be difficult to control insert sizes in long mate pair libraries.

Given these considerations, genome sequencing projects typically develop a mixture of paired-end libraries that vary in insert lengths and often employ sequencing platforms that generate reads of different lengths (Dalloul et al. 2010). For sunflower, we combined 101 bp Solexa reads (Fig. 1) with relatively longer 454 reads (Fig. 2). While longer Solexa reads are possible, error rates increased with sequence length (Fig. 1), which limits the utility of the longer reads. Read lengths are more difficult to control on the 454 platform. For example, with the standard 200 cycles, read lengths range from 50 to 600 bp, with a median read length of ~400 bp (Fig. 2). However, read lengths can be increased substantially with more cycles, and these longer reads (in the 600–800 bp range) are available at Roche's Service Centre as part of an early access program. With respect to library development,

Fig. 1. Example of Solexa data showing increase in error rate with position along read. Mate-pair library with 2 kb inserts.

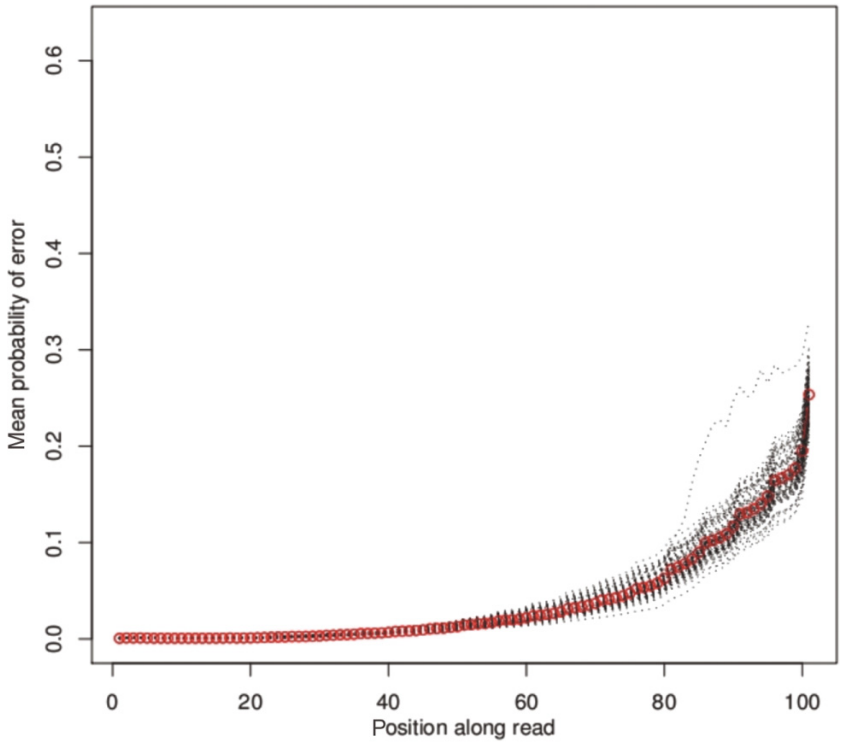
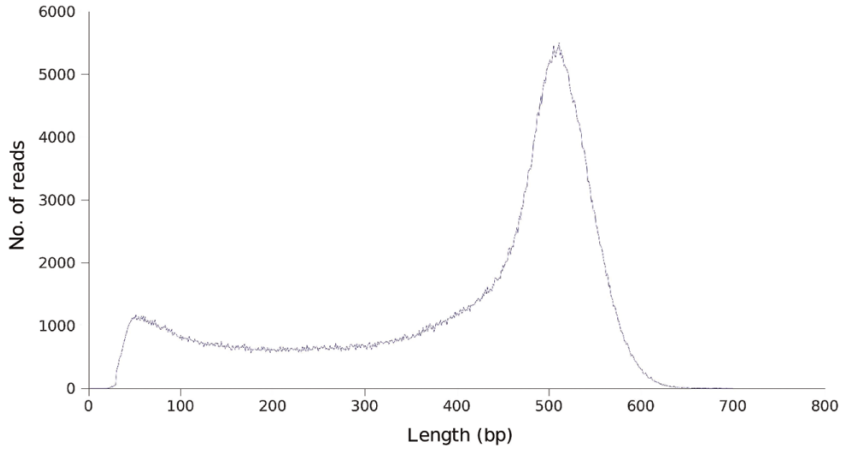


Fig. 2. Distribution of read lengths in a typical 454 sequencing run.

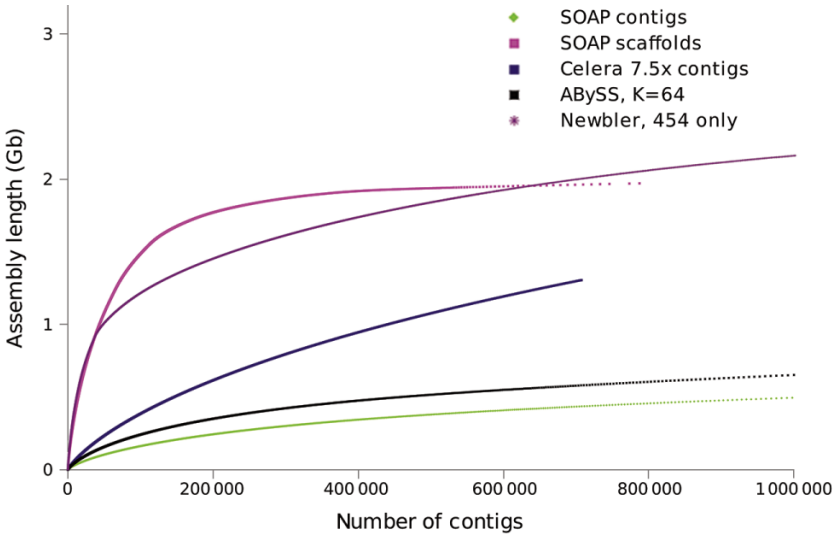


we developed Solexa libraries for five different insert sizes (0.2, 0.5, 2, 5, and 10 kb). Both single-read and paired-end libraries with 3, 6, 8, 10, and 14 kb inserts have been constructed for 454 sequencing.

As of March 2010, we had obtained a total of 304 Gb of Solexa sequence, including 35× 200 bp inserts, 35× 500 bp inserts, 7.5× 2 kb inserts, 2.8× 5 kb inserts, and 2.9× 10 kb inserts. After filtering to only include full-length reads (i.e., pairs where no bases had quality <4), which will be the most useful for de novo assembly, and removing redundant reads (identical copies), we have a total of 28× coverage of Solexa sequence. While this provides sufficient depth for a high-quality assembly, we are preparing additional long mate-pair libraries because of high levels of redundancy in the libraries sequenced thus far. We also have 35.8 Gb short single end, 9.5 Gb long single read (from the Roche early access program), and 9.0 Gb paired-end 454 sequence, for a total of 15.5× coverage with 454 sequence. We are aiming for 20× depth with 454 reads, which we expect to achieve by mid-2011.

In addition to the WGS sequencing, cDNA from 10 tissues has been sequenced as paired-end libraries using Illumina mRNA Seq protocols. Sequencing of small RNA from the same samples is under way. We obtained between 4.4 and 10.5 Gb of trimmed transcriptomic data for each tissue: stems, leaves, roots, bracts, ray floret ligules, disc floret ligules, ovaries, seeds, pistils (pooled stigmas and styles), stamens (pooled anthers and filaments), and pollen. The assembly of these samples will allow us to identify nearly all sunflower transcripts and therefore greatly improve the genome annotation. Lastly, we sequenced over 100 BACs to validate the WGS contigs.

Fig. 3. Summary of preliminary assemblies of the sunflower genome.



Assembly

The assembly of large eukaryotic genomes is challenging because of the repeat structure of these genomes, as well as the computational requirements for handling very large numbers of reads. For these reasons, methods have been developed that partition assemblies into distinct phases (Schatz et al. 2010). Typically, an initial phase assembles reads with unambiguous overlaps into "unitigs" (Myers et al. 2000), whereas later phases use mate-pair information to assemble the unitigs into larger contigs and then the contigs into scaffolds (scaffolds represent contigs that are linked by mate pairs, but separated by repeats or sequence gaps).

For assembly of the sunflower genome, we have been testing several different assemblers that implement these phases in different ways. These include the Celera Genome Assembler (Myers et al. 2000) and NEWBLER (Margulies et al. 2005), which use an "overlap–layout–consensus" approach and perform well for longer reads (>100 bp), as well as the ABySS (Simpson et al. 2009) and SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) assemblers, which are de Bruijn Graph methods best suited for short reads (Simpson et al. 2009; Schatz et al. 2010). We have generated very preliminary assemblies with all four assemblers (Fig. 3), with the most complete assemblies coming from SOAPdenovo for the Illumina data and the Newbler Assembler for the 454 data. However, even the most complete assemblies generated to date include several hundred thousand scaffolds and cover approximately 60% of the genome.

There are several factors contributing to the incomplete assemblies produced to date. First, they include only a subset of the data (either Solexa or 454 data), and we suspect that combined assemblies that include both 454 and Solexa data will be much more complete. We are currently working on ways to combine and scaffold contigs generated from different assemblies and sequencing platforms. Second, the assemblies are based on only 50% (Celera) or 75% (Newbler) of the 454 WGS data to be generated by the project. Third, the amount of long paired-end sequence is currently quite limited (and highly redundant), so a priority during the spring of 2011 is to generate more long paired-end or mate-pair reads on both the 454 and Illumina platforms. Finally, we have only just begun to use the unique sequence tags in our BAC contigs to scaffold the sequence data. However, this approach appears to be very powerful and should allow us to generate very large scaffolds (see below).

Despite the preliminary nature of our assemblies, we do have enough contigs to evaluate the value of the genetic and physical maps for creating scaffolds and for assigning sequences to sunflower chromosomes. Comparisons with the Celera assembly data indicate that a substantial fraction of contigs can be assigned to either the genetic or physical map or to both (Fig. 4). Currently, 7% of contigs >5 kb can be mapped to the genetic map, and 73% of contigs >5 kb contain at least one unique sequence from the physical map. When considered together, 76% of contigs >5 kb can be assigned to at least one of the two maps. This indicates that a large proportion of the final contigs should be mappable, greatly facilitating creation of large scaffolds and improving the quality of the final assembly.

Repeat structure of the sunflower genome

A major challenge associated with sequencing large eukaryotic genomes is the abundance of repetitive sequences. Long repeats that exceed the length of sequence reads are especially problematic for assembly programs. Recent expansions of repeat families can be troublesome as well, since there has not been sufficient time for mutational differences to accumulate among repeats.

A preliminary analysis indicated that 78.5% of the sunflower genome consists of repetitive sequences (Table 1), which is higher than the previously reported estimate of ~62% (Cavallini et al. 2009). Our analysis was based on a 5.8% random sample of the genome (207 534 291 bp, excluding gaps). De novo repetitive sequences were identified using RepeatScout (Price et al. 2005), annotated using The Institute for Genomic Research (TIGR) all plant repeat database (Ouyang and Buell 2004), and used as a custom library for RepeatMasker (http://repeatmasker.org; A.F.A. Smit and P. Green, RepeatMasker, version 3.1.9). It is likely that the high fraction of repeats characterized as novel in sunflower

Fig. 4. Proportion of contigs in draft assembly that contain sequences employed for genetic and physical map construction.

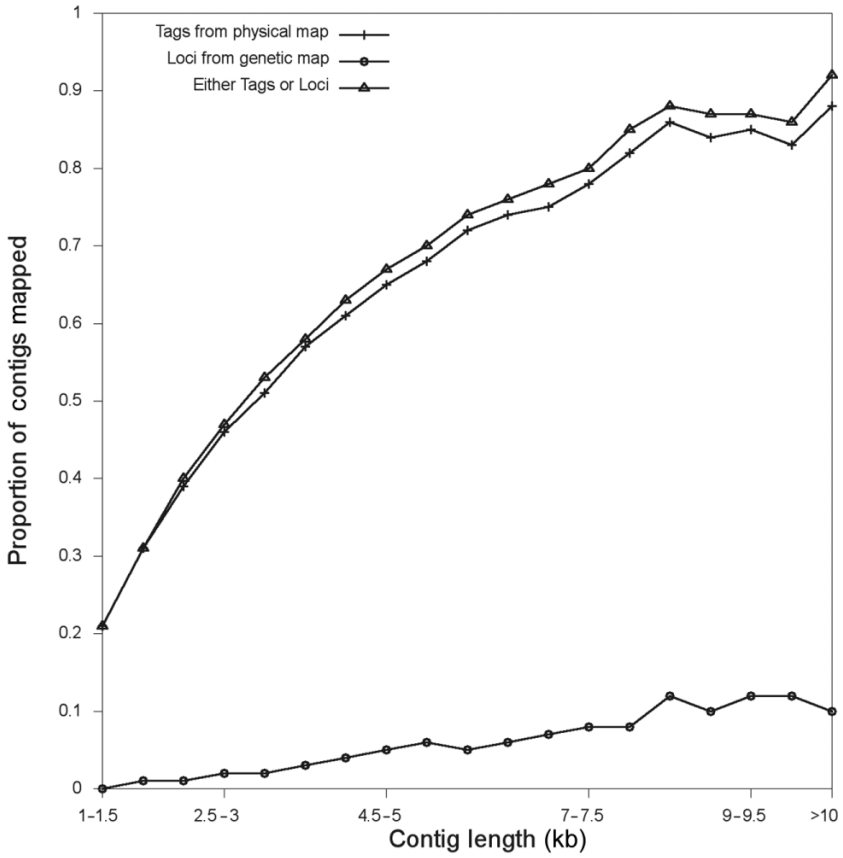


Table 1. Repeat composition of the sunflower genome.

Repeat class	Repeat type	Total sequence length (bp)	% of the genome*	% of total repetitive
Class I elements	LTR retrotransposon	6 591 789	3.18	4.05
	Non-LTR retrotransposon	148 963	0.07	0.09
	Unclassified	4 322 702	2.08	2.65
	Class I subtotal	**11 063 454**	**5.33**	**6.79**
Class II elements	Non-MITEs	541 951	0.26	0.33
	MITEs	17 664	0.01	0.01
	Class II subtotal	**559 615**	**0.27**	**0.34**
Centromeric		324 647	0.16	0.20
Telomeric		44 988	0.02	0.03
Ribosomal		274 270	0.13	0.17
Low-complexity sequences		2 465 632	1.19	1.51
Simple repeats		657 013	0.32	0.40
Unclassified[†]		4 896 654	2.36	3.01
Sunflower novel[‡]		142 617 437	68.72	87.55
Total for all repeat classes		**162 903 710**	**78.49**	

*Percentages are based on a random 5.8% sample of the sunflower genome.

[†]Unclassified repeats present in the TIGR all plant repeat database.

[‡]Novel repeats present in the sunflower genome that lack similarity to TIGR plant repeats.

is an overestimation, because many of these elements could not be annotated owing to the fragmentary nature of the data.

Overall, our results indicate that sunflower is one of the highly repetitive large plant genomes, such as those of wheat and barley, with 70%–80% repetitive content (Rostoks et al. 2002; Wicker et al. 2002, 2007). Sunflower repeat content is also considerably higher as compared with that of other model flowering plants such as *Arabidopsis*, rice, and maize, where the repetitive content (primarily transposable elements) is at least 10%, 35%, and 66%, respectively (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Haberer et al. 2005).

Chloroplast genome

We assembled the chloroplast genome from the 454 WGS genome based on 58 348 reads and 165× average coverage.

Because pyrosequencing is prone to homopolymer errors, putative indel differences were confirmed by mapping four lanes of Illumina WGS data onto the chloroplast genomes of HA383 and HA412. The HA412 chloroplast genome is of course very similar to the HA373 chloroplast genome sequence in GenBank. There are a total of three bases that are different, based on our assembly, which excludes one copy of the inverted repeat. Differences include two 1 bp indels in mononucleotide repeats, one between *ndhD* and *ccsA* and one between *atpF* and *atpA*, and one SNP in the intergenic region between *atpB* and *rbcL*. However, at the majority of mononucleotide repeats, some of which are exceedingly long, the two chloroplasts are identical.

Conclusions

In summary, we have made significant progress toward the generation of a fully sequenced sunflower genome. We are optimistic that our strategy, which combines WGS sequence data from multiple second-generation sequencing platforms with high-density genetic and physical maps, will prove effective for assembling large and repetitive eukaryotic genomes. Our preliminary data suggest that the sequence tags in our physical map will be especially useful for scaffolding and for assigning sequence contigs to chromosomes. While knowledge of the physical location of DNA sequences is not essential for some genome projects, many applications in crop species, including map-based cloning, marker-assisted selection, and synteny-based homology determination, require map information. Also, physical and genetic mapping data can greatly reduce the misassembly of recently duplicated regions.

While several other genome projects are underway in the Compositae (R. Michelmore, personal communication, 5 June 2010; D. Soltis, personal communication, 6 October 2010; N. Stewart, personal communication, 22 January 2010), as far as we are aware only for sunflower will both genetic and physical mapping data be available for assembly and scaffolding. Thus, the sunflower genome will be a fundamentally important resource, enabling major advances for sunflower and the entire Compositae and providing the data necessary for functional and comparative genomic analyses related to agricultural, biological, and environmental research.

Acknowledgements

This research was supported by Genome Canada, Genome British Columbia, the French National Institute for Agricultural Research (INRA), the US National Science Foundation (award 0421630), and the US Department of Agriculture (2008-35300-19263, 2008-35300-04579). We thank the BAC development group at the French Plant Genomic Resource Centre, the physical mapping group at Keygene N.V., the 454 sequencing group at Genome Quebec, and the Solexa sequencing group at the Michael Smith Genome Sciences Centre for their contributions to the project.

References

Baack, E.J., Whitney, K.D., and Rieseberg, L.H. 2005. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. New Phytol. **167**(2): 623–630. doi:10.1111/j.1469-8137.2005.01433.x. PMID:15998412.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, W., Knapp, S.J., and Rieseberg, L.H. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. **25**(11): 2445–2455. doi:10.1093/molbev/msn187. PMID:18728074.

Berry, S.T., Leon, A.J., Hanfrey, C.C., Challis, P., Burkholz, A., Barnes, S.R., et al. 1995. Molecular marker analysis of *Helianthus annuus* L. 2. Construction of an RFLP linkage map for cultivated sunflower. Theor. Appl. Genet. **91**(2): 195–199. doi:10.1007/BF00220877.

Blackman, B.K., Strasburg, J.L., Raduski, A.R., Michaels, S.D., and Rieseberg, L.H. 2010. The role of recently derived FT paralogs in sunflower domestication. Curr. Biol. **20**(7): 629–635. doi:10.1016/j.cub.2010.01.059. PMID:20303265.

Blackman, B.K., Rasmussen, D.A., Strasburg, J.L., Raduski, A.R., Burke, J.M., Knapp, S.J., et al. 2011. Contributions of flowering time genes to sunflower domestication and improvement. Genetics, **187**(1): 271–287. doi:10.1534/genetics.110.121327. PMID:20944017.

Burke, J.M., Tang, S., Knapp, S.J., and Rieseberg, L.H. 2002. Genetic analysis of sunflower domestication. Genetics, **161**(3): 1257–1267. PMID:12136028.

Burke, J.M., Lai, Z., Salmaso, M., Nakazato, T., Tang, S., Heesacker, A., Knapp, S.J., and Rieseberg, L.H. 2004. Comparative mapping and rapid karyotypic evolution in the genus *Helianthus*. Genetics, **167**(1): 449–457. doi:10.1534/genetics.167.1.449. PMID:15166168.

Cavallini, A., Natali, L., Zuccolo, A., Giordani, T., Jurman, I., Ferrillo, V., et al. 2009. Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. Theor. Appl. Genet. **120**(3): 491–508. doi:10.1007/s00122-009-1170-7. PMID:19826774.

Chapman, M.A., Pashley, C.H., Wenzler, J., Hvala, J., Tang, S.X., Knapp, S.J., and Burke, J.M. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). Plant Cell, **20**(11): 2931–2945. doi:10.1105/tpc.108.059808. PMID:19017747.

Church, S.A., Livingstone, K., Lai, Z., Kozik, A., Knapp, S.J., Michelmore, R.W., and Rieseberg, L.H. 2007. Using variable rate models to identify genes under selection in sequence pairs: Their validity and limitations for EST sequences. J. Mol. Evol. **64**(2): 171–180. doi:10.1007/s00239-005-0299-5. PMID:17200807.

Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., Blomberg, L., et al. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biol. **8**(9): e1000475. doi:10.1371/journal.pbio.1000475. PMID:20838655.

Dempewolf, H., Rieseberg, L.H., and Cronk, Q.C. 2008. Crop domestication in the Compositae: a family-wide trait assessment. Genet. Resour. Crop Evol. **55**(8): 1141–1157. doi:10.1007/s10722-008-9315-0.

Engler, F.W., Hatfield, J., Nelson, W., and Soderlund, C.A. 2003. Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res. **13**(9): 2152–2163. doi:10.1101/gr.1068603. PMID:12915486.

Funk, V., Susanna, A., Stuessy, T., and Bayer, R. (*Editors*). 2009. Systematics, evolution and biogeography of the Compositae. International Association of Plant Taxonomy, Vienna, Austria.

Gedil, M.A., Wye, C., Berry, S., Segers, B., Peleman, J., Jones, R., et al. 2001. An integrated restriction fragment length polymorphism – amplified fragment length polymorphism linkage map for

cultivated sunflower. Genome, **44**(2): 213–221. doi:10.1139/g00-111. PMID:11341731.

Gentzbittel, L., Vear, F., Zhang, Y.X., Berville, A., and Nicolas, P. 1995. Development of a consensus linkage RFLP map of cultivated sunflower (*Helianthus annuus* L.). Theor. Appl. Genet. **90**(7–8): 1079–1086.

Gentzbittel, L., Mestries, E., Mouzeyar, S., Mazeyrat, F., Badaoui, S., Vear, F., et al. 1999. A composite map of expressed sequences and phenotypic traits of the sunflower (*Helianthus annuus* L.) genome. Theor. Appl. Genet. **99**(1–2): 218–234. doi:10.1007/s001220051228.

Haberer, G., Young, S., Bharti, A.K., Gundlach, H., Raymond, C., Fuks, G., et al. 2005. Structure and architecture of the maize genome. Plant Physiol. **139**(4): 1612–1624. doi:10.1104/pp.105.068718. PMID:16339807.

Harter, A.V., Gardner, K.A., Falush, D., Lentz, D.L., Bye, R.A., and Rieseberg, L.H. 2004. Origin of extant domesticated sunflowers in eastern North America. Nature (Lond.), **430**(6996): 201–205. PMID:15241413.

Heesacker, A., Kishore, V.K., Gao, W.X., Tang, S.X., Kolkman, J.M., Gingle, A., et al. 2008. SSRs and indels mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. Theor. Appl. Genet. **117**(7): 1021–1029. doi:10.1007/s00122-008-0841-0. PMID:18633591.

Heesacker, A.F., Bachlava, E., Brunick, R.L., Burke, J.M., Rieseberg, L.H., and Knapp, S.J. 2009. Karyotypic evolution of the common and silverleaf sunflower genomes. Plant Genome, **2**(3): 233–246. doi:10.3835/plantgenome2009.05.0015.

Jan, C.C., Vick, B.A., Miller, J.F., Kahler, A.L., and Butler, E.T., III. 1998. Construction of an RFLP linkage map for cultivated sunflower. Theor. Appl. Genet. **96**(1): 15–22. doi:10.1007/s001220050703.

Lai, Z., Livingstone, K., Zou, Y., Church, S.A., Knapp, S.J., Andrews, J., and Rieseberg, L.H. 2005*a*. Identification and mapping of SNPs from ESTs in sunflower. Theor. Appl. Genet. **111**(8): 1532–1544. doi:10.1007/s00122-005-0082-4. PMID:16205907.

Lai, Z., Nakazato, T., Salmaso, M., Burke, J.M., Tang, S.X., Knapp, S.J., and Rieseberg, L.H. 2005*b*. Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. Genetics, **171**(1): 291–303. doi:10.1534/genetics.105.042242. PMID:16183908.

Lai, Z., Gross, B.L., Zou, Y., Andrews, J., and Rieseberg, L.H. 2006. Microarray analysis reveals differential gene expression in hybrid sunflower species. Mol. Ecol. **15**(5): 1213–1227. doi:10.1111/j.1365-294X.2006.02775.x. PMID:16626449.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. **24**(3): 133–141. doi:10.1016/j.tig.2007.12.007. PMID:18262675.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature (Lond.), **437**(7057): 376–380. PMID:16056220.

Miller, J.F., Gulya, T.J., and Vick, B.A. 2006. Registration of three maintainer (HA 456, HA 457, and HA 412 HO) high-oleic oilseed sunflower germplasms. Crop Sci. **46**(6): 2728–2728. doi:10.2135/cropsci2006.06.0437.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., et al. 2000. A whole-genome assembly of *Drosophila*. Science (Washington, D.C.), **287**(5461): 2196–2204. doi:10.1126/science.287.5461.2196. PMID:10731133.

Natali, L., Santini, S., Giordani, T., Minelli, S., Maestrini, P., Cionini, P.G., and Cavallini, A. 2006. Distribution of Ty3–Gypsy- and Ty1–Copia-like DNA sequences in the genus *Helianthus* and other Asteraceae. Genome, **49**(1): 64–72. doi:10.1139/g05-058. PMID:16462902.

Ouyang, S., and Buell, C.R. 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. **32**(1): D360–D363. doi:10.1093/nar/gkh099. PMID:14681434.

Price, A.L., Jones, N.C., and Pevzner, P.A. 2005. De novo identification of repeat families in large genomes. Bioinformatics, **21**(Suppl. 1): i351–i358. doi:10.1093/bioinformatics/bti1018. PMID:15961478.

Rieseberg, L.H., Choi, H.C., Chan, R., and Spore, C. 1993. Genomic map of a diploid hybrid species. Heredity, **70**(3): 285–293. doi:10.1038/hdy.1993.41.

Rieseberg, L.H., Vanfossen, C., and Desrochers, A.M. 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature (Lond.), **375**(6529): 313–316. doi:10.1038/375313a0.

Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., et al. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science (Washington, D.C.), **301**(5637): 1211–1216. doi:10.1126/science.1086949. PMID:12907807.

Rostoks, N., Park, Y.-J., Ramakrishna, W., Ma, J., Druka, A., Shiloff, B.A., et al. 2002. Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. Funct. Integr. Genomics, **2**(1–2): 51–59. doi:10.1007/s10142-002-0055-5. PMID:12021850.

Rounsley, S., Marri, P.R., Yu, Y., He, R.F., Sisneros, N., Goicoechea, J.L., et al. 2009. De novo next generation sequencing of plant genomes. Rice, **2**(1): 35–43. doi:10.1007/s12284-009-9025-z.

Schatz, M.C., Delcher, A.L., and Salzberg, S.L. 2010. Assembly of large genomes using second-generation sequencing. Genome Res. **20**(9): 1165–1173. doi:10.1101/gr.101360.109. PMID:20508146.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F.S., Pasternak, S., et al. 2009. The b73 maize genome: complexity, diversity, and dynamics. Science (Washington, D.C.), **326**(5956): 1112–1115. doi:10.1126/science.1178534. PMID:19965430.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature (Lond.), **436**(7052): 793–800. doi:10.1038/nature03895. PMID:16100779.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, I. 2009. ABYSS: a parallel assembler for short read sequence data. Genome Res. **19**(6): 1117–1123. doi:10.1101/gr.089532.108. PMID:19251739.

Soderlund, C., Humphray, S., Dunham, A., and French, L. 2000. Contigs built with fingerprints, markers, and FPCv4.7. Genome Res. **10**(11): 1772–1787. doi:10.1101/gr.GR-1375R. PMID:11076862.

Stevens, P. 2010. Angiosperm phylogeny Website. Available from http://www.Mobot.Org/mobot/research/apweb/ [accessed 2 December 2010].

Tang, S., Yu, J.K., Slabaugh, M.B., Shintani, D.K., and Knapp, S.J. 2002. Simple sequence repeat map of the sunflower genome. Theor. Appl. Genet. **105**(8): 1124–1136. doi:10.1007/s00122-002-0989-y. PMID:12582890.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature (Lond.), **408**(6814): 796–815. doi:10.1038/35048692. PMID:11130711.

Timms, L., Jimenez, R., Chase, M., Lavelle, D., McHale, L., Kozik, A., et al. 2006. Analyses of synteny between *Arabidopsis thaliana* and species in the Asteraceae reveal a complex network of small syntenic segments and major chromosomal rearrangements. Genetics, **173**(4): 2227–2235. doi:10.1534/genetics.105.049205. PMID:16783026.

Ungerer, M.C., Strakosh, S.C., and Zhen, Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. **16**(20): R872–R873. doi:10.1016/j.cub.2006.09.020. PMID:17055967.

Wicker, T., Matthews, D.E., and Keller, B. 2002. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. **7**(12): 561–562. doi:10.1016/S1360-1385(02)02372-5.

Wicker, T., Yahiaoui, N., and Keller, B. 2007. Contrasting rates of evolution in *PM3* loci from three wheat species and rice. Genetics, **177**(2): 1207–1216. doi:10.1534/genetics.107.077354. PMID: 17720914.

Wieckhorst, S., Bachlava, E., Dussle, C.M., Tang, S., Gao, W., Saski, C., et al. 2010. Fine mapping of the sunflower resistance locus PL (ARG) introduced from the wild species *Helianthus argophyllus*. Theor. Appl. Genet. **121**(8): 1633–1644. doi:10.1007/s00122-010-1416-4. PMID:20700574.

Wu, C.C., Nimmakayala, P., Santos, F.A., Springman, R., Scheuring, C., Meksem, K., et al. 2004. Construction and characterization of a soybean bacterial artificial chromosome library and use of multiple complementary libraries for genome physical mapping. Theor. Appl. Genet. **109**(5): 1041–1050. doi:10.1007/s00122-004-1712-y. PMID:15164176.

Yu, J.K., Tang, S., Slabaugh, M.B., Heesacker, A., Cole, G., Herring, M., et al. 2003. Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Sci. **43**(1): 367–387. doi:10.2135/cropsci2003.0367.

Copyright of Botany is the property of Canadian Science Publishing and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.