The First Reference Genome of Sunflower (*Helianthus annuus* L.)

A Domesticated Compilospecies

by

Christopher J. Grassa

B.S. in Zoology, University of Florida, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Botany)

The University of British Columbia
(Vancouver)

April 2015

Abstract

I present the first reference genome for sunflower, *Helianthus annuus*. The reference is 3.6 billion base pairs long and is divided into seventeen lines of text representing the DNA of sunflower's seventeen chromosomes. This reference was constructed via DNA sequencing and assembly of sunflower line HA412, physical mapping using a sequence-based barcoding approach, and genetic mapping based on low coverage DNA sequencing of a highly polymorphic mapping population. I also assembled and annotated a reference genome of sunflower's mitochondrial genome. Sunflower and its wild relatives are a useful system for studying ecology and evolution. *Helianthus annuus* may be regarded as a natural compilospecies; adaptive introgressive hybridization with related species has facilitated the expansion of its range over a variety of soils and climates. In addition, the compatibility of sunflower with its extremophile wild relatives offers the opportunity to breed environmentally resilient sunflower cultivars that can cope with global climate change. The resource described in this thesis will be a useful tool for evolutionary biologists and crop breeders with interests pertaining to sunflower genetics.

Preface

The Sunflower Genome Project received $10 million in funding over a period of five years and was headed by five co-Principle Investigators (PIs) in three countries. It follows that the work was highly collaborative. Loren H. Rieseberg, Nolan C. Kane, John M. Burke, Patrick Vincourt, and Steve Knapp conceived the Sunflower Genome Project. They were responsible for its high-level design under the supervision of Scientific Advisory Board members: Scott Jackson, Brad Barbazuk, Carl Douglas, Conrad Brunk, and Catherine Feuillet.

My intellectual contributions to the project involved understanding the high-level design of the PIs and the technical details of its individual components. My practical contributions mainly involved performing bioinformatics work to process DNA sequencing data into biologically meaningful information. The work described in this thesis would have been impossible in the absence of a team of people. I performed most of the bioinformatics work for several of the project's major components and was responsible for assuring the quality of others. My most important contributions to the project, however, were: persistent involvement, developing a detailed understanding of how the components fit together, and integrating them to meet the project's high-level design.

The plant material employed for Section 2.1 was prepared by Shunxue Tang. The sequencing described in Section 2.2 was carried out by at Genome Quebec in Montreal, QC, Canada. I carried out the genotyping described in Section 2.2. I constructed the genetic map described in Section 2.3 which was hand-curated by John Edward Bowers. Bowers provided the verbal model for matching incomplete segregation patterns to a template map described in Section 2.3. I formalized the model in computer code.

The DNA sequencing libraries described in Section 3.1.1 were prepared by at: Genome B.C., Genome Quebec, the French National Institute for Agricultural Research (INRA), and the Beaty Biodiversity Research Centre's NextGen Sequencing facility. I curated the data and was responsible for its quality control. I performed all of the work described in Section 3.1.2. Nolan C. Kane and I worked together to configure the genome assembly described in Section 3.1.3. Nolan C. Kane and Thuy Nguyen monitored the assemblys computation. After three months of processing, the assembly failed in its final stage of converting binary files to FASTA-formatted text, but Thuy Nguyen wrote custom computer code to recover from the failure. Sariel Hubner carried out some of the bioinformatics work described in Section 3.1.4. Navdeep Gill assigned Allpaths and Celera scaffolds to linkage groups using the physical map. I assigned the Allpaths and Celera scaffolds, as well as mate-pair reads, to linkage groups using information from genetic maps. The plant material and sequencing libraries employed in Section 3.2 were prepared by Dan Ebert. I performed most of bioinformatics work described in this section, including configuring the SOAP assembly and writing the custom computer code used to align the assembly contigs to the restriction map. Nolan C. Kane filled gaps in the mitochondrial genome assembly using 454 reads. I hand-curated and annotated the mitochondrial genome assembly.

The physical map described in Section 4.2 was constructed by KeyGene, Inc. and hand-curated by Navdeep Gill. Thuy Nguyen wrote custom computer code to assign BACs to linkage groups and break chimeric contigs. Navdeep Gill then assembled physical maps for each linkage group independently. I designed the algorithm for aligning the genome assembly scaffolds to physical map contigs within the constraints of the genetic map, which was implemented in custom computer code written by Frances Raftis and me. I designed and implemented the algorithm described in Section 4.3. I was responsible for the quality control described in Section 4.4. Jerome Gouzy and Sebastien Carrere at INRA performed the genome annotation described in Section 4.4.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Loren H. Rieseberg gave me the opportunity to be a part of a whole that exceeds the sum of its parts.

Nolan C. Kane believed that I could grow as a scientist.

Rose L. Andrew demonstrated to me that the true reward of doing good science is truth.

Rob J. Kulathinal taught me the value of being a protégé.

Connor Morgan-Lang and Nadia Chadir taught me the value of being a mentor.

Michael C. Whitlock is responsible for much of my understanding of population genetics. He also provided the most comprehensive review of this thesis.

John E. Bowers taught me most of what I know about meiotic mapping.

Matt King, John M. Burke, Navdeep Gill, and S. Evan Staton taught me much of what I know about sunflower DNA.

Armando Geraldes and Sebastien Renaut taught me that a scientist is responsible for disseminating knowledge in addition to creating it.

Quentin C.B. Cronk taught me some concepts of plant genomics. More importantly, he encouraged my curiosity.

Austin Davis-Richardson demonstrated the power that lies in UNIX mastery to me.

Thuy Nguyen, Daisie Huang, and Frances Raftis demonstrated the value of elegant algorithm design to me.

Gregory J. Baute was always available to listen to my ideas.

Kathryn G. Turner, Brook T. Moyers, Gregory L. Owens, Kate Ostevick, and Kieren Samuk taught me much of what I know about ecology and evolution.

Genome Canada, Genome B.C., and the Sunflower Genome Consortium funded my work.

Joshua Chang Mell taught me to question my assumptions.

Jayne Elizabeth Knight brought me to Vancouver, B.C., Canada.

Rogers Thompson Brewer, Marjesca Brown, Rebecca Deist, Jonathan S. Griffiths, Nikta Fay, and Katie Elizabeth Berns encouraged me to keep trying when I felt like giving up.

Carl J. Douglas and Shawn D. Mansfield influenced my cerebral model of nature at the cellular scale.

Dylan Orion Burge influenced my cerebral model of nature at the geologic scale.

My parents, Dreama Andersen, Bill Grassa, and Bruce Andersen, and grandparents, Sarah and Thomas Grassa, provided me with DNA and an environment for early development.

Cristina Maria Moya reminded me of the reasons I first fell in love with science and renewed my enthusiasm for a scholarly life. She also provided helpful comments for this thesis.

Chapter 1

Introduction

This thesis describes my part in producing a reference genome for sunflower. A genome is the entire DNA (deoxyribonucleic acid) belonging to a single organism. It contains the information needed to grow from a zygote to mature adult. A reference genome is a textual model of this information. Bare DNA is useless in the absence of the cellular machinery needed to transcribe and translate the information it encodes into proteins. Similarly, a reference genome is meaningless outside the context in which it will be used; it is a resource. Sunflowers are an important oilseed crop and also an important model for studying ecology and evolution. As such, I begin with introductions of the sunflower system and the domestication of the common sunflower before introducing the methods and resources I used to craft a resource for future research.

Darwin's sketches of the phylogenetic relationship of species resemble the branching of a tree (Darwin 1859). As time progresses in the sketches, biodiversity increases via the division and differentiation of populations, eventually leading to speciation. Edgar Anderson suspected that the topology of phylogenetic relationships could be more complex than this and closed his manuscript "Internal Factors Affecting Discontinuity between Species" with:

> *I have taken asexual propagation, polyploidy series and physiological isolation as representatives of the internal factors which affect specific isolation and which whole genera or even families of plants may have in common. There must be many other such factors. May we not there-*

> *fore logically expect that, even though species prove to be biological units, their relationships with each other and the relationships of individuals within species will vary from genus to genus and from family to family?* (Anderson 1931)

This commences his explorations of reticulate evolution (e.g. Anderson 1936, Anderson and Hubricht 1938). Reticulate evolution refers to a phylogenetic topology in which branches not only bifurcate, but also interweave and rejoin to form new branches. He first researched allopolyploid speciation, but would later develop the concept of (and write a book titled) *Introgressive Hybridization* (Anderson 1949). Introgressive hybridization is recombinant reticulate evolution, whereby some portion of the genome of one species is introduced to another's via meiotic recombination of the two in a first generation hybrid, followed by backcrossing in later generations.

Around the time that he published these ideas, his student, Charlie Heiser, took an interest in hybridizing sunflowers (Heiser Jr 1947) and would go on to conduct a comprehensive inventory and key of *Helianthus* (Heiser et al. 1969), the sunflower genus. Morphometry and cytogenetics of the clade suggested widespread and ongoing hybridization resulting in polyploid speciation (Heiser and Smith 1954) and introgression (Heiser et al. 1962). In hindsight, it is clear that this collaboration gave birth to sunflower as a system in which to study reticulate evolution in the context of a variety of geographies (Renaut et al. 2013).

The genus includes approximate fifty species endemic to North America. The area covered by their combined ranges includes most of the geography bounded by the Unites States, the prairies of southern Canada, and northern and central Mexico, including Baja. Heiser split the genus into three sections: Annui (the annuals), Ciliares (western perennials), and Divaricati (eastern perennials). While several allopolyploid origins of perennial species have been documented, the majority of sunflower evolution and ecology research focuses on section Annui, comprised of approximately fourteen diploid species. The most ancestral node in the section dates to approximately two million years (Sambatti et al. 2012), splitting the Annuus group from the Petiolaris group. In addition to its namesake, *H. annuus*, the Annuus group includes: *H. argophyllus*, the silverleaf sunflower, endemic to Texas,

H. bolanderi, a California sunflower colonizing serpentine soils, and *H. winterii* (Stebbins et al. 2013), a derived tree that has reverted to a perennial life history and grows dense woody stems. *H. petiolaris*, the prairie sunflower, is broadly sympatric with *H. annuus*. The Petiolaris group also includes *H. debilis*, split into several subspecies clustered in the southeast U.S., *H. neglectus*, and *H. niveus*, hypothesized to be the ancestral type (Beckstrom-Sternberg et al. 1991) of the annuals and divided into polyphyletic subspecies (Rieseberg et al. 1991). In general, some interspecific gene flow may be expected wherever annual sunflowers are sympatric (Yatabe et al. 2007, Kane et al. 2009, Scascitelli et al. 2010).

Heliantus annuus and *H. petiolaris* are the most broadly sympatric annual sunflower species. Mosaic hybrid zones (made up of first-generation (F1) crosses and various backcrossed generations) often form when these two species are in very close proximity (Rieseberg et al. 1998), and so it is perhaps unsurprising that they have some of the highest rates of gene flow documented in the clade. *Helianthus annuus* and *H. petiolaris* are also notable as the progenitor species of three homoploid hybrid species: *H. anomalus*, *H. deserticola*, and *H. paradoxus* (Rieseberg 1991). Homoploid hybrid speciation is hybrid speciation without a change in chromosome number; the derived genomes, which stabilize after about 1,000 generations, are mosaic chimeras of the ancestral genomes (Buerkle and Rieseberg 2008). The ancestry of chromosomal tiles making up the mosaics matches the parental direction of quantitative traits segregating in synthetic interspecific crosses (Rieseberg et al. 2003), suggesting the possibility that they harbor multi-gene complexes coadapted to produce phenotypes matched to some small sub-niche of ecological space. The homoploid hybrid sunflowers inhabit extreme environments, for example, the salt marshes where *H. paradoxus* grows (Karrenberg et al. 2006). The transgressive phenotypes needed to survive in these extreme environments may be caused by positive epistatic interactions between the tiles or additive gene action.

Such hybridization is not only a historical process, but also occurs frequently in many places where interfertile species co-occur (e.g. Figure 1.1). In most cases, however, hybridizaiton is rare and leads to only low levels of gene flow among species (Kane et al. 2009). Still, because of their extremely large effective population sizes, the widespread sunflower species such as *H. annuus* harbor genetic variation derived from introgression from even quite distant lineages.

Given all this hybridization and gene flow, a critical reader might ask if the sunflower species named above qualify as separate species at all. This is a legitimate question, but it is important to remember that speciation is, more often than not, a gradual process that occurs through time (Schluter 2009, Feder et al. 2012) (N.B. a hybrid fern, albeit reproducing asexually, has recently been found to be the product of an intergeneric cross of lineages separated by 60 million years (Rothfels et al. 2015)). That sunflower species lie within various ranges of the speciation continuum is what makes them so useful for studying the process. Examples of the early stages of speciation may be found in sunflowers (e.g. dune populations of *H. petiolaris* diverging from those living on the nearby sandsheet (Andrew et al. 2013)), but interfertility between the named species is quite low; the fertility of interspecific crosses is usually less than 5% (Chandler et al. 1986). Much of this reproductive barrier may be attributed to chromosomal rearrangements (Burke et al. 2004). Sunflowers have some of the highest known rates of chromosomal evolution, a factor likely contributing to their rapid diversification (Barb et al. 2014).

The sunflower may be regarded as a natural compilospecies (Harlan and De Wet 1963); adaptive introgressive hybridization with related species has facilitated the expansion of its range over a variety of soils and climates.

Not only is sunflower an important model system for studying evolution and ecology, but it is also an important crop. The common sunflower (*Helianthus annuus macrocarpus*) was domesticated approximately 5,000 years ago in the area of what is now Tennessee (Blackman et al. 2011). Native Americans selected for increased head size and lack of shattering in their crop, and the farming practice spread via social transmission (Harter et al. 2004). They mostly used sunflowers as food, but the Hopi also developed a second line high in anthocyanin that is still used to produce dye (Heiser 1951).

Sunflowers were introduced to Europe in the 16th century by Spanish explorers returning from the new world. They first became popular there as a horticultural novelty that was easy to care for and grew larger than a child in a single year. By the 17th century, they had reached Russia. There, sunflower became popular in part because the Russian Orthodox Church did not include it in the list of fats that could not be eaten during Lent. Russian breeders selected for larger seed size and higher oil content, establishing it as an oilseed crop. Germplasm resulting from

their efforts returned to North America in the late 1800s and became the primary stock from which most modern elite lines are derived (Blamey et al. 1997).

Pedigrees kept by breeders indicate that synthetic introgression of alleles from wild relatives has been used to improve the germplasm of elite lines several times (Rieseberg and Seiler 1990). Modern genome scans confirm this, for example: reintroduction of the branching allele from *H. annuus ssp. texanus*, downy mildew resistance from *H. argophyllus*, and cytoplasm from *H. petiolaris* (Baute et al. 2015, Dussle et al. 2004, Horn et al. 1991). Investigation of the ecological niches (as modeled by bioclimatic variables) that sunflowers occupy suggest that current elite germplasm may be grown in conditions covering less than half the variance that their wild relatives occupy (M. Kantar, per. comm.). The vast genetic diversity present in wild relatives of cultivated sunflower (Seiler 1992, Mandel et al. 2011, Hodgins et al. 2014) will continue to be an important resource in crop breeding, with ongoing efforts to breed lines resistant to drought, flood, salt, and parasites (Rauf 2008, Wan et al. 2013, Ahmed et al. 2013, Seiler and Jan 2014). These projects are helping to ensure that humans nutritional requirements will be met in the face of global climate change (McCouch et al. 2013, Dempewolf et al. 2014).

The domestic sunflower's nuclear genome is estimated to contain approximately 3.6 billion base pairs (bp) (Baack et al. 2005) with a guanosine + cytosine (G+C) content of 40%. Karyotype analyses report seventeen chromosome pairs. Generally, thirteen of these are categorized as meta- or submeta-centric and four as acrocentric with a total of three nucleolus-organizing regions (NORs) (Feng et al. 2013). The genome is highly redundant: approximately 80% of the DNA is retrotransposon sequence. Most of this derives from recent proliferation of the *Ty-3/Gypsy* type (Staton et al. 2012). Approximately half of the genome consists of a single *Ty-3/Gypsy* element less than 10kbp in length with an average pairwise divergence of 1% between copies. Additional redundancy in the gene space has been attributed to a number of paleopolyploidy events (Barker et al. 2008). Line HA412HO (Miller et al. 2006) was chosen for sequencing because it is highly inbred.

Cooperation and communication between evolutionary biologists and plant breeders expedites the practical application of pure science. The goal of my work, described below, is to facilitate knowledge synthesis by providing a common axis

for all sunflower researchers. To do so, I produced an ultra-high density genetic map, assembled a genome *de novo* from short reads, and integrated these with a physical map of the genome. All three information sources were necessary to complete this reference genome. The product of DNA sequencing is millions or billions of very short reads. I assembled these into hundreds of thousands of contiguous sequences. I scaffolded these into tens of thousands of sequences using a physical map and anchored them to chromosomes with a genetic map. This furthers the work of several collaborators (Kane et al. 2011). Here I mainly describe my contributions to generating a reference sequence for sunflower, but I also include brief summaries of work by others as needed for context.

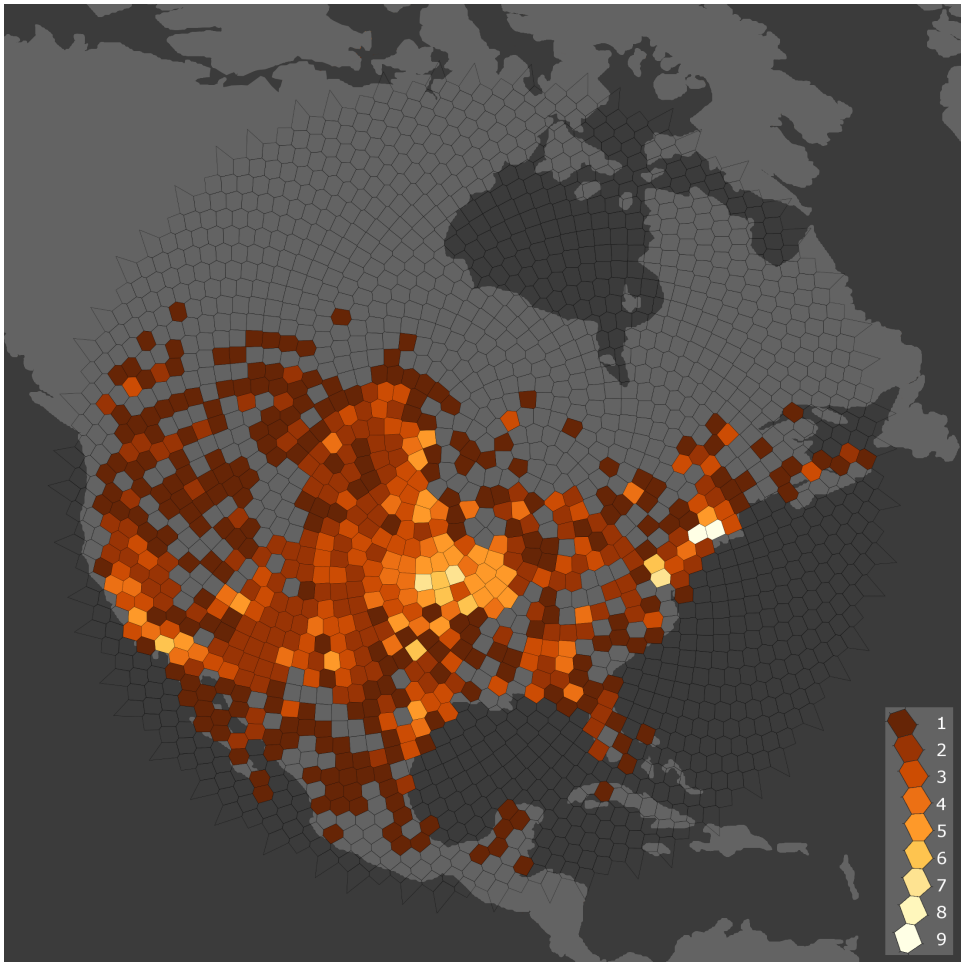

Figure 1.1: *H. annuus* is interfertile with nine other species of sunflower. The fill color of each polygon indicates the number of species interfertile with *H. annuus* found in the area. North American occurrence records for the nine species were downloaded from The Global Biodiversity Information Facility (GBIF) (Lane 2003). Polygons are defined by a tessellation (Dirichlet 1850) around points generated from a model of sunflower seed packing (Vogel 1979).

Chapter 2

Ultra-high Density Genetic Map

2.1 Plant Material and Construction of Mapping Population

The sunflower reference mapping population is derived from a cross between *Helianthus annuus* cultivars RHA280 and RHA801. RHA280 was first registered in 1974 (Fick et al. 1974) and is derived from the open-pollenated Sundak germplasm. A typical confectionary line, it produces large black seeds with white stripes containing relatively low oil concentration. It is also a fertility restorer of male-sterile cytoplasm, midseason maturing, and rust-resistant. RHA801 was first registered in 1981 (Roath et al. 1981) and is derived from a population of lines RHA271, RHA273, RHA274, R344, R494 after selection for improved yield and three generations of selfing. RHA801 is a dominant fertility restorer and has moderate rust resistance. It is also resistant to *Verticillium* wilt and downy mildew. RHA801 is a high-oil cultivar with a single apical inflorescence.

Coancestry analysis based on pedigree indicates that the confectionary restorer lines and oilseed restorer lines to be highly inbred within each group with strong separation between groups (Cheres and Knapp 1998). RHA280 and RHA801 adhere to this pattern. Principle component analysis (PCA) of simple sequence repeat (SSR) markers revealed RHA280 as very dissimilar to other elite lines, and especially distant from RHA801 (Yu et al. 2002). The cross is thus ideal for generating a highly polymorphic mapping population.

The mapping population began with hand emasculation of RHA280, followed by pollination with RHA801 to produce an F1 (Tang et al. 2002). F1 seeds were then grown to begin the generation of the recombinant inbred lines (RILs). Each RIL lineage is of single seed decent from this F1 (Figure 2.2). Self-pollination of the RILs was carried out for seven generations in summer and winter nurseries and/or greenhouses located in Corvallis, Oregon and Balcarce, Argentina between 1995 and 1998. The RIL population segregates for apical branching as well as several seed traits, including: hull pigment, seed oil concentration, overall seed weight, and seed dimensions (Tang et al. 2006).

2.2 Sequencing and Genotyping

Whole genome shotgun sequencing was carried out with 100 base pair paired-end Illumina reads at Genome Quebec in Montreal, Canada. One lane of Illumina sequence was generated for each parent. 172,086,364 read pairs were generated for RHA280, for a total of 34,417,272,800 sequenced bases. 160,718,566 read pairs were generated for RHA801, for a total of 32,143,713,200 sequenced bases. We sequenced a total of 96 RILs to low depth. Eight lanes were each multiplexed with twelve barcoded RILs. As coverage was a little lower than we expected for some samples, an additional lane was sequenced. The ninth lane of RIL sequencing included the samples with the lowest count for each barcode tag, except for Index_8. In all, the number of bases obtained for the RILs ranged from 1.859,549,200 to 6,971,326,000 with a mean of 3,692,104,758 and standard deviation of 881,568,716. Assuming a genome size of 3.6 Gbp, RHA280 and RHA801 were sequenced to a depth of coverage of approximately 9.6x and 8.9x, respectively, with the average depth of coverage obtained for the RILs approximately 1.0x (Figure 2.3).

I aligned parental reads to our draft reference assembly using the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) and called genotypes using SAMtools mpileup (Li et al. 2009). I used fixed Single Nucleotide Polymorphisms (SNPs) with a genotype quality more than 20 and a mapping quality more than 30 on the Phred scale (Ewing and Green 1998) as candidate sites for calling genotype blocks in the RILs. The RIL reads were aligned to our draft reference assembly using

BWA. I used SAMtools to convert the alignments to the pileup format. As the RILs were sequenced to very low coverage, I did not apply the strict quality cut-offs used for the parental reads to them. Instead, I called each candidate site as inherited from either or both parents based on the presence of fewer than three aligned reads and used a quality-control heuristic later in the process (Section 2.3). In all, I identified 2,726,257 SNPs on 273,422 contigs.

2.3 Construction of Genetic Map

In each individual, I then called genomic contigs as descended from one or the other parent based on the presence of at least nine genotype calls at candidate sites. As no quality filters were applied at the read level, I also required at least 90% of the genotype calls to indicate descent from the same parent. I used this cut-off to allow contigs containing small repetitive regions (potentially attracting reads from distant loci), distal recombination breakpoints (allele switching at the contig ends), or small regions of gene conversion (allele switching internal to the contig) to be mapped. I used contigs meeting these requirements in at least 75% of the RILs and with a minor allele frequency greater than 30% as map markers. I used MSTmap (Wu et al. 2008) to order the markers in linearly. MSTmap groups markers based on the minimum sum of recombination events (Hamming 1950) between their segregation patterns and divides them into linkage groups if the sum is significantly different than observed across all markers. MSTmap then orders markers on each linkage group using a recursive minimum spanning tree algorithm. I calculated the map distance between adjacent pairs of markers that were ordered by MSTmap with Kosambi's mapping function (Kosambi 1943) (Figure 2.1) (N.B. John E. Bowers later pointed out to me that a mapping function is not needed for a saturated genetic map).

A template map of 1629 mapped bins was curated from the map generated by MSTmap. The initial template map was manually curated by collaborator John E. Bowers based on results of testing all SNPs and contigs to fill in gaps that may have been missed with the initial 4200 contigs. Each bin represented all loci that showed an identical segregation pattern for 93 RILs. Plants representing three RILs (RIL10, RIL46, and RIL255) appeared to be highly heterozygous and showed an

excessive number of apparent recombinations. These plants were assumed to represent outcrosses and were not used in the map. The apparent number of recombinations on the three excluded lines ranged from four to ten times the number seen on the other 93 lines. I suspect a bee or some other insect may have contaminated these lines with non-self pollen. The template map contains 2531 recombination events and is 1361 centimorgans long.

A primary goal of constructing the genetic map was to anchor *de novo* assembled contigs to chromosomes. I compared all contigs containing segregating SNPs to the template map. Comparisons were made in forward and reverse order and the best match was stored for each direction. A contig was placed with an upper distance of the best forward match and a lower distance of the best reverse match if both were found on the same linkage group. This allowed me to anchor contigs to chromosomes even if they did not contain complete segregation patterns or if they contained some level of error in genotyping. A total of 243,048 contigs were placed to an accuracy of 5 centiMorgans (cM) (Figure 2.1).

2.4 Consensus with Other Genetic Maps

This ultra-high density genetic map was just the most recent of several constructed using the core mapping population of sunflower. The prior map of highest density (Bowers et al. 2012) used RHA280 x RHA801 markers genotyped with a 10,640 SNP Infinium array that we developed in collaboration with Advanta Seeds, Dow Agrosciences, Syngenta AG, and Pioneer Hi-Bred. The array's probe sequences were matched by BLAST (Altschul et al. 1997) to the contigs in the sunflower assembly. The cM positions on the two maps were compared. The two maps agreed very well in terms of synteny and ordering even though they were completely independently constructed (Figure 2.4). The chromosomes from the sequence-based map were then named and oriented relative to the previous literature.

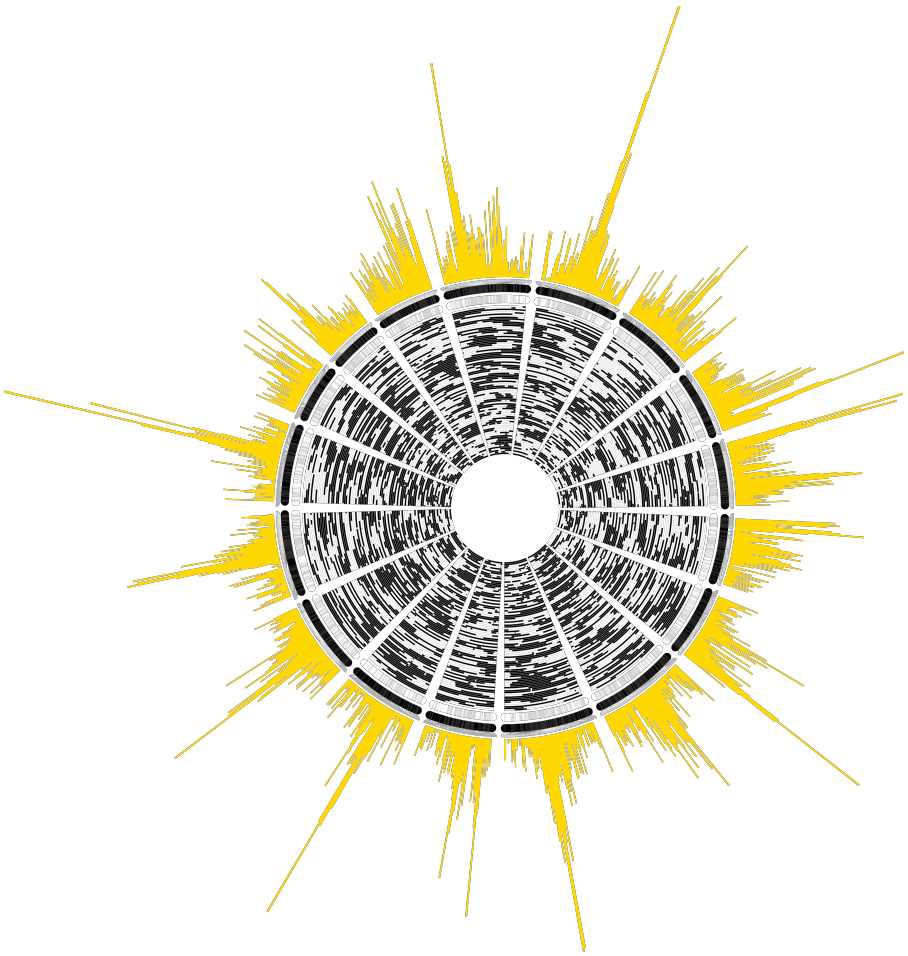

Figure 2.1: Genetic map for *H. annuus*. Interior radii show the segregation of chromosome segments in ninety-three RILs. Black segments indicate RHA280 ancestry and white segments RHA801 ancestry, with transitions locations of chromosomal crossover. The genetic map is drawn along the outer two sets of radii. The ray length (yellow) is proportional to the sum of *de novo* base pairs assigned to 1 cM bins.

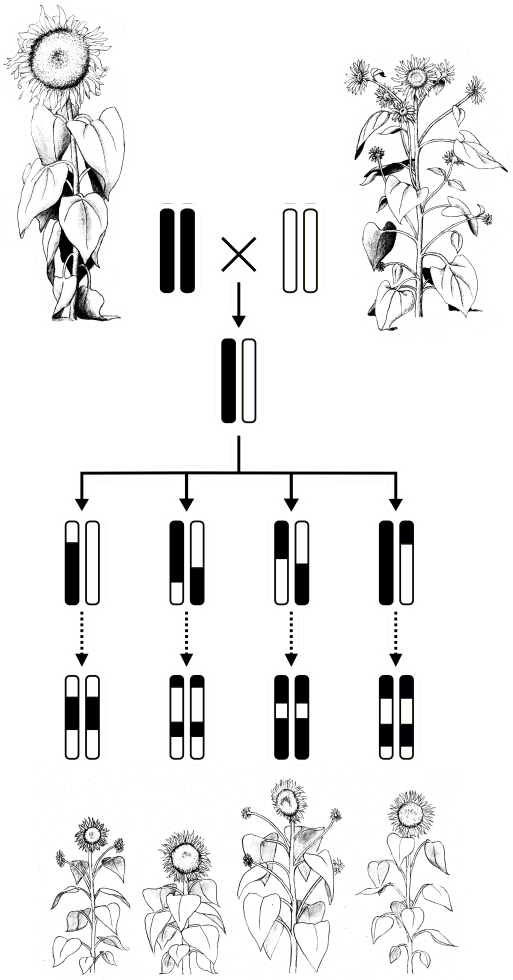

Figure 2.2: Illustration of RIL crossing design employed for making the RHA280 x RHA801 genetic map for *H. annuus*. (Courtesy of Kasia Stepien)

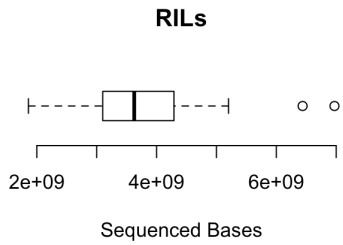

Figure 2.3: Box plot showing distribution of sequencing depth for 93 RILs. The sunflower's genome size is estimated to be 3.6 Gbp. The RILs were sequenced to approximately 1x depth.

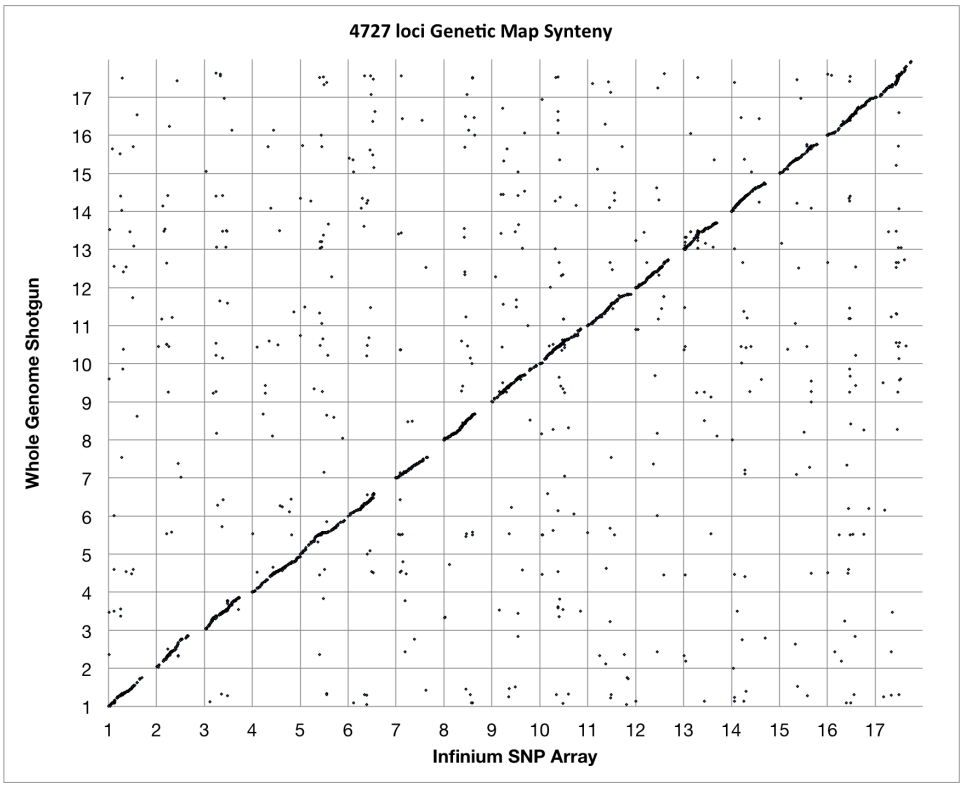

Figure 2.4: Comparison of synteny between Illumina Infinium SNP array and Whole Genome Shotgun sequence-based map of RHA280 x RHA801 RILs. Approximately 90% of hits are in 17 syntenic blocks. The roughly 10% of non-syntenic hits can be explained by picking the second best hit if the true homolog is fragmented into several contigs, or if the sequence is multi-copy.

Chapter 3

Genome Assembly

3.1 Nuclear Genome

3.1.1 Preparation and Sequencing of DNA Libraries

Two sequencing technologies were dominant during the life of the project: Roche 454 (Margulies et al. 2005) and Illumina (Bentley 2006). Both of these produce reads of DNA fragments about 500bp in length. Mate-pair libraries (Van Nieuwerburgh et al. 2011) were prepared in order to achieve paired reads separated by up to about 20Kbp. I will briefly describe the properties of each technology and method of library preparation, as the unique properties affect the choice of appropriate algorithms used to assemble them.

454 reads are generated as follows (Rothberg and Leamon 2008). Organismic DNA is fractionated. Oligonucleotide adapters are attached to denatured fragments. The fragments are diluted in an emulsion along with beads that the adapters bind to. The emulsion is prepared such that one fragment of organismic DNA and one bead lie within a drop of oil. A polymerase chain reaction (PCR) occurs within the drop of oil so that many single-stranded copies of the original DNA molecule exist within it, hybridized to the beads via the adapter sequence. The beads are then drawn into wells etched in a fiber optic plate.

A solution of nucleotides, polymerase, sulfurylase, and luciferase are added to the wells. Polymerization of the complementary strand results in the addition of

a nucleoside to the strand and pyrophosphate. The pyrophosphate is converted to adenosine triphosphate (ATP) by the sulfurylase. ATP and luciferin are converted to oxyluciferin by the luciferase, emitting a photon. A digital camera captures the photon emissions from the wells. This process is repeated for each nucleotide, with washes in between, to make one cycle. The template sequence contained in each well may be inferred by determining which nucleotide addition in the cycle caused photon emissions from the well to be captured by the camera.

Illumina reads are generated via methods fundamentally similar to 454 sequencing (Metzker 2010). Two sequence primer templates are ligated to DNA fragments; each end receives a different primer template sequence with an adapter, either complementary or the same as that of the flow cell, at its extremity. The molecules are placed directly on a flow cell and polymerized via bridge amplification. Oligonucleotide primers complementary to one of the templates are added to the flow cell, initiating polymerization. The different nucleotides are added to the flow cell in solution together. They are engineered such that a specific fluorescent label is bonded to the base. Additionally, a trinitrogen monoxide (rather than alcohol) is bonded to carbon 3 of the pentose, preventing polymerization. A digital camera records the fluorescence as the flow cell is excited with a laser. The labels are cleaved and the trinitrogen monoxide is replaced with an alcohol, completing one cycle. This is repeated for one hundred cycles. The second primer oligonucleotides are then added to the flowcell, and the process is repeated.

While 454 and Illumina technologies are fundamentally similar, there two differences with significant consequences. One is the trinitrogen monoxide on the pentose that is later replaced by an alcohol (termed reversible terminator) (Bentley et al. 2008). While nucleotides are added individually in the 454 process, it is still possible for more than one to be added if the present region of the template is a homopolymer. The intensity of the luminescence is used to estimate the homopolymer length, but the estimation is not precise enough to determine the exact length of long homopolymers. The other major difference of the Illumina process is that both strands of DNA are sequenced, each from the opposite end's primer. This gives paired-end reads.

Reconstructing the contiguous sequence of a genome from reads of this size is impossible if the genome contains repeated sequence longer than the longest read.

Mate-pair libraries help overcome this limitation. Mate-pair library preparation begins by size-selecting DNA fragments ranging in length from about 2,000bp to about 20,000bp. The fragments are then circularized (via biotinylation for Illumina libraries, or, for 454 libraries, a 42-44bp linker sequence). The circular molecules are then broken on either side of the join to give a small fragment of DNA containing sequence from the extremities of the original long fragment. These fragments are then sequenced via the aforementioned methods.

Fragmentation of the circular molecules does not always occur on either side of link. 454 mate pairs are easily identified as the linker sequence is wholly contained within the read, flanked by each mate. Illumina mate pair libraries include reads representing proper mate pairs (if the link was located at a distance from either end that is greater than the read length), typical short-fragment paired-end reads (if the fragmentation did not include the link), and chimeric reads (if the link was located at a distance from either end that is less than the read length). As Illumina mate pair library preparation does not include a linker sequence, proper mate pairs must be isolated via the removal of improper pairs. This can be accomplished by aligning them to a draft assembly from which they were excluded.

We prepared five short-fragment libraries and seventeen mate-pair libraries, totalling approximately 60x depth of coverage, for sequencing on the Illumina platform. Seven short-fragment libraries and fourteen mate-pair libraries, totaling approximately 24x depth of coverage, were sequenced using 454 technology. Library statistics are summarized in Table 3.1 and Table 3.2.

3.1.2 Allpaths-LG Assembly

The de Bruin graph has emerged as the most popular method for assembling high-volume, short read sequencing data with a low error rate (Illumina reads) (Zerbino and Birney 2008). Most assemblers based on the de Bruin graph divide the assembly process into three general steps: error correction of the reads, contig construction using the graph, and scaffolding.

Error correction (Kelley et al. 2010) involves first tiling the reads into short (e.g. 25bp) words of length k, or k-mers, while keeping track of how often they appear (multiplicity). Whole genome shotgun data involves the random shearing

and sequencing of organismic DNA. The frequency distribution of sampling multiplicities of k-mers that are unique in the genome is expected be Gaussian (Galton 1894) and centered on the mean depth of coverage. In practice, repetitive or polymorphic k-mers affect the distribution, but this is not relevant for correcting errors. Errors in high quality reads are rare, introducing a frequency spike of low multiplicities in the distribution. In other words, we may expect to find many errors in billions of reads, but it is unlikely to find the same error many times. The frequency minimum between this spike and the Gaussian peak suggests a multiplicity cut off between suspicious and trusted k-mers. A second tiling pass is made over the reads. Erroneous bases are identified as present in untrusted k-mers, flanked by trusted k-mers. If trusted k-mers with a low edit distance (1) from the untrusted k-mers can be found, the error can be changed. Otherwise, the read may be discarded.

Contig construction (Compeau et al. 2011) involves tiling the corrected reads into k-mers. In this phase of the assembly process, it is important for the k-mers to be unique in the genome, and so a larger value of k is used. A directed graph is built as the reads are tiled. Each k-mer is an edge in the graph connecting the k-1 word at the start of the tile to the k-1 word ending the tile. The full graph is explored in parallel. Each exploration is constrained such that it is only allowed to visit an edge once. Under this constraint, if every node has the same number edges entering it there are leaving it (i.e. it is balanced), an exploration will end at the same node that it began. The explorations are combined to form a path that visits each node once. The first nucleotide of each edge is added to a growing sequence as this path is traversed, giving the genome sequence. In practice, genomes may contain true repeats that are much longer than values of k suitable for use with short reads. Consequently, some nodes may be unbalanced. The number of edges entering the first node of a long repeat will be equal to the number of biological copies, but the number of exiting edges may be just one. The resulting final path is thus no longer linear, and must therefore be broken to give several contiguous sequences, or contigs, rather than one. Mate pairs are then used to rejoin the contigs where it is possible to do so unambiguously.

Access to a high performance computer and the sequencing of new sunflower DNA libraries recently provided me with the opportunity to use the Broad's AllPaths-LG genome assembler (Gnerre et al. 2011). AllPaths-LG estimated the sunflower

genome size to be 3.151 Gb with a GC content of 39.5% and 77% present as repetitive sequence. The final assembly size was 1.154 Gb in 99,439 scaffolds. The assembler incorporates strict sequence quality control including: cleaning reads of sequencing artifacts, trimming of low-quality sequence, base quality score normalization, and removal of low frequency kmers. The fraction of reads used from each library ranged from 13.1% to 39.9%. This filtering brought the genome sequence coverage to 40.1x in fragment libraries and 10.2x in jumping libraries. The suggested coverage is 45x for both required library classes.

Increases in library coverage, insert size, and diversity are expected to improve the performance of AllPaths-LG. The library insert sizes obtained for the sunflower differed from the sequencing model proposed by the assembler's authors. The inserts for the fragment libraries are recommended to be about 1.8 times the read length; those obtained ranged from 1.22 to 1.47 times the read length. The recommended long jumping library insert size is 6,000 bp; the longest insert size obtained had a mean insert size of 4,447 bp.

The French National Institute for Agricultural Research (INRA) provided use of their GenoBigMem server to compute the assembly. The server has approximately 1 TB of available physical RAM and 32 CPUs. AllPaths-LG completed the assembly in 206.47 hours, with a peak memory usage of 913.51 GB and an effective parallelization factor of 15.36. The assembler estimated the memory required for each module. If the required memory exceeded the available memory, the module was divided into a number of passes. This suggests that the assembler could be used on a computer with lesser resources.

3.1.3 Celera Assembly

Although not part of my thesis, for the Roche 454 reads were assembled by former postdoc Nolan Kane using the Celera Genome Assembler (CABOG) (Miller et al. 2008). It is an overlap-layout-consensus assembler. The best assembly had an N50 of 25kb, and a total assembly length of 3.1 Gb.

3.1.4 Merge of Allpaths-LG and Celera Assemblies

Because many of the Allpaths scaffolds were not found in the Celera assembly (and vice versa), postdoc Sariel Hubner employed the computer program Minimus2 (Sommer et al. 2007) was used to merge the two assemblies. To reduce the complexity of the merger (and to minimize false merges), scaffolds assigned to each linkage group were merged independently. Next, he employed the computer program SSPACE (Boetzer et al. 2011) to increase scaffold lengths with new long mate pair Illumina libraries (20 and 40 kb) and bacterial artificial chromosome (BAC) -end sequences. Again, this was done for each linkage group independently to reduce the likelihood of generating chimeric scaffolds. The scaffolding resulted in a total of 155,000 scaffolds, which were used to generate pseudomolecules (Chapter 4).

3.2 Mitochondrial Genome

Leaf tissue from ten-day-old HA412 seedlings was enriched for mitochondria by centrifugation. DNA was extracted from the enriched tissue, barcoded, and sequenced on 1/48th of an Illumina lane, producing 2,727,097 pairs of 101 bp reads. Reads were quality trimmed and cleaned of sequencing artifacts using Trimmomatic (Bolger et al. 2014).

Reads with exact matches of at least 50bp to the chloroplast genome and their mates were removed from the dataset. SOAPdenovo (Luo et al. 2012) was used to assemble the reads, producing an assembly 387,493 bp in length with an N50 of 562 bp and an N90 of 11,390 bp. Next, reads from the mate pair libraries prepared for the AllPaths-LG assembly with exact matches of at least 50 bp to this assembly without matches to the chloroplast genome were added to the scaffolding steps of a second SOAPdenovo assembly. This produced an assembly 466,799 bp in length with an N50 of 500 bp and an N90 of 46,247 bp. Some scaffolds in both assemblies are of nuclear origin and were identified based on coverage.

Scaffolds from the second assembly were digested *in silico* using the recognition sites of these restriction enzymes: PstI, SalI, KpnI, BglI, BstEII, and SacI. Each scaffold's restriction enzyme cut site sequence was aligned to the sequence of cut sites of a previously published fragment map (Figures 1 and 6 of Siculella and

Palmer 1988) using the Smith-Waterman algorithm (Smith and Waterman 1981). Alignments were confirmed by comparing the order and size of digested fragments in the region. Agreement of fragment sizes between the chemical and computational digests was very good: usually within 100 bp (Figure 3.1).

Gaps within and between scaffolds were filled using long (typically 500-1000 bp) 454 reads. Reads with exact matches to sequence on both sides of a gap were found using the UNIX grep (Kernighan and Mashey 1979) command and aligned to the super scaffold by hand. The same procedure was used to close the 300,945 bp master circle. Raw reads were aligned to the reference and a few substitution and small indel errors were fixed by hand.

I annotated the mitochondrial genome by hand and using the software Mitofy (Alverson et al. 2010). I searched for open reading frames using the National Center for Biotechnology Information's (NCBI) online BLAST aligner to identify the genes based on homology and used the software Mitofy to identify transfer and ribosomal ribonucleic acid (RNA) sequences. I identified repetitive regions by aligning the finished reference to itself using BLAST. To further verify that the assembly was correct, I aligned an independent Whole Genome Shotgun (WGS) to it using BWA and inspected the alignments by eye. As expected, I found coverage spikes at regions with high homology to the plastid genome. I found drops in coverage (although never to a depth of less than thirty reads) near the boundaries of the large repeat copies. This supports the hypothesis that the sunflower's mitochondrial genome is typically arranged as two equimolar subcircle chromosomes, each containing one of the two repeat copies found in the master replication circle (Siculella and Palmer 1988). I also aligned a sequenced RNA library to the reference. The alignment coordinates with the highest depth of coverage overlapped with gene coordinates. Some unannotated intergenic regions also attracted alignments at low coverage, which suggests the possibility that they may have some functional role. The mitochondrial genome was submitted to GenBank and is publicly available as NCBI Reference Sequence: NC_023337. Protein-coding features are summarized in Table 3.3; transfer RNA (tRNA), ribosomal RNA (rRNA), and structural features are summarized in Table 3.4; all features are plotted in Figure 3.2.

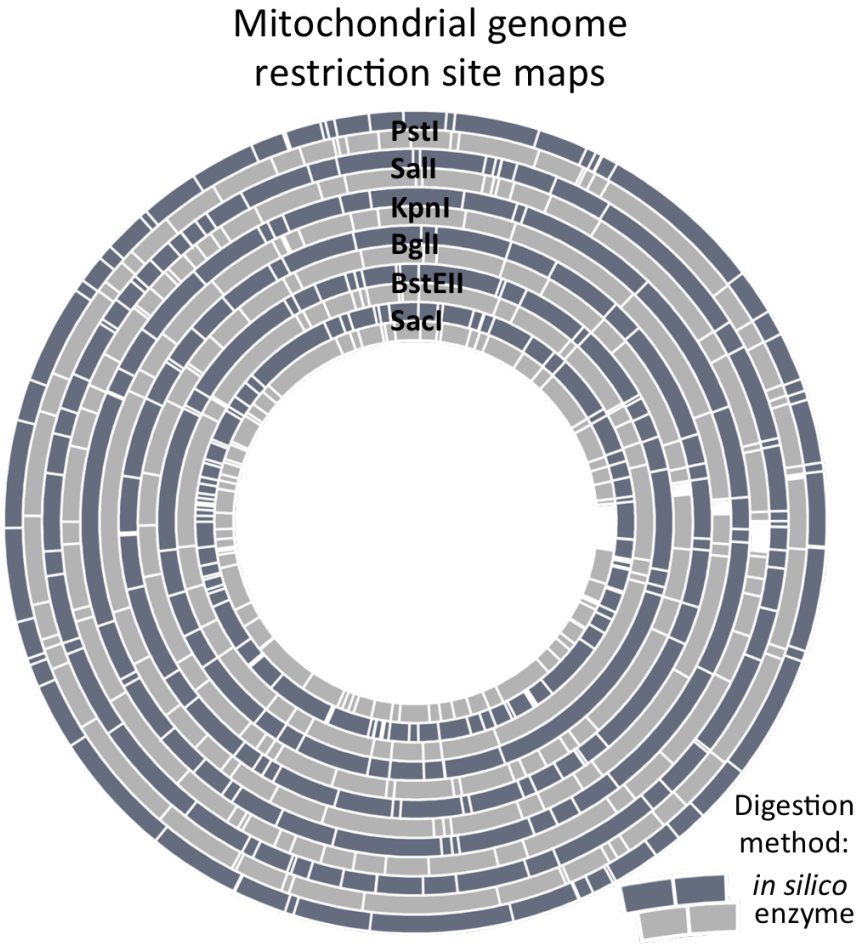

Figure 3.1: Comparison of restriction fragment maps. Dark grey segments show an *in silico* digestion of sunflower mitochondrial genome NCBI Reference Sequence: NC_023337. Light grey segments show fragment lengths of the enzyme digestion reported by Siculella and Palmer 1988.

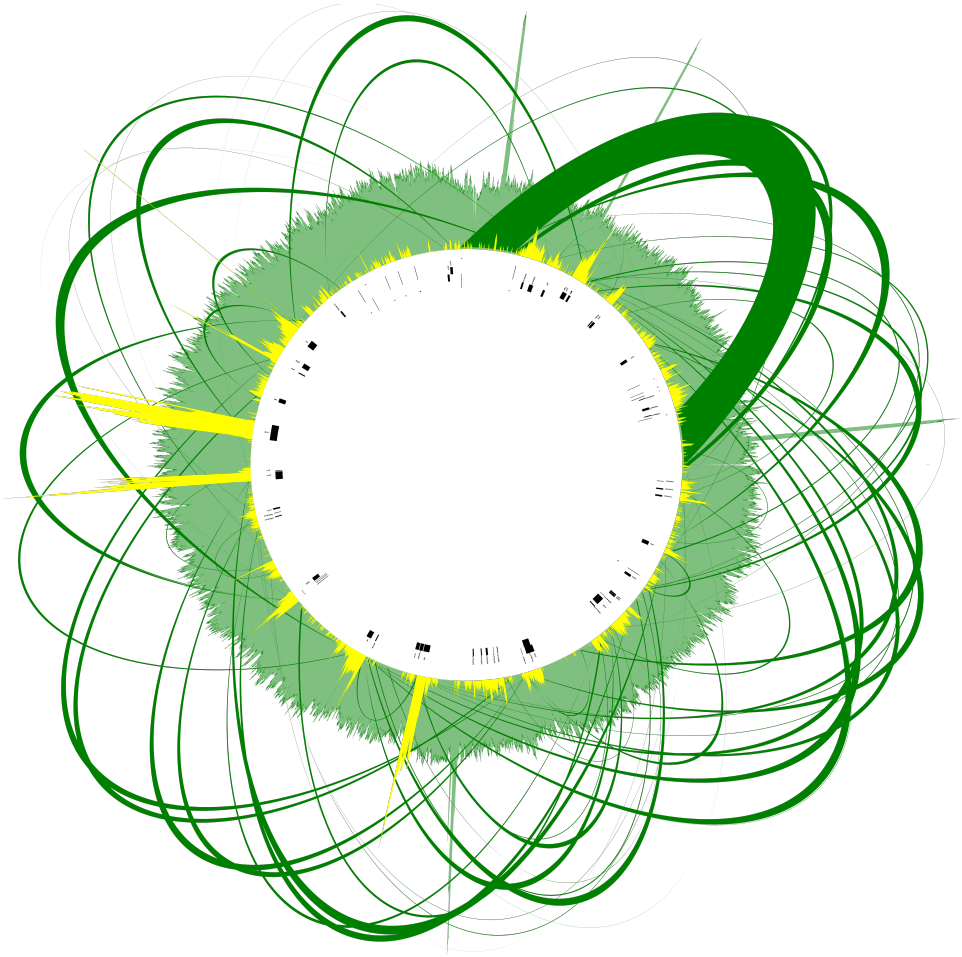

Figure 3.2: Gene and repeat map of sunflower mitochondrial genome NCBI Reference Sequence: NC_023337. Genic loci of Table 3.3 are indicated by black rectangles. Dark green ribbons attach to repeats. Light green rays show alignment depth of a WGS library to the reference mitochondrial genome. Note the coverage spikes, which indicate regions of high homology to the plastid genome. Drops in coverage near the large repeat boundary suggest that the cellular molarity of the mitochondrial genome's master replication circle is lower than the alternative configuration of two sub-circles. Yellow rays show alignment depth of coverage of an RNAseq library.

Table 3.1: Illumina Reads: Fragment Sizes

Type	*Library*	$Mean(bp)$	$Std.Dev.(bp)$
PairedEnd	$A1$	136	23
	$A2$	139	28
	$A5$	160	34
	$200bp\_HA0001$	192	21
	$500bp\_HA0002$	408	46
MatePair	$2kbp\_HA0003\_61YEJAAXX\_1$	1510	321
	$MP1$	2062	1744
	$MP2.BD0TEHACXX\_3$	2451	295
	$MP3.AC0C9VACXX\_4$	2500	272
	$LBM11326\_GFI-529\_3kb\_LJD$	2550	760
	$MP4.BD0TEHACXX\_5$	3320	443
	$HA412\_GGCTAC\_40kb\_LJD$	3458	1910
	$MP5$	3848	339
	$5kbp\_HA0004\_626E6AAXX\_5$	4418	846
	$MP6.BD0TEHACXX\_7$	4653	468
	$INX517*$	4394	321
	$LBM11326\_GFI-546\_40kb\_LJD$	5084	3642
	$INX518*$	5286	2016
	$LBM11325\_GFI-530\_8kb\_LJD$	7114	1090
	$LBM\_CAGATC\_8kb.LJD$	7132	1057
	$LBM\_GATCAG\_20kb.LJD$	13887	5153
	$LBM1481\_GFI-531\_20kb\_LJD$	16863	4578

Table 3.2: Roche 454 Reads: Fragment Sizes

Type	*Library*	*Mean*(bp)	*Std.Dev.*(bp)
ShortFragment	01*V*17*GRL*2	368	133
	*MPS*004761454*RL*	373	124
	*MPS*006655454*RL*	383	120
	01*V*17*G*454*RL*	384	136
	*MPS*004762454*RL*	394	125
	*HA*412*Long*	592	169
	*MAYha*412*long*	648	205
MatePair	01*V*17*G*454*PE*1	2890	485
	01*V*17*G*454*PE*2	2929	515
	*MPS*008920454*PE*55*kb*	3259	1148
	*MPS*008921454*PE*6*kb*	3517	1412
	*MPS*006655454*PE*20*Kb*	4491	3390
	*MPS*004761454*PE*38*kb*	7272	1060
	*MPS*008922454*PE*8*kb*	7584	1360
	*MPS*004761454*PE*210*kb*	7897	1146
	*MPS*008923454*PE*10*kb*	10041	1587
	*MPS*004761454*PE*10*kb*	10441	1932
	*MPS*004761454*PE*15*kb*	11157	5060
	*MPS*009917454*PE*20*kb*	12507	4667
	*MPS*008924454*PE*20*kb*	12955	4944
	*MPS*009918454*PE*20*kb*	13463	5190

Table 3.3: Sunflower Mitochondrial Genome Protein-Coding Features.

Class	*Start*	*End*	*Strand*	*Gene*	*Product*
CDS	16027	15284	−	*ccmC*	*cytochrome c biogenesis C*
	28498	27923	−	*atp*4	*ATPase subunit* 4
	28950	28678	−	*nad*4*L*	*NADH dehydrogenase subunit* 4*L*
	36771	37250	+	*atp*8	*ATPase subunit* 8
	37820	38617	+	*coxIII*	*cytochrome c oxidase subunit*
	43497	42934	−	*rpl*5	*ribosomal protein L*5
	66603	67223	+	*ccmB*	*cytochrome c biogenesis B*
	68019	67531	−	*rpl*10	*ribosomal protein L*10
	106128	107822	+	*coxI*	*cytochrome c oxidase subunit*
	112934	111735	−	*nad*5	*NADH dehydrogenase subunit* 5
	114601	114341	−	*atp*9	*ATPase subunit* 9
	122115	123110	+	*rps*4	*ribosomal protein S*4
	149093	149443	+	*rps*13	*ribosomal protein L*13
	169722	168793	−	*nad*6	*NADH dehydrogenase subunit* 6
	188450	189643	+	*cob*	*apocytochrome B*
	201645	200761	−	*ccmFc*	*cytochrome c biogenesis FC*
	202665	201790	−	*orf*873	*hypothetical protein*
	204362	202830	−	*atp*1	*ATPase subunit* 1
	215079	213361	−	*ccmFn*	*cytochrome c biogenesis FN*
	228434	230110	+	*ccmFn*	*cytochrome c biogenesis FN*
	230001	230516	+	*rpl*16	*ribosomal protein L*16
	251892	249925	−	*matR*	*maturase*
	254008	254364	+	*nad*3	*NADH dehydrogenase subunit* 3
	254416	254793	+	*rps*12	*ribosomal protein L*12
	260202	260774	+	*nad*9	*NADH dehydrogenase subunit*9
	269075	269980	+	*atp*6	*ATPase subunit* 6

Table 3.4: Sunflower Mitochondrial Genome RNA and Structural Features.

Class	*Start*	*End*	*Strand*	*Gene*	*Product*
tRNA	5785	5703	$-$	*trnY*	*tRNA* $-$ *Tyr*
	6659	6588	$-$	*trnN*	*tRNA* $-$ *Asn*
	8761	8691	$-$	*trnC*	*tRNA* $-$ *Cys*
	51553	51626	$+$	*trnD*	*tRNA* $-$ *Asp*
	75504	75585	$+$	*trnM*	*tRNA* $-$ *Met*
	79517	79589	$+$	*trnG*	*tRNA* $-$ *Gly*
	82782	82853	$+$	*trnQ*	*tRNA* $-$ *Gln*
	87906	87833	$-$	*trnH*	*tRNA* $-$ *His*
	89834	89905	$+$	*trnE*	*tRNA* $-$ *Glu*
	64558	64486	$-$	*trnK*	*tRNA* $-$ *Lys*
	170075	170001	$-$	*trnP*	*tRNA* $-$ *Pro*
	170454	170381	$-$	*trnF*	*tRNA* $-$ *Phe*
	170923	170836	$-$	*trnS*	*tRNA* $-$ *Ser*
	261753	261826	$+$	*trnW*	*tRNA* $-$ *Trp*
	300889	300817	$-$	*trnK*	*tRNA* $-$ *Lys*
rRNA	128775	132510	$+$	*rrn*26	26*S ribosomal RNA*
	139908	140023	$+$	*rrn*5	5*S ribosomal RNA*
	140166	142111	$+$	*rrn*18	18*S ribosomal RNA*
repeat	51682	64614	*n*/*a*	*n*/*a*	*large repeat copy*1
	288012	300945	*n*/*a*	*n*/*a*	*large repeat copy*2

Chapter 4

Pseudomolecules

4.1 What is a Pseudomolecule?

The utilities of a reference genome include: representing the entire genome of one representative individual from a species, providing a common axis for inter-study comparisons, and contextualizing loci. The traditional representation of genetic sequence is as text, with individual letters corresponding to individual bases. An ideal reference genome would include a contiguous sequence of letters for each chromosome of the organism. These sequences are often referred to as pseudomolecules. In practice, many factors affect how closely a set of reference pseudomolecules matches the ideal. These factors include: genome content and degree of repetition, sequencing read length, and other positional information, such as genetic and physical maps.

The genome of sunflower line HA412HO was sequenced with high volume, short read technologies and assembled with algorithms described in a previous section. Many genetic maps have been created for sunflower, including the ultra-high density genetic map described in a previous section. A single physical map developed by postdoc Navdeep Gill using Keygene's sequence-based BAC fingerprinting approach, (Van 2011) was available during the span of this project. Our pseudomolecules are a synthetic amalgamation of these resources. We also applied quality control steps to remove technical artifacts of the sequencing and assembly process.

While we had a variety of high-quality resources available, our final pseudomolecules contain many gaps. The factor limiting the achievement our goal of producing a highly contiguous reference is the sunflower genome's biology; it is highly repetitive. Approximately 85% of the sunflower genome is high-copy sequence (Staton et al. 2012). Approximately 50% of the genome is a *Ty3/gypsy* LTR retrotransposon, about 10kbp in length, with an estimated 1% divergence between element copies. We were unable to place many of the sunflower genome's repeats within pseudomolecules and they are included in the reference artificially concatenated as the so-called sequence Q. We have however, made progress towards quantifying and localizing some repetitive genomic features, namely the centromeres, telomeres, and ribosomal repeats.

4.2 Combining Genetic and Physical Maps

Our ultra-high density genetic map and Finger Printed BAC Contig (FPC) physical map were both useful for ordering genomic loci. They are, however, of maximum utility at different scales. We believe our genetic map to be nearly saturated; that is, we have accounted for all observable recombination events. The lengths of the *de novo* assembled scaffolds were usually shorter than the distance between any two consecutive and observable recombination events. The result is that several scaffolds may share the same genetic position, which may alternatively be referred to as a genetic bin. Within a bin, relative scaffold ordering is unknown. Recombination rate varies widely throughout the genome and genetic distance is not well correlated to physical distance at the chromosome scale.

The physical map is a collection of contigs constructed from fingerprinted BACs. The fingerprinting and contig construction are described briefly below. A library of BACs is constructed covering the genome to approximately 12x depth. The BACs are digested with a restriction enzyme. Digested fragments are barcoded with oligonucleotides such that the BAC they originated from may be determined later. They are sequenced using short Illumina reads that begin at the cut site. We refer to a set of reads originating from a BAC as physical map tags. This set of tags is the BAC's fingerprint. Contigs are constructed by comparing fingerprints to each other. Fingerprints partially shared between BACs indicate overlap. Several

tiled fingerprints form a contig. Note that the tag order within a contig is inferred from the presence or absence of tags within adjacent tiled BACs, therefore only a partial ordering may be inferred for some tag subsets.

Our physical map is useful at a more granular scale than the genetic map. It provides an estimate of the number of base pairs between scaffolds and their relative orientation. Physical map contigs, however, are not ordered or oriented relative to each other, nor are they anchored to a chromosome. Integrating the genetic and physical maps exploits the complementary information gained from each to overcome their individual limitations. The *de novo* assembly is used to do so.

First, the *de novo* genome assembly is searched for physical map tag sequences using BLAST (blastn -evalue 10000 -outfmt 7 -dust no -word_size 7 -perc_identity 96). For each tag, all hits with a bit score equal to the highest for the tag are retained. For each scaffold, a candidate set of matching physical map contigs is generated by searching for those sharing matching tags. The tag-to-scaffold bit scores are summed for all tags shared between a contig and a scaffold. The three highest scores are retained. In case of ties, all contigs with the three highest scores are retained as candidate matches.

I wrote a small piece of software to order and orient scaffolds using the physical map. A scaffold is matched to a contig using the alignment-scoring scheme described below; an example alignment is provided in Table 4.1. For a given scaffold-to-contig alignment, only tags shared between both are considered. That is, there is no mismatch penalty. First, tags are ordered according to their starting position in the scaffold. Note that some tags may share the same position in a physical map contig. The tag holding the lowest position in the scaffold receives a score of one and its contig position index is recorded. The contig position index of each subsequent tag is checked. For tag_i to tag_{i+1}: if there is no change in contig position index, the match score is increased by one; if the contig position index increases by one, the match score is increased by two; if the contig position index increases by more than one, the match score is not changed; if the contig position index decreases, the match score is decreased by two. The tag orders are then reversed and the process is repeated. After searching all candidate contigs in both directions, the highest score is chosen, giving both the final matching contig and the scaffold orientation within the contig (Figure 4.1).

4.3 The Golden Path

The FPC assembly software (Nelson and Soderlund 2009) provides distances measured in custom units. The mean BAC length for all physical map contigs was 21.23503 FPC units. We fully sequenced and assembled 100 BAC clones in order to estimate their physical size in base pairs Figure 4.3. The mean assembly length of sequenced BACs was 149678.9 bp. The pseudomolecules were constructed using the conversion of 7049 base pairs per 1 FPC unit.

I determined the initial gap length between member scaffolds using the following method. I first summed the lengths of member scaffolds. This sum was subtracted from the estimated length of the physical map contig. If the difference was greater than twice the number of scaffolds, each member scaffold was padded with Ns on either side with the difference divided by twice the number of scaffolds. When the lengths of member *de novo* assembled scaffolds assigned to super-scaffold lengths were compared to the estimated lengths of the physical map contigs they were based on, the distribution of differences were centered above zero, but some differences were negative. If the difference was less than twice the number of scaffolds, each scaffold was padded with a single N on either side.

Introducing gaps between scaffolds such that the super-scaffold matches its corresponding physical map contig without compensating for cases in which the sum of scaffold lengths exceeded the physical map contig size would have the effect of inflating the length of a pseudomolecule above that physical map estimate. I also included scaffolds in the pseudomolecules if we could anchor them to a linkage group, even if they could not be assigned to a physical map contig. Thus, I adjusted gap length between *de novo* assembled scaffolds such that the total length of a pseudomolecule matched the sum of physical map contigs assigned to it.

Scaffold orders within a genetic bin are taken from the physical map. Scaffold lengths are summed for a genetic bin. The genetic distance of a bin is divided among member scaffolds proportional to their length to assign them pseudo-cM positions. The desired length of a chromosome's pseudomolecule is obtained by summing the length of all physical map contigs assigned to it. The chromosome is initially divided into 1cM windows. The difference of lengths estimated via the physical map and the sum of scaffold lengths is taken for each window. If the differ-

ence is positive for all windows, it is divided among scaffolds in proportion to their pseudo cM position. If the smallest proportion is less than two, window expansion continues. Otherwise, optimal window size has been found. All between-scaffold gaps are thus positive and the sum of all gap and scaffold lengths is equal to the sum of all physical map contig lengths. The pseudomolecule is then printed. Figure 4.2 shows the positions of *de novo* assembled scaffolds in physical and genetic space.

4.4 Masking

Preliminary analysis suggested that some non-biological duplicated sequence was present in the merged assembly. When I looked at alignments of EST sequences to the genome, about a third of the alignments were duplicates covering the entire transcript at over 99% identity. These duplicates are not present in the Allpaths-LG assembly, nor are they present in the Celera assembly. I thus took measures to remove these technical artefacts. I applied the following methods were to each linkage group separately.

A database of know repetitive elements was constructed by concatenating the SUNREP database (Natali et al. 2013), Repbase (Jurka et al. 2005) repeats classified as present in asterids, sunflower full-length LTR-RT families, transposable elements known to be active in sunflowers (Gill et al. 2014), ribosomal DNA (Bock et al. 2014), and cytoplasmic reference genomes (Chapter 2, Timme et al. 2007). This database was used to hard-mask (replacing ATCGs with Ns) the Allpaths and Celera subassemblies with RepeatMasker (Tarailo-Graovac and Chen 2009). Masked subassembly scaffolds were split into contigs at runs of Ns longer than nine. For each subassembly, contigs were aligned to themselves with BLAST (-dust no -perc_identity 99). The coordinates of non-self matches were hard-masked, retaining single-copy sequence. Masked contigs were again split into contigs at runs of Ns longer than nine. Single copy sequence entries from each subassembly were concatenated into file and then clustered at 99% identity using cd-hit-est (Li and Godzik 2006) to remove redundant sequences. In order to ensure the resulting non-redundant sequences were single copy in both sub-assemblies, they were aligned to both subassemblies separately using BLAST. Query sequence with more than

one match longer than 100 base pairs were masked from the non-redundant set, and again split into contigs at runs of Ns longer than nine.

In all, 465,802,996bp assigned to a chromosome were identified as single copy sequence at 99% identity. These sequences were used to identify technical artifacts in the merged assembly. Single copy sequences were aligned to the reference assembly using BLAST. Matches with greater than 99% identity, at least 100bp in length, and spanning at least 90% of the query length were inspected for copy number. If a query sequence matched the subject sequence twice, the pair of subject matches was flagged as containing a technical artefact. The genetic position of the query sequence was compared to both matches in the pair. If the genetic position of one match of the pair differed from that listed in the subassembly, it was flagged for masking. If the genetic positions of the subject matches were the same, we chose one at random to be masked. The regions determined to be technical artefacts were masked using BEDTools (Quinlan and Hall 2010).

4.5 Seventeen Pseudomolecules

The final reference set of pseudomolecules is similar to the expected genome length (3.64 Gbp versus 3.6 Gbp expected) (Baack et al. 2005), with a super-scaffold N50 of 210kbp. The genome includes greater than 98% of CEGMA (Core Eukaryotic Genes Mapping Approach) (Parra et al. 2007) genes, of which approximately 90% are full length, indicating that the gene space is well covered. The genome has been fully annotated by colleagues at INRA and includes approximately 39k strongly supported protein-coding gene models (excluding transposable elements). It is displayed in JBrowse (Skinner et al. 2009) and is accompanied by numerous tools for searching, mapping, and functional analyses (http://www.sunflowergenome.org). The number of protein-coding genes, length in centiMorgans, length in base pairs, and number of nucleotides assigned to each pseudomolecule is tabulated in Table 4.2.

Figure 4.1: Diagram showing integration of genetic map, *de novo* genome assembly, and physical map on Linkage Group 4. Genetic map bins positions are shown in dark grey. Scaffolds are shown in green, with the scaffold base pair position and physical map tag sequences in the neighboring columns. The red bar at the top of the diagram shows the physical map contig of the member tags with FPC units in the row below. Yellow bars indicate a minimum tiling path of BACs. Orange rectangles indicate alignment matches of scaffold tag positions with the corresponding position in the FPC physical map contig.

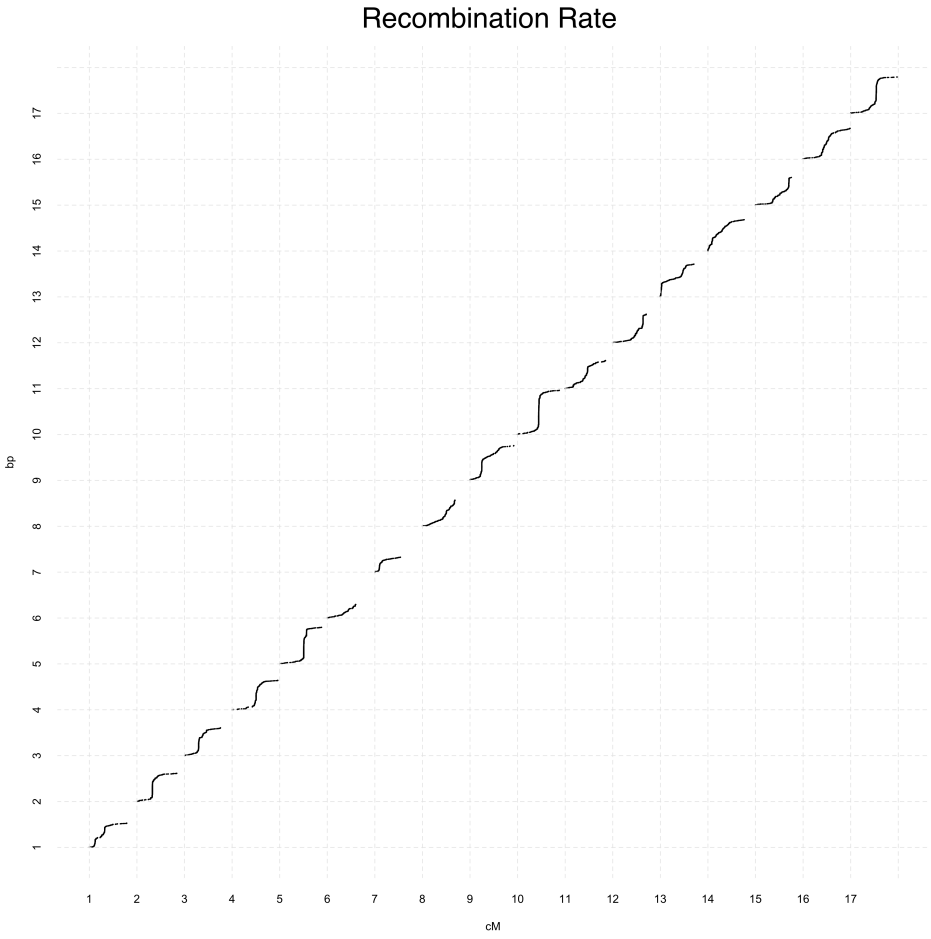

Figure 4.2: Scaffold positions plotted in RHA280 x RHA801 cM (x-axis) and HA412 physical map bp (y-axis). Each numbered cell contains a chromosome's plot. Regions of extreme recombination suppression are shown where the slope of a line is close to infinity. I suspect these regions harbor centromeric loci. Conversely, regions of the chromosome that recombine frequently are shown where the slope of a line is close to zero.

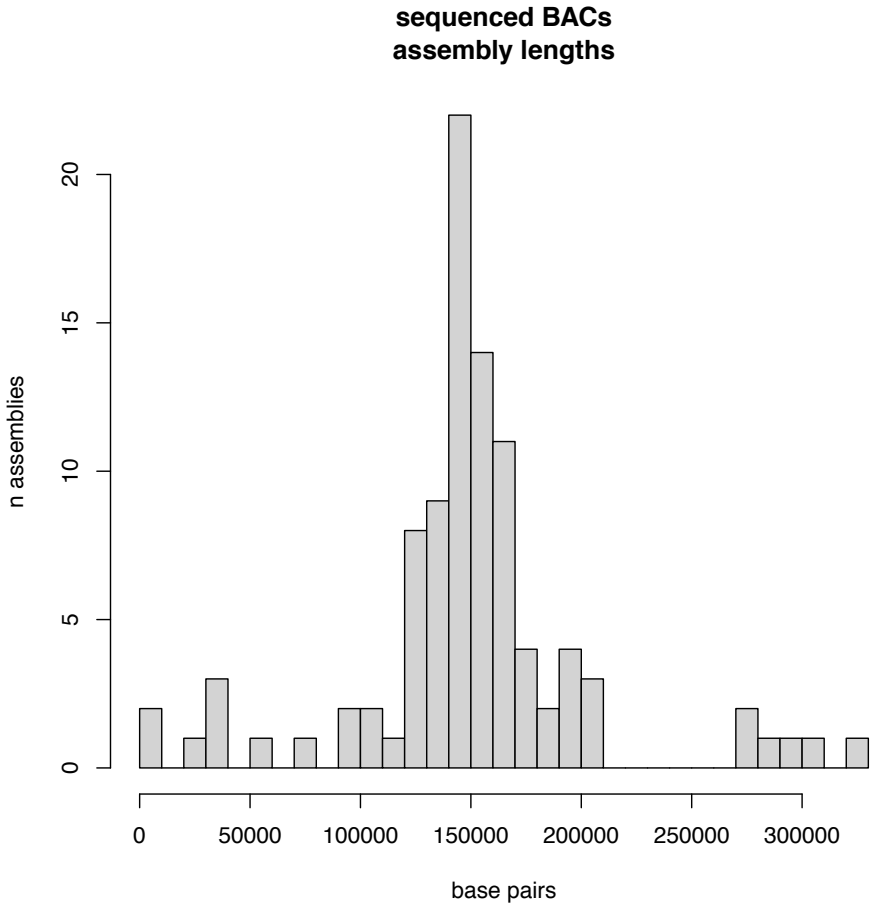

Figure 4.3: Frequency distribution of length in base pairs for 100 fully sequenced and assembled BACs.

Table 4.1: Example Alignment: Scaffold403 to LG8-Ctg66

$FPCunits:$	100	102	115		
$\sum score$	Δ	Δ	Δ	$forward$	
				$start(bp)$	$tag\_sequence$
1	0	0	+1	17724	*GAATTCCGAACACACTGATGTGATTA*
2	0	0	+1	17746	*GAATTCGTTGTAAAACAGAGATATGATTTC*
3	0	0	+1	18170	*GAATTCTAGAATATCCTTGAATACAACCAT*
4	0	0	+1	18974	*GAATTCAAGGAAACACGAAATGAGTGGTTT*
5	0	0	+1	20734	*GAATTCATTTTCATCAACATGCATCATCTT*
6	0	0	+1	20758	*GAATTCAAGGTTGATTTTGAAGAAGAACTG*
4	0	−2	0	25585	*GAATTCGAGCTAGCTCGGCTTGGCTCGATC*
2	−2	0	0	25609	*GAATTCTAATCAAGCCGAGCTCGAGCCTCA*
				$reverse$	
1	+1	0	0	25609	*GAATTCTAATCAAGCCGAGCTCGAGCCTCA*
3	0	+2	0	25585	*GAATTCGAGCTAGCTCGGCTTGGCTCGATC*
5	0	0	+2	20758	*GAATTCAAGGTTGATTTTGAAGAAGAACTG*
6	0	0	+1	20734	*GAATTCATTTTCATCAACATGCATCATCTT*
7	0	0	+1	18974	*GAATTCAAGGAAACACGAAATGAGTGGTTT*
8	0	0	+1	18170	*GAATTCTAGAATATCCTTGAATACAACCAT*
9	0	0	+1	17746	*GAATTCGTTGTAAAACAGAGATATGATTTC*
10	0	0	+1	17724	*GAATTCCGAACACACTGATGTGATTA*

Table 4.2: Final Pseudomolecule Statistics

Chromosome	*ngenes*	$length(cM)$	$length(bp)$	*nATCG*
1	2535	78.52	175,985,764	99,635,607
2	2050	83.61	209,013,747	116,957,742
3	2567	75.70	203,472,901	111,263,426
4	2486	95.99	216,026,857	114,464,986
5	2538	88.06	271,056,985	147,484,857
6	1654	59.68	100,519,666	57,620,576
7	1569	54.03	109,221,022	60,893,579
8	2240	68.46	192,129,815	105,634,875
9	3300	91.98	253,478,808	139,276,314
10	3233	87.89	327,788,049	183,694,265
11	2168	84.69	208,730,832	109,503,895
12	2591	70.22	208,068,730	114,409,345
13	2732	70.56	239,367,298	137,400,774
14	2613	76.30	230,295,834	119,919,823
15	2326	75.34	202,246,870	110,705,372
16	2350	99.13	226,777,971	115,811,864
17	2699	100.77	267,415,242	144,655,486
total	41,651	1360.94	3,641,596,391	1,989,332,786

Chapter 5

Conclusion

I have produced a set of pseudomolecules representing the seventeen chromosomes of sunflower. Additionally, I have closed the mitochondrial genome's master circle. Together with the previously assembled plastid genome, nearly all the DNA of a sunflower can now be easily browsed as graphics online.

This project required integrating many sources of information to deliver a good that will aid knowledge synthesis. The physical map, genetic map, and *de novo* assemblies all provide useful information, but at different scales. At this point in time, leveraging all three was necessary to model the sunflower's genome as stretched out strings of DNA.

The delivery of a reference genome is largely a technical achievement. The immediate question it answers (i.e. what is the linear sequence of DNA in a single, highly inbred, line?) is narrow. On its own, it allows us to address other biological questions, such as: How do repeat families cluster in space? How do recombination rates vary? What is their relationship to sequence features? Where are protein-coding genes located? What is the syntenic relationship of the sunflower's paleologs?

While these are interesting questions, reference genomes are most powerful when used as an x-axis against which to plot measures pertinent to the study of macroevolution, population genetics, and functional morphology. Perhaps a population geneticist will gain new insight into the mechanisms driving differentiation by viewing the Fixation Index (*Fst*) outliers between two populations on an axis

shared with the tests for selection of a multi-species comparison, transcript expression measures values of a gene expression experiment, or the Quantitative Trait Loci of a mapping cross.

The reference described here is currently being used to genotype nearly 500 accessions of sunflowers via the alignment of low-coverage short reads. Alignment to a common reference is facilitating the use of accurate Bayesian models of genotyping. The resulting matrix of genotypes will allow researchers to model reticulate evolution in the genus and help understand mechanisms of speciation in the face of high levels of gene flow. Additionally, companies involved in the sunflower genome consortium have favorably reviewed the reference genome as useful in elite breeding.

Recent technological breakthroughs in sequencing technology (i.e. PacBio (Eid et al. 2009) and Oxford Nanopore (Bayley 2015)) have resulted in read lengths measured in kilobases. New methods for long-range scaffolding using read libraries prepared from precipitated chromatin (Lieberman-Aiden et al. 2009) are being developed as well (Burton et al. 2013). The next generation of sunflower genomes assembled *de novo* will be based on these technologies. These will likely supplant the reference discussed here. However, many of the tools, methods, and resources we developed for the HA412 reference will be reused to produce new sets of pseudomolecules from the new *de novo* assemblies.

Bibliography

Ahmed, R., Yousaf, J., Nadeem, I., Saleem, M., and Ali, A. (2013). Response of sunflower (*Helianthus annuus L.*) hybrids to population of different insect pests and their bio-control agents. *Journal of Agriculutural Research*, 51(1).

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., Barry, K., and Palmer, J. D. (2010). Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (cucurbitaceae). *Molecular Biology and Evolution*, 27(6):1436–1448.

Anderson, E. (1931). Internal factors affecting discontinuity between species. *American Naturalist*, pages 144–148.

Anderson, E. (1936). The species problem in *Iris*. *Annals of the Missouri Botanical Garden*, pages 457–509.

Anderson, E. (1949). *Introgressive hybridization.* John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London.

Anderson, E. and Hubricht, L. (1938). Hybridization in *Tradescantia*. III. the evidence for introgressive hybridization. *American Journal of Botany*, pages 396–402.

Andrew, R. L., Kane, N. C., Baute, G. J., Grassa, C. J., and Rieseberg, L. H. (2013). Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, 22(3):799–813.

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 Phytologist*, 167(2):623–630.

Barb, J. G., Bowers, J. E., Renaut, S., Rey, J. I., Knapp, S. J., Rieseberg, L. H., and Burke, J. M. (2014). Chromosomal evolution and patterns of introgression in *Helianthus*. *Genetics*, 197(3):969–979.

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. 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. *Molecular Biology and Evolution*, 25(11):2445–2455.

Baute, G. J., Kane, N. C., Grassa, C. J., Lai, Z., and Rieseberg, L. H. (2015). Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytologist*, 206(2):830–838.

Bayley, H. (2015). Nanopore sequencing: From imagination to reality. *Clinical Chemistry*, 61(1):25–31.

Beckstrom-Sternberg, S., Rieseberg, L. H., and Doan, K. (1991). Gene lineage analysis in populations of *Helianthus niveus* and *H. petiolaris* (Asteraceae). *Plant Systematics and Evolution*, 175(3-4):125–138.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinions in Genetics & Development*, 16(6):545–552.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Blackman, B. K., Scascitelli, M., Kane, N. C., Luton, H. H., Rasmussen, D. A., Bye, R. A., Lentz, D. L., and Rieseberg, L. H. (2011). Sunflower domestication alleles support single domestication center in eastern North America. *Proceedings of the National Academy of Sciences*, 108(34):14360–14365.

Blamey, F., Zollinger, R. K., and Schneiter, A. A. (1997). Sunflower production and culture. *Sunflower Technology and Production*, pages 595–670.

Bock, D. G., Kane, N. C., Ebert, D. P., and Rieseberg, L. H. (2014). Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytologist*, 201(3):1021–1030.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4):578–579.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, page btu170.

Bowers, J. E., Bachlava, E., Brunick, R. L., Rieseberg, L. H., Knapp, S. J., and Burke, J. M. (2012). Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *G3: Genes— Genomes— Genetics*, 2(7):721–729.

Buerkle, C. A. and Rieseberg, L. H. (2008). The rate of genome stabilization in homoploid hybrid species. *Evolution*, 62(2):266–275.

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.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125.

Chandler, J. M., Jan, C.-C., and Beard, B. H. (1986). Chromosomal differentiation among the annual *Helianthus* species. *Systematic Botany*, pages 354–371.

Cheres, M. T. and Knapp, S. J. (1998). Ancestral origins and genetic diversity of cultivated sunflower: coancestry analysis of public germplasm. *Crop Science*, 38(6):1476–1482.

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991.

Darwin, C. (1859). *On the origins of species by means of natural selection*. London: Murray.

Dempewolf, H., Eastwood, R. J., Guarino, L., Khoury, C. K., Müller, J. V., and Toll, J. (2014). Adapting agriculture to climate change: A global initiative to collect, conserve, and use crop wild relatives. *Agroecology and Sustainable Food Systems*, 38(4):369–377.

Dirichlet, G. L. (1850). Über die reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *Journal für die Reine und Angewandte Mathematik*, 40:209–227.

Dussle, C., Hahn, V., Knapp, S., and Bauer, E. (2004). Pl Arg from *Helianthus argophyllus* is unlinked to other known downy mildew resistance genes in sunflower. *Theoretical and Applied Genetics*, 109(5):1083–1086.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research*, 8(3):186–194.

Feder, J. L., Egan, S. P., and Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7):342–350.

Feng, J., Liu, Z., Cai, X., and Jan, C.-C. (2013). Toward a molecular cytogenetic map for cultivated sunflower (*Helianthus annuus L.*) by landed BAC/BIBAC clones. *G3: Genes— Genomes— Genetics*, 3(1):31–40.

Fick, G., Zimmer, D., and Kinman, M. (1974). Registration of six sunflower parental lines (Reg. No. PL 1 to 6). *Crop Science*, 14(6):912–912.

Galton, F. (1894). *Natural inheritance*. Macmillan.

Gill, N., Buti, M., Kane, N., Bellec, A., Helmstetter, N., Berges, H., and Rieseberg, L. H. (2014). Sequence-based analysis of structural organization and composition of the cultivated sunflower (*Helianthus annuus L.*) genome. *Biology*, 3(2):295–319.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.

Harlan, J. R. and De Wet, J. (1963). The compilospecies concept. *Evolution*, pages 497–501.

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*, 430(6996):201–205.

Heiser, C. B. (1951). The sunflower among the North American Indians. *Proceedings of the American Philosophical Society*, pages 432–448.

Heiser, C. B., Martin, W. C., and Smith, D. (1962). Species crosses in *Helianthus*: I. Diploid species. *Brittonia*, 14(2):137–147.

Heiser, C. B. and Smith, D. M. (1954). New chromosome numbers in *Helianthus* and related genera (Compositae). In *Proceedings of the Indiana Academy of Science*, volume 64, pages 250–253.

Heiser, C. B., Smith, D. M., Clevenger, S. B., and Martin, W. (1969). *North American sunflowers (Helianthus).*, volume 22 of *Memoirs of the Torrey Pines Botanical Club*. Durham.

Heiser Jr, C. B. (1947). Hybridization between the sunflower species *Helianthus annuus* and *H. petiolaris*. *Evolution*, pages 249–262.

Hodgins, K. A., Lai, Z., Oliveira, L. O., Still, D. W., Scascitelli, M., Barker, M. S., Kane, N. C., Dempewolf, H., Kozik, A., Kesseli, R. V., et al. (2014). Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives. *Molecular Ecology Resources*, 14(1):166–177.

Horn, R., Köhler, R. H., and Zetsche, K. (1991). A mitochondrial 16 kDa protein is associated with cytoplasmic male sterility in sunflower. *Plant Molecular Biology*, 17(1):29–36.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467.

Kane, N., Gill, N., King, M., Bowers, J., Berges, H., Gouzy, J., Bachlava, E., Langlade, N., Lai, Z., Stewart, M., et al. (2011). Progress towards a reference genome for sunflower. *Botany*, 89(7):429–437.

Kane, N. C., King, M. G., Barker, M. S., Raduski, A., Karrenberg, S., Yatabe, Y., Knapp, S. J., and Rieseberg, L. H. (2009). Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution*, 63(8):2061–2075.

Karrenberg, S., Edelist, C., Lexer, C., and Rieseberg, L. (2006). Response to salinity in the homoploid hybrid species *Helianthus paradoxus* and its progenitors *H. annuus* and *H. petiolaris*. *New Phytologist*, 170(3):615–629.

Kelley, D. R., Schatz, M. C., Salzberg, S. L., et al. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11(11):R116.

Kernighan, B. W. and Mashey, J. R. (1979). *The* UNIX *programming environment*, volume 9. Wiley Online Library.

Kosambi, D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics*, 12(1):172–175.

Lane, M. A. (2003). The global biodiversity information facility. *Bulletin of the American Society for Information Science and technology*, 30(1):22–24.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, W. and Godzik, A. (2006). CD-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18.

Mandel, J., Dechaine, J., Marek, L., and Burke, J. (2011). Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus L. Theoretical and Applied Genetics*, 123(5):693–704.

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

McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., Burke, J. M., Charest, D., Cloutier, S., Cole, G., et al. (2013). Agriculture: feeding the future. *Nature*, 499(7456):23–24.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Miller, J., Gulya, T., and Vick, B. (2006). Registration of three maintainer (HA 456, HA 457, and HA 412 HO) high-oleic oilseed sunflower germplasms. *Crop science*, 46(6):2728–2728.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824.

Natali, L., Cossu, R. M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N. C., Rieseberg, L., et al. (2013). The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics*, 14(1):686.

Nelson, W. and Soderlund, C. (2009). Integrating sequence with FPC fingerprint maps. *Nucleic Acids Research*, 37(5):e36–e36.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067.

Quinlan, A. R. and Hall, I. M. (2010). BEDtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rauf, S. (2008). Breeding sunflower (*Helianthus annuus L.*) for drought tolerance. *Communications in Biometry and Crop Science*, 3(1):29–44.

Renaut, S., Grassa, C., Yeaman, S., Moyers, B., Lai, Z., Kane, N., Bowers, J., Burke, J., and Rieseberg, L. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, 4:1827.

Rieseberg, L. H. (1991). Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. *American Journal of Botany*, pages 1218–1237.

Rieseberg, L. H., Baird, S. J., and Desrochers, A. M. (1998). Patterns of mating in wild sunflower hybrid zones. *Evolution*, pages 713–726.

Rieseberg, L. H., Beckstrom-Sternberg, S. M., Liston, A., and Arias, D. M. (1991). Phylogenetic and systematic inferences from chloroplast DNA and isozyme variation in *Helianthus sect. Helianthus* (Asteraceae). *Systematic Botany*, pages 50–76.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., and Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301(5637):1211–1216.

Rieseberg, L. H. and Seiler, G. J. (1990). Molecular evidence and the origin and development of the domesticated sunflower (*Helianthus annuus*, Asteraceae). *Economic Botany*, 44(3):79–91.

Roath, W., Miller, J., and Gulya, T. (1981). Registration of RHA 801 sunflower germplasm (Reg. No. GP 5). *Crop Science*, 21(3):479.

Rothberg, J. M. and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10):1117–1124.

Rothfels, C. J., Johnson, A. K., Hovenkamp, P. H., Swofford, D. L., Roskam, H. C., Fraser-Jenkins, C. R., Windham, M. D., and Pryer, K. M. (2015). Natural hybridization between genera that diverged from each other approximately 60 million years ago. *The American Naturalist*, 185(3):433–442.

Sambatti, J., Strasburg, J. L., Ortiz-Barrientos, D., Baack, E. J., and Rieseberg, L. H. (2012). Reconciling extremely strong barriers with high levels of gene exchange in annual sunflowers. *Evolution*, 66(5):1459–1473.

Scascitelli, M., Whitney, K., Randell, R., King, M., Buerkle, C., and Rieseberg, L. (2010). Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H. debilis*) reveals asymmetric patterns of introgression and small islands of genomic differentiation. *Molecular Ecology*, 19(3):521–541.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323(5915):737–741.

Seiler, G. J. (1992). Utilization of wild sunflower species for the improvement of cultivated sunflower. *Field Crops Research*, 30(3):195–230.

Seiler, G. J. and Jan, C.-C. (2014). Wild sunflower species as a genetic resource for resistance to sunflower broomrape (*Orobanche cumana Wallr.*). *Helia*, 37(61):129–139.

Siculella, L. and Palmer, J. D. (1988). Physical and gene organization of mitochondrial DNA in fertile and male sterile sunflower. CMS-associated alterations in structure and transcription of the *atpA* gene. *Nucleic Acids Research*, 16(9):3787–3799.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., and Holmes, I. H. (2009). JBrowse: a next-generation genome browser. *Genome Research*, 19(9):1630–1638.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, 8(1):64.

Staton, S. E., Bakken, B. H., Blackman, B. K., Chapman, M. A., Kane, N. C., Tang, S., Ungerer, M. C., Knapp, S. J., Rieseberg, L. H., and Burke, J. M. (2012). The sunflower (*Helianthus annuus L.*) genome reflects a recent history of biased accumulation of transposable elements. *The Plant Journal*, 72(1):142–153.

Stebbins, J. C., Winchell, C. J., and Constable, J. V. (2013). *Helianthus winteri* (Asteraceae), a new perennial species from the southern Sierra Nevada foothills, California. *Aliso: A Journal of Systematic and Evolutionary Botany*, 31(1):19–24.

Tang, S., Leon, A., Bridges, W. C., and Knapp, S. J. (2006). Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower. *Crop Science*, 46(2):721–734.

Tang, S., Yu, J.-K., Slabaugh, M., Shintani, D., and Knapp, S. (2002). Simple sequence repeat map of the sunflower genome. *Theoretical and Applied Genetics*, 105(8):1124–1136.

Tarailo-Graovac, M. and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, pages 4–10.

Timme, R. E., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2007). A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany*, 94(3):302–312.

Van Nieuwerburgh, F., Thompson, R. C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P., and Head, S. R. (2011). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Research*, page gkr1000.

Vogel, H. (1979). A better way to construct the sunflower head. *Mathematical Biosciences*, 44(3):179–189.

Wan, S., Jiao, Y., Kang, Y., Jiang, S., Tan, J., Liu, W., and Meng, J. (2013). Growth and yield of oleic sunflower (*Helianthus annuus L.*) under drip irrigation in very strongly saline soils. *Irrigation Science*, 31(5):943–957.

Wu, Y., Bhat, P. R., Close, T. J., and Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genetics*, 4(10):e1000212.

Yatabe, Y., Kane, N. C., Scotti-Saintagne, C., and Rieseberg, L. H. (2007). Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics*, 175(4):1883–1893.

Yu, J.-K., Mangor, J., Thompson, L., Edwards, K. J., Slabaugh, M. B., and Knapp, S. J. (2002). Allelic diversity of simple sequence repeats among elite inbred lines of cultivated sunflower. *Genome*, 45(4):652–660.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829.