ABSTRACT

Understanding how genomes evolve can help us disentangle the evolutionary history of species, which has practical applications in agriculture. The Camelina genus contains nine species including the emerging oilseed crop, Camelina sativa, and its wild relative, Camelina microcarpa. Hexaploid C. microcarpa is almost perfectly sexually compatible with C. sativa, and diploid C. microcarpa is a potential descendant of one of C. sativa’s progenitors. Thus, these species are likely to have a complex history of hybridization and polyploidy that is unexplored. We have sequenced and assembled a draft genome of diploid C. microcarpa into 1036 scaffolds which cover 73% of the 264 Mb estimated genome size, and evaluated its evolutionary relationship with the chromosomes of C. sativa through the construction of phylogenies for each conserved ancestral crucifer karyotype (ACK) block. Based on these phylogenies, I propose revisions to the previously described C. sativa karyotype and that C. microcarpa is most closely related to subgenome 2 of C. sativa. Sequencing C. microcarpa is the key step to understand genome evolution in the Camelina genus.

KEY WORDS

Genome Assembly, Genome Evolution, Sequencing, Phylogeny, Polyploidy, Oilseed
ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Sara Martin for her mentorship and support throughout my studies. Also, my co-supervisor Dr. Owen Rowland for his guidance and encouragement during my degree. I would also like to extend my thanks to my coworkers at Agriculture and Agri-Food Canada, Connie Sauder and Tracey James for making my days at Neatby more enjoyable.

Above all I would like to thank my family and friends that have always been my support and drive in life. My parents Antonio Lujan and Beatriz Toro, you have always encouraged me to follow my dreams, and made sure I was able to do so, you build the strong foundation in which I stand on today, and for that I am extremely thankful. My siblings Alejandro, Andreina and Daniela, thank you for being my pillars, you have always been there for me, heard me and lent a hand every time I needed it. I am also extremely lucky to have had the support of my friends Ahlam and Prabh that helped at the start of my thesis and made sure I could take off on this wonderful journey. Lastly, to Danyal, who made sure I didn’t give up, trusted me, and gave me the strength to get all the way to the finish line, thank you.
STATEMENT OF CONTRIBUTIONS

In this study, Connie Sauder raised the plants, extracted the DNA and sent it for sequencing to the Centre for Applied Genomics (The Hospital for Sick Children, Toronto, Canada), the Microbial Molecular Technologies Laboratory (Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Canada) and the Genome Quebec Innovation Centre (McGill University, Quebec, Canada). Dr. Sara Martin wrote and executed the R program used to find and align overlapping fragments between $C. \text{sativa}$, $C. \text{microcarpa}$ and $A. \text{lyrata}$. All other bioinformatic analyses were carried out by me.
# TABLE OF CONTENTS

**ABSTRACT** ................................................................................................................................. II

**ACKNOWLEDGMENTS** .................................................................................................................... III

**STATEMENT OF CONTRIBUTIONS** ................................................................................................ IV

**TABLE OF CONTENTS** ........................................................................................................................ V

**LIST OF TABLES** ............................................................................................................................... VI

**LIST OF FIGURES** ............................................................................................................................. VII

**LIST OF APPENDICES** ........................................................................................................................ VIII

**INTRODUCTION** .................................................................................................................................. 1

1. *Genome Evolution in the Brassicaceae* ....................................................................................... 4

2. *Tribe Camelineae: History and Importance* ................................................................................. 11

3. *Camelina microcarpa a Weedy Relative of C. sativa* .............................................................. 16

**MATERIALS AND METHODS** ........................................................................................................ 18

1. *Plant Material and Nucleic Acid Isolation* .................................................................................. 18

2. *Sequencing* ..................................................................................................................................... 19

   Illumina HiSeq ................................................................................................................................. 19

   Illumina MiSeq .................................................................................................................................. 20

   PacBio .............................................................................................................................................. 20

3. *Quality Control and de novo Genome Assembly* ........................................................................ 21

   Illumina and Hybrid Assembly ........................................................................................................ 21

   Correction of PacBio Reads ............................................................................................................. 21

   PacBio de novo Assembly ................................................................................................................ 22

4. *Evaluation of the Assembly* .......................................................................................................... 22

5. *Genetic Distance between C. sativa Subgenomes and C. microcarpa* ..................................... 23

**RESULTS** .......................................................................................................................................... 24

1. *Whole Genome Assembly of Diploid C. microcarpa* ................................................................. 24

2. *Genetic Distance between C. sativa Subgenomes and C. microcarpa* ..................................... 29

**DISCUSSION** .................................................................................................................................. 35

1. *Draft Assembly of Diploid C. microcarpa* .................................................................................... 36

2. *PacBio-based Assembly is Superior to Hybrid-based Assembly* ............................................. 37

3. *The Evolutionary Relationship between C. microcarpa and C. sativa’s Subgenomes* ............. 41

   Genome Changes after Polyploidization ......................................................................................... 41

   Division of Chromosomes into Subgenomes ................................................................................. 42

   Reevaluation of C. sativa Subgenomes ........................................................................................... 43

   Summary and Future Work on Subgenome Structure .................................................................... 49

4. *Insights into Plant Evolution can aid Crop Development* ......................................................... 50

5. *Next Steps* ....................................................................................................................................... 53

**REFERENCES** ................................................................................................................................. 57

**APPENDICES** ................................................................................................................................. 67
LIST OF TABLES

Table 1. Genome size and chromosome counts for the Camelina species ........................................12
Table 2. Coverage of the sequencing data before and after quality control.....................................25
Table 3. Metrics on the three best assemblies using Illumina, Pacbio or both (hybrid) sequencing data for a whole genome assembly of C. microcarpa .................................................................26
LIST OF FIGURES

Figure 1. Phylogeny of the Brassicaceae showing representative species from each lineage ....... 7
Figure 2. The Ancestral crucifer karyotype (ACK) is composed of eight chromosomes and 22 conserved ancestral blocks (A-X)................................................................................................................. 10
Figure 3. Taxon-annotated GC-coverage (TAGC) plot of the C. microcarpa draft assembly ....... 28
Figure 4. Collinearity between the draft C. microcarpa assembly and A. lyrata ......................... 30
Figure 5. Collinearity between the draft C. microcarpa assembly C. sativa ................................. 31
Figure 6. Phylogenetic analysis of C. microcarpa, C. sativa and A. lyrata ................................. 34
Figure 7. Camelina sativa subgenome structure as proposed by Kagale et al. and the structure proposed in this study ................................................................................................................................. 48
LIST OF APPENDICES

Flower picture references for figure 1 and 2 ................................................................. 66
Script used to run Mr. Bayes .......................................................................................... 67
Table A1. Coordinates used to define the 24 Ancestral Crucifer Karyotype (ACK) blocks .... 68
Table A2. The 24 ACK blocks and their placement on the chromosomes of A. lyrata (Alyr) and
the three subgenomes of C. sativa ............................................................................... 69
Table A3. Concatenated sequence length for each Ancestral Crucifer Karyotype (ACK) block
used to construct the phylogenies............................................................................... 70
INTRODUCTION

A key goal in evolutionary biology is to understand the evolution of genomes - how they change in structure and content through time and how this affects the rate of evolution (Otto and Whitton, 2000). Three closely linked processes play major roles in shaping genome structure and the evolution of plants: hybridization (Anderson and Stebbins, 1954; Stebbins, 1959, Abbott et al. 2013), polyploidization (de Wet, 1971; Levin, 1983; Soltis and Soltis, 1999; Husband et al. 2013) and chromosomal rearrangements (Rieseberg, 2001; Rieseberg and Willis, 2007). The sequencing and publication of crop genomes such as maize, canola, soybean, sugarcane and wheat (Schnable et al., 2009; Chalhoub et al., 2014; Schmutz et al., 2010; The International Wheat Genome Sequencing Consortium; 2014) have underscored that these processes occur frequently and have fundamentally shaped plant lineages. As a result, understanding these processes are major areas of interest in plant whole genome-scale studies (Koenig and Weigel, 2015). While there has been less effort focused on the wild diploid relatives of these crops (Michael and VanBuren, 2015), sequencing the extant representative of potential diploid progenitors of these crop species has the potential to allow for the changes resulting from each of these processes to be reconstructed. This then allows for the development of additional resources for crop improvement and weed control. For example, the genome of the emerging oil seed crop Camelina sativa (L.) Crantz (camelina), a hexaploid, has
been sequenced and well described (Kagale et al., 2014), but if the diploid genomes were available, the changes resulting from each of these processes within the genus could be determined. Given that the Brassicaceae family contains the model organism *Arabidopsis thaliana* (L.) Heyhn., important crops such as *Brassica napus* L. (canola), and the well-studied allopolyploids of the triangle of U (U, 1935), there is both a framework for understanding how these processes have shaped *Camelina* and the tools to do so.

Hybridization, defined as the result of crossing between species or, more broadly, between genetically divergent populations (Soltis and Soltis, 2009), can provide a source of adaptations and contribute to speciation (Anderson and Stebbins, 1954; Grant, 1971; Rieseberg, 1997). In fact, one of the key difficulties in understanding plant evolution is that when hybrid speciation occurs the most accurate representation of the phylogenetic history of the group will be a network rather than a tree (Linder and Riesberg, 2004). Hybridization is believed to facilitate adaptation to new or disturbed habitats by allowing for recombination of the characteristics of the parental species and emergence of novel traits (Arnold, 1997). Detection of these events is difficult without genomic resources and there are only few examples that are considered to be documented examples of hybrid speciation (Gross and Rieseberg, 2005), though there are more examples of genetic exchange as a result of hybridization (introgression, reviewed by Ellstrand, 2003). In contrast, allopolyploidization, the combination of hybridization and polyploidization, is much more common (Rieseberg and Willis,
Polyploidy is rare in animals, but some polyploids are found within insects, fish, amphibians and reptiles; whereas polyploidy is widespread amongst plants (Otto, 2007). Polyploidization is the increase of an organism’s genome to include more than two sets of chromosomes through duplication within a species (autopolyploidy) or through hybridization with another species (allopolyploidy) (Grant, 1971, Ramsey and Schemske, 1998). Different mechanisms can give rise to polyploids, which involve anomalous mitotic or meiotic events that result in gametes with an extra set of chromosomes (Comai, 2005). Some effects of polyploidy include increased diversity and heterozygosity, changes in chromatin remodeling, difficulty in meiosis or mitosis, and changes in gene regulation (Comai, 2005). There are potential disadvantages to being a polyploid, but polyploidy still prevails in plants. In fact, many of our crops are polyploids arising from recent polyploidy events (mesopolyploid), and all angiosperms show evidence of ancestral polyploidy (paleopolyploidy) (De Bodt et al., 2005).

Indeed, the proportion of polyploids in angiosperms has been continuously revised upward. Grant (1971) initially estimated 47%, while Masterson (1994) gave a larger estimate of 70%. It was then suggested that there might be a whole genome duplication event that occurred before the angiosperm-gymnosperm divergence, implying that all seeding plants are ancient polyploids (De Bodt et al. 2005, Soltis et al. 2008, Soltis et al. 2009), which was later confirmed by Jiao et al. (2011). Polyploidy is now considered a major force driving speciation,
with estimates showing that 2-4% of angiosperm speciation events may be due to polyploidy (Otto and Whitton, 2000). It is also thought that polyploidy has had an essential role in the rapid diversification of angiosperms (Proost et al., 2011, De Bodt et al., 2005, Soltis et al., 2008, Soltis et al., 2009).

Following polyploidization, genomes often undergo a process of diploidization and genome downsizing, a process that involves chromosomal loss, fusion, translocations and rearrangements. This forms part of the changes resulting from a period of genomic instability where a set of rapid changes occur in the genomes involving genetic and epigenetic mechanisms, gene loss (fractionation), genome restructuring and downsizing, and dominance of one of the genomes (Comai, 2005; Soltis et al. 2016). Other than gene loss, duplicated genes that are retained can receive new functions (neofunctionalization) or the function might get fractionated into the duplicated genes (subfunctionalization) (Renny-Byfield and Wendel, 2014). This process contributes to a return to a diploid-like state and is believed to facilitate stabilization of meiotic behaviour (Feldman and Levy, 2012). More broadly, chromosomal rearrangements may contribute to reproductive isolation between populations, giving the process a role in the formation of species (Riesberg, 2001).

1. **Genome Evolution in the Brassicaceae**

The Brassicaceae, also known as the mustard family or the Cruciferae, is a large plant family comprising 338 genera and 3700 species (Bailey et al., 2006). The crucifers are an
economically and scientifically important plant family. The evolutionary history of the
Brassicaceae has been strongly shaped by the three above described processes involved in
geno\-me evolution (Marhold and Lihova, 2006). For example, there is large variation in
chromosome number suggesting lineages within the family frequently experiences
polypl\-o\-idization and 37% of the species are considered recent polyploids (Warwick and Al-
Shehbaz 2006). In addition, the family has had extensive karyotype evolution and diploidization
as it is clear from the species with lower chromosome number that are derived from species
with higher chromosome number (e.g. A. thaliana) (The Arabidopsis Genome Initiative, 2000;
Lysak et al., 2005). Further, work within the Brassicaceae has made a strong contribution to
understanding genome processes in angiosperms because of the availability of molecular tools
for the group. For example, many advances in our understanding of the role of hybridization
and polypl\-o\-idization in plant evolution have been as a result of whole genome sequencing. In
2015, there were 10 genomes sequenced and 26 other genome projects in progress within the
family (Koenig and Weigel, 2015). The availability of these genomes, representing different
ploidy levels, lifestyles (wild or cultivated), and mating systems (selfing or outcrossing), is
allowing researchers to study genome evolution within the Brassicaceae and ask questions that
were not approachable a few years ago. Indeed, these studies provided evidence of whole
geno\-me duplications (WGD), which gave rise to angiosperms and all extant crucifers (Franzke et
al., 2011, Proost et al., 2011).
Currently the evolutionary history of the Brassicaceae is understood to have included at least three polyploidization events. The number and dates of these ancestral duplications was investigated using expressed sequence tag (EST) data as well as genomic data (Soltis et al., 2009). Bowers et al. (2003) proposed three ancestral polyploidy events, α and probably β are specific to the Brassicaceae, while γ (which may have been a triplication) was found to predate dicot-monocot divergence (Soltis et al. 2009, figure 1). Jiao and collaborators (2011) have also proposed two more ancestral duplications ε and ζ. The ancestral duplication predating the angiosperm-gymnosperm split (ζ) was confirmed with the sequencing of *Amborella trichopoda* Baill., the sister of all angiosperms, which showed evidence of only one duplication ζ, placing it before angiosperm divergence (Amborella Genome Project, 2013). These genome studies provide evidence of the complex history of polyploidy in the Brassicaceae and the role of polyploidization in the evolution of angiosperms.
Figure 1. Phylogeny of the Brassicaceae showing representative species from each lineage. The star denotes the whole genome duplication α that predates the diversification of the Brassicaceae. Phylogeny based on Franzke et al. (2011). Images shown are under the creative commons license (see appendix A).

The *Brassica* genus, from lineage II of Brassicaceae, has been important in the study of polyploidy and hybridization within the Brassicaceae (figure 1) because three different species and their allopolyploid descendants have been known since the 1930s (U, 1935). The relationship between the diploids *Brassica rapa* L. (AA genome), *Brassica nigra* (L.) Koch (BB) and *Brassica oleracea* L. (CC), and the allotetraploid species *Brassica juncea* (L.) Czern. (AABB), *Brassica napus* (AACC), and *Brassica carinata* A. Braun (BBCC) is called the “Triangle of U”. The genomes of one of the allotetraploids, *B. napus*, and the diploid progenitors, *B. rapa* and *B.
*oleracea*, have been sequenced (Chalhoub et al., 2014; The *Brassica rapa* Genome Sequencing Project Consortium, 2011 and Liu et al., 2014; respectively). Sequencing of these genomes has revealed some of the effects of polyploidy such as gene loss (Liu et al., 2014), and rearrangements in the structure of a polyploid’s genome (Cheng et al., 2013). Additional genomes from within the crucifer family will allow for a better understanding of how polyploidization has shaped the evolution of these species and whether these effects are general or lineage specific.

The Brassicaceae have also played a key role in facilitating our understanding of chromosome evolution. In the early 2000’s researchers found large conserved collinear regions or blocks among crucifer genomes (Schranz et al., 2006). In addition, it was hypothesized that the ancestral karyotype of the crucifers, had 8 chromosomes and was more similar to the genomes of *A. lyrata* (L.) O’Kane & Al-Shehbaz and *C. rubella* Reut. than *A. thaliana* (Koch and Kiefer, 2005). Based on this knowledge and genomic information from five *Arabidopsis* species, *C. rubella*, and *Brassica* species, Schranz and collaborators defined an Ancestral Crucifer Karyotype (ACK), composed of 8 chromosomes divided in 24 conserved genomic blocks (2006). This ACK is a theoretical representation of the genome structure of the ancestor from which all crucifers diverged (figure 2). Our understanding of the ACK has been revised using information from 35 other crucifer genomes by Lysak and collaborators to contain 22 conserved blocks (2016) and Murat et al. have provided a high resolution definition of the ancestral karyotypes
In this comprehensive study, the authors defined the genes that are found in the conserved blocks and simplified the blocks to a 16 block system. Describing the genes in the conserved regions provided a resource for comparative studies for the evaluation of genome rearrangements and gene loss (Murat et al., 2015). The description of the ACK has allowed the development of an understanding of the genome rearrangements that have occurred within crucifer species from the ancestral karyotype, including what types of rearrangements are most common, the rate at which they occur, and how they can contribute to a better understanding of the family’s evolutionary history (Lysak et al., 2016). For example, evaluation of the 24 ACK blocks described by Schranz et al. in 2006, facilitated the reconstruction of the subgenomes of *B. rapa* (Cheng et al., 2013) and *C. sativa* (Kagale et al., 2014).
Figure 2. The Ancestral crucifer karyotype (ACK) is composed of eight chromosomes and 22 conserved ancestral blocks (A-X). From lineage I the Arabidopsis thaliana’s genome has experienced rearrangements of the genomic blocks that resulted in the five chromosomes genome, while Capsella rubella and A. lyrata have no rearrangements with respect to the ACK. The ancestor of lineage II plant species is thought to have had a seven chromosome genome in what has been described as the proto-calepinae karyotype. Edited from Lysak et al. 2016
2. Tribe Camelineae: History and Importance

An additional opportunity to examine how genomes evolve as a result of polyploidization and hybridization within the Brassicaceae is presented by the genus *Camelina*, which contains hexaploid oilseed crop *C. sativa* that has a fully sequenced genome (Kagale et al., 2014). In contrast to the well-studied Brassica’s, *C. sativa* is from the Camelineae tribe within Lineage I (figure 1). This lineage also includes: the model plant *A. thaliana*; and the genus *Capsella*, which is a model for mating systems’ evolution (selfing vs outcrossing) (Al-Shehbaz, 2012). Consequently, knowledge acquired from the model plants in lineage I is more easily transferable to *Camelina* species to improve the development of the *C. sativa* crop, than to the *Brassica* crops.

The *Camelina* genus contains eight species (table 1), three of which were introduced in Canada: *C. sativa* (sometimes referred to as ‘false flax’ or ‘gold-of-pleasure’), *Camelina microcarpa* Andrz. Ex DC, and the third *Camelina alyssum* (Miller) Thellung which appears to not have persisted in Canada (Martin et al., 2017). From the species introduced in Canada, *C. microcarpa* is the most abundant in the wild, and *C. sativa* is an ancient oilseed crop receiving renewed interest (Gugel and Falk, 2006). The rest of the species have apparently remained restricted to the likely centre of origin for the genus in the south-eastern Mediterranean (Davis 1965; Al-Shehbaz 1987).
Table 1. Genome size and chromosome counts for the *Camelina* species recognized by Al-Shehbaz (personal communication August 2014), the 2CDNA, cytotype and chromosome counts were obtained from Martin et al. (2017) except the count for *C. alyssum* which was obtained from Warwick and Al-Shehbaz (2006). The genome size was estimated based on the 1 pg = 978Mb estimate by Dolezel and collaborators (2003). Species where no information was available are indicated with a dash (-)

<table>
<thead>
<tr>
<th>Species name</th>
<th>2CDNA (pg)</th>
<th>Genotype size (1C Mb)</th>
<th>Cytotype (base x = 6,7)</th>
<th>Chromosome count (2n)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. alyssum</em> (Mill.) Thell.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40</td>
</tr>
<tr>
<td><em>C. anomal</em> Boiss &amp; Hausskn.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>C. hispida</em> var. <em>grandiflora</em> (Boiss.) Hedge</td>
<td>0.65</td>
<td>318</td>
<td>2x</td>
<td>14</td>
</tr>
<tr>
<td><em>C. hispida</em> var. <em>hispida</em> Boiss.</td>
<td>0.73</td>
<td>357</td>
<td>2x</td>
<td>14</td>
</tr>
<tr>
<td><em>C. hispida</em> var. <em>lasiocarpa</em> (Boiss. &amp; Blanche) Post</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>C. laxa</em> C. A. Mey.</td>
<td>0.56</td>
<td>274</td>
<td>2x</td>
<td>12</td>
</tr>
<tr>
<td><em>C. microcarpa</em> Andrz. ex DC.</td>
<td>0.54</td>
<td>264</td>
<td>2x</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>489</td>
<td>4x</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>1.50</td>
<td>734</td>
<td>6x</td>
<td>40</td>
</tr>
<tr>
<td><em>C. rumelica</em> subsp. <em>rumelica</em> Velen.</td>
<td>1.26</td>
<td>616</td>
<td>4x</td>
<td>26</td>
</tr>
<tr>
<td><em>C. rumelica</em> subsp. <em>transaspica</em> (Fritsch) Hedge</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>C. sativa</em> (L.) Crantz</td>
<td>1.54</td>
<td>753</td>
<td>6x</td>
<td>40</td>
</tr>
</tbody>
</table>

Archeological data shows that *C. sativa* has been cultivated since the Neolithic (Hovsepyan and Willcox, 2008). *Camelina sativa* oil was used by the Romans as lamp oil. In Europe, it continued to be grown as an agricultural crop for lamp fuel and cooking oil, until World War II after which cultivation of *C. sativa* declined as it was replaced by canola (Francis and Warwick, 2009). *Camelina* species may have been introduced into North America as a contaminant of flax, but there are also reports from Manitoba of *C. sativa*’s deliberate
introduction into Canada (Francis and Warwick, 2009). *Camelina sativa* is an oilseed crop with many potential applications in the food, cosmetic and biofuel industries (Bansal and Durrett, 2016). Development of *C. sativa* is in the early stages, with research and field trials expected in the following years for *C. sativa* oil to meet industry standards. However, the re-emergence of *C. sativa* as a crop is expected due to the search for alternatives to meet the challenges associated with increasing world population and diminishing resources, including sustainable production of high value products such as jet fuel and biolubricants (Bansal and Durett, 2016), or for renewable production of essential long chain fatty acids such as DHA and EPA (Ruiz-Lopez et al., 2014).

Drought, cold and disease resistance, low water input, and ability to grow on marginal lands makes *C. sativa* an excellent candidate oilseed crop in current climate conditions and with limited arable land (Gugel and Falk, 2006). Since *C. sativa* oil has an exceptionally high omega-3 content, a promising application for *C. sativa* is in the fish oil industry where it could serve as a renewable alternative for fish oil. There is high demand for fish oil due to the beneficial effects of omega-3 fatty acids in preventing cardiovascular and other diseases, which is placing large pressure on this resource causing a decline in fish stocks. Production of fish oils requires feeding fish with fish meal or fish oil and efforts to replace fish oil in fish feed with vegetable oil have been unsuccessful (Lenihan-Geels et al., 2013). Recently, genetically modified *C. sativa* rich in
EPA and DHA was shown to successfully replace fish oil in fish feed, which is a step towards making fish oil production more sustainable (Betancor et al., 2016).

*Camelina sativa* is also suitable for sustainable production of biofuels and high-value industrial oils (Reaney and Furtan, 2006; Shonnard et al., 2010; Liu et al., 2015; Snapp et al., 2014) and its oil or seed meal used in animal feed and cosmetic products (Aziza et al., 2010; Francis and Warwick, 2009). *Camelina sativa* is being developed in Canada and worldwide, however oil yield and content needs to be substantially improved for commercial success (Lu and Kang, 2009). The sequencing of the *C. sativa* genome and transcriptome have provided key information for the improvement of *C. sativa* oil (Kagale et al. 2014, Kagale et al. 2016).

The *Camelina* genus contains three known polyploids: *C. sativa*, *C. rumelica* and *C. microcarpa* (table 1). Unfortunately, *Camelina* species have been overlooked as one of the important polyploids in the Brassicaceae. For example, Franzke and collaborators ignore *Camelina* spp. when reviewing the whole genome duplications in the Brassicaceae (Franzke et al., 2011). Sequencing of hexaploid *C. sativa* (2n=40) was the first step to understand the polyploidy events that occurred within the genus. *Camelina sativa* was sequenced by a hybrid Illumina and Roche 454 sequencing approach and was completed to a reference genome using a genetic linkage map (Kagale et al., 2014). Kagale et al. (2014) suggested that two allopolyploidy events gave rise to hexaploid *C. sativa*, but the diploid progenitors that hybridized to give rise to *C. sativa* are unknown. The diploid ancestor of the three parental
genomes of *C. sativa* is estimated to have diverged from *A. thaliana* ~17 million years ago (Mya), and the three parental genomes are thought to have diverged from each other ~5.4 Mya (Kagale et al., 2014). The dates of the hybridizations are unknown, but Kagale and collaborators suggest that they occurred between 5-10,000 years ago.

Kagale and collaborators reconstructed the three *C. sativa* subgenomes (2014) based on synteny with *A. lyrata*, and performed a phylogenetic analysis using expressed sequence tag (EST) data to understand the evolutionary relationship between them. From these analyses they suggest that an initial allopolyploidy event occurred between a 7 chromosome species (subgenome 2) and a 6 chromosome species (subgenome 1), since these were the most closely related. A later hybridization occurred with a 7 chromosome containing species, represented in subgenome 3. Interestingly, the three subgenomes of *C. sativa* were found to have retained a large proportion of their genes and have little rearrangements, suggesting that the two polyploidy events occurred recently and with little time in between them (Kagale et al. 2014).

The *C. sativa* genome shows a conserved ACK structure in each of the three subgenomes (figure 2), with only a few chromosomal rearrangements and in-block breaks. Overall, subgenome 3 has experienced the most rearrangements, followed by subgenome 2. Most ancestral chromosomes are found unaltered in *C. sativa* except for the fusion that occurred between the ancestral chromosomes 2 and 4. This fusion is what resulted in the reduction from an 8 chromosome ancestor, to the 7 chromosomes in subgenomes 2 and 3 of *C. sativa*. In
addition to the 2/4 ancestral chromosome fusion, subgenome 1 experienced the merging of ancestral chromosomes 7 and 8, resulting in the 6 chromosome genome. Within the chromosomes, 20 of the 24 ACK blocks are found almost intact, except for D, E, I, J (Kagale et al. 2014). In total, Lysak et al. (2016) report 21 ACK in-block breaks in the *C. sativa* genome. These rearrangements are believed to have primarily occurred in diploid progenitors of *C. sativa* (Lysak et al., 2016).

The phylogenetic analysis by Kagale et al. using EST data, also included other *Camelina* species: *C. laxa, C. hispida* and *C. rumelica*. However, *C. microcarpa* an important member of the genus, which is also a polyploid that was not included in the analysis. Studies including *C. microcarpa* are necessary to get a better understanding of polyploidy within the genus.

3. *Camelina microcarpa* a Weedy Relative of *C. sativa*

The most abundant wild relative of *C. sativa* in Canada is *C. microcarpa* or littlepod false flax. There is very little literature on *C. microcarpa*, a search on PubMed returns only four papers compared to 125 for *C. sativa* (March, 2017). *C. microcarpa* is a naturally weedy species, and a potential descendant of one of the diploid ancestors of *C. sativa*, but the extent of their evolutionary relationship is unexplored. *C. microcarpa* is currently listed by the USDA (United States Department of Agriculture) as a potentially weedy or invasive species. *Camelina microcarpa* and *C. sativa* are very similar morphologically (Francis and Warwick, 2009), generally *C. microcarpa* is taller, has smaller seeds and requires vernalization compared to *C.
sativa. To date three different cytotypes of C. microcarpa have been identified, a diploid (2n=12), tetraploid (4n=26) and hexaploid (6n=40) (Martin et al., 2017). Forced crosses were performed to evaluate the fitness of hybrids between hexaploid C. microcarpa and C. sativa (Martin et al., unpublished) and we determined there is near perfect sexual compatibility between hexaploid C. microcarpa and C. sativa (96%). Hybrids were morphologically similar to C. microcarpa but have reduced pollen viability and seed output. Release of transgenic C. sativa lines should consider the possibility of transgene escape to C. microcarpa wild populations. In addition, crosses between C. sativa and C. microcarpa could be used to introduce variability into the C. sativa crop.

The Camelina genus contains three known extant diploid species: Camelina laxa C. A. Mey., Camelina hispida Boiss. and C. microcarpa (table 1). Their small genomes make them tractable for determining not only whether they are descendants of the diploids that contributed to the synthesis of hexaploid C. sativa genome, but to better understand how a diploid genome changes following polyploidization. Camelina microcarpa provides an excellent opportunity to study whole genome evolution from diploidy to hexaploidy, since there are also tetraploid and hexaploid cytotypes of this wild species (Martin et al., 2017). This opportunity would complement work on polyploids in Brassica as Camelina belongs to a separate evolutionary lineage within the Brassicaceae.

Here I make the first step toward the goal of extending our understanding of genome
evolution using the opportunity available within *Camelina* by assembling a draft sequence for a
diploid wild relative of *C. sativa, C. microcarpa*, and examining its relationship with *C. sativa*.
This study will also provide genomic information valuable for the development of *C. sativa* as a
viable Canadian oilseed crop. To obtain the draft genome, I have used data from different Next
Generation Sequencing (NGS) platforms: Illumina paired-end, mate-pair and PacBio. This draft
genome was used to answer the following question: What subgenome of *C. sativa* is most
closely related to diploid *C. microcarpa*?

**MATERIALS AND METHODS**

1. **Plant Material and Nucleic Acid Isolation**

   *C. microcarpa* seeds were obtained from the North Central Regional Plant Introduction
   Station (NCRPIS) and were originally collected from Lozere, France (PI650135). These individuals
   were morphologically identified as *C. microcarpa* by I.A. Al-Shehbaz (Missouri Botanical
   Garden). Original seeds obtained from NCRPIS were stratified in petri dishes using filter paper
   that was moistened with 0.2% KNO$_3$, sealed with Parafilm (Pechiney Plastic Packaging
   Company, Illinois, USA), and placed at 4°C in the dark for 2 weeks. For seed germination, the
   plates were then placed at room temperature under growth lights with a 16 h/8 h day/night
   light cycle. Seedlings were then sown on soil (soil, peat, and sand; 1:2:1) in 2X3 in, 2 in deep
   trays, and placed in growth chambers with a photoperiod of 16 h 20°C days/8 h 18°C nights.
   After 6 weeks, the plants were placed for another 6 weeks at 4°C with an 8 h/16 h day/night
photoperiod for vernalization. Plants were then transplanted to 5 in pots with the same soil and allowed to self-pollinate and set seed in 16 h/8 h day/night photoperiod with 20°C days and 18°C nights. Following approximately 3 months, the mature seeds were collected and this process was repeated to obtain a fifth-generation inbred line.

Young leaves from the rosette stage of this inbred line were used for DNA extraction. Total DNA was extracted using a FastDNA Spin Kit (MP Biomedicals, Solon, OH), grinding was done in the FastPrep (MP Biomedical) at 4.0 (no units) for 20 s, with the addition of one ceramic bead. Two DNA extractions were pooled for a total volume of 200 μl and precipitated with the addition of 20 μl 3 M NaOAc and 200 μl 100% ethanol. Following an overnight incubation at -20°C, the DNA was centrifuged at 13,000 rpm at 4°C for 30 min. The ethanol was decanted and the DNA pellet was washed with cold 70% ethanol, dried at 37°C for approximately 20 min and resuspended in 100 μl 5 mM Tris-HCl (pH 8.5). The DNA concentration was determined by Qubit was 110 ng/μl. DNA quality was determined by running 1.0 μl on a 0.8% E-gel (Invitrogen by ThermoFisher) beside a 0.2 μg of 20kb ladder (GeneRuler 1kbPlus, ThermoFisher).

2. Sequencing

*Illumina HiSeq*

From one individual plant, DNA was sent to The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada for sequencing. Four libraries were prepared: one
paired-end (PE) library with 550 bp inserts using a Nano kit (Illumina, San Diego, CA, USA) and three mate-pair (MP) libraries: 3kb, 5kb and 10kb inserts using the Nextera mate-pair kit (Illumina). All four libraries were sequenced on an Illumina HiSeq-2000 using v4 chemistry and flow cells with runs of 2 × 126 bases.

*Illumina MiSeq*

From a second individual plant, DNA was sent to the sequencing facility at the Microbial Molecular Technologies Laboratory (MMTL) in Ottawa (Ottawa Research and Development Centre, Agriculture and Agri-Food Canada) for paired-end sequencing. A single library was prepared using the Ovation Ultralow Library Systems (NuGEN, San Carlos, CA, USA), with 500bp inserts, and sequenced using Illumina MiSeq v3 chemistry (Illumina) with runs of 2 x 300 bases.

*PacBio*

From a third plant, DNA was sent to the McGill University Genome Quebec Innovation Centre for PacBio sequencing using P6-C4 chemistry (Pacific Biosciences, Menlo Park, CA, USA). Since DNA was slightly degraded, the library was prepared without shearing. A total of 7 Single Molecule Real-Time (SMRT) cells were used for sequencing. A subset of the PacBio reads obtained were circular consensus sequencing (CCS), which is the consensus sequence from reads where DNA polymerase has done at least two full passes on the read. All other reads were regular reads: continuous long reads (CLR).
3. Quality Control and \textit{de novo} Genome Assembly

\textit{Illumina and Hybrid Assembly}

A conservative quality control was performed following Zhou and Rokas' approach (2014). Mate-pair data was processed using nextclip v1.3.1 (Leggett et al., 2014) to remove adapters and de-duplicate reads. The minimum read length to keep a read after the adapter was removed was set to 25 bases. Trimmomatic 0.33 (Bolger et al. 2014) was used for quality based trimming of all the MP and PE reads. A sliding window trimming approach was done by trimming all four bases that had an average quality lower than 15 (phred score), and reads shorter than 36 bases were filtered out. Musket 1.1 (Liu et al. 2013), which takes advantage of the high coverage to perform substitution-error correction, was used to correct MP and PE reads.

An Illumina-only assembly and a hybrid assembly (with the PacBio raw reads) were performed using SPAdes v3.7.0 (Bankevich et al., 2012).

\textit{Correction of PacBio Reads}

PacBio reads were corrected with the Illumina PE data, using proovread 2.13.11 (Hackl et al., 2014). The PE HiSeq data used was processed as described in the previous section. For the PE MiSeq data there was a 100 bp overlap since the fragment length was 500 bp and the forward and reverse reads were 300 bp long. Following the recommendation from the proovread author, the PE MiSeq reads were processed by correcting with musket (see previous
section), overlapped using FLASH v1.2.11 (Magoc and Salzberg, 2011), and then trimmed with trimmomatic (see previous section). Processed PacBio reads were used for scaffolding and gap closing of the PacBio assembly.

**PacBio de novo Assembly**

Raw PacBio reads were self-corrected and assembled using Canu v1.3 (Koren et al, 2016), with the error rate set to 2.5%. The contigs were scaffolded using SSPACE Long Read v1-1, using corrected PacBio reads (Boetzer et al., 2011). Then, PBJelly was used for gap closure and finishing of the scaffolds using corrected PacBio reads (English, 2012). To further improve the assembly, a second step of scaffolding was performed with all processed Illumina data using SSPACE v3.0. Gap closure with PBJelly was repeated after scaffolding.

The quality of the assembly was evaluated using QUAST 2.2 to obtain the assembly metrics (Gurevich et al., 2013). The completion of the assembly’s gene space was evaluated using CEGMA 2.5 to identify conserved eukaryotic genes (Parra et al., 2007). In addition GMAP version 2011-03-28.v3 (Wu and Watanabe, 2005), was used to align the leaf transcriptome of *C. sativa* (Liang et al., 2013) and estimate the proportion of *C. sativa*-like genes in the assembly. The transcriptome was obtained from NCBI (version GABO0000000.1)

4. **Evaluation of the Assembly**

After evaluating the assembly metrics and the completeness of the gene space, the best assembly was selected for further analysis. To evaluate if there was contamination in the
assembly, a taxon-annotated GC-coverage (TAGC) plot was generated using blobtools v0.9.19.5 (Kumar et al., 2013). For the TAGC plot, the coverage of each scaffold was obtained by aligning all raw PacBio reads to the PacBio-only draft assembly using bwa v.0.7.10-r789 (Li and Durbin, 2009). To annotate the scaffolds for the TAGC, a megablast search was done using BLASTn 2.26.0+ against the nt database (downloaded on January 26, 2017), and the expected value was set to 1e-25.

To evaluate collinearity between the C. microcarpa draft genome and both C. sativa and A. lyrata as a reference, the MUMmer 3.23 toolset was used to align the genomes, filter the alignment results and plot the alignment (Kurtz et al., 2004). The C. sativa genome was obtained from NCBI (version JFZQ0000000.1) and the A. lyrata genome was obtained from Ensembl Genomes (version 1.0). First, the scaffolds from C. microcarpa were ordered with respect to the reference genome using mauve 2.4.0 (Rissman et al., 2009). Then, from the MUMmer toolset Nucmer was used for the alignment. Using the delta-filter tool, the alignment was filtered to keep only the longest alignments for each query. Mummerplot was used to plot the alignments.

5. Genetic Distance between C. sativa Subgenomes and C. microcarpa

For the estimation of genetic distance between C. microcarpa and the three subgenomes of C. sativa, we first choose homologous regions between the genomes by finding collinear regions to each ACK block as described in the A. lyrata genome (appendix table A1).
Pairwise collinearity was computed between *A. lyrata* and *C. microcarpa* and each of the three subgenomes of *C. sativa* using MUMmer (see section above). Then, an in-house R script was used to find the collinear fragments that overlapped between all genomes within the 24 blocks. For a fragment to be included in the analysis it had to: 1) be at least 1000bp long (before trimming), 2) pass the alignment test (based on pairwise genetic distance) and 3) be found in the expected group of *C. sativa* chromosomes (appendix table A2). The alignment test was implemented to filter out alignments that had excessive gaps or where one or more genomes did not align well. For each ACK block, 30 fragments were chosen at random to perform an alignment with MAFFT v7.164b (Katoh and Standley, 2013), then trimmed and concatenated by gblock v0.91b (Castresana, 2000). Using the concatenated blocks, phylogenetic trees were estimated by Bayesian Inference with MrBayes v3.2.1 (Ronquist et al., 2012) following recommendations by Hall (2011). The analysis was run for a maximum of 10,000,000 generations, using 4 independent chains, the frequency of chain swapping set to 0.2 and the analysis was stopped after reaching 0.01 convergence (appendix B).

Unless otherwise specified all tools were run using default settings.

**RESULTS**

1. **Whole Genome Assembly of Diploid *C. microcarpa***

*Camelina microcarpa* plants were inbred for five generations to reduce heterozygosity to 3.1% of the original levels, genomic DNA was extracted and sequenced using Illumina and
PacBio technologies. The genome was sequenced to a total 190x raw coverage, and a final 86x coverage after quality control (table 2). Quality control was done directly by the canu assembler for PacBio data, while separate tools (nextclip, trimmomatic and musket) were used for quality control of the Illumina data. Mate-pair data from the Illumina technology was of particularly low quality, and thus most data were filtered out in the quality control steps (table 2).

Table 2. Input paired-end (PE), mate-pair (MP), PacBio circular consensus sequencing (CCS) and PacBio continuous long reads (CLR) libraries, showing the corresponding coverage before and after quality control (QC). For PacBio reads, the average read length is shown (*).

<table>
<thead>
<tr>
<th>Library</th>
<th>Insert Size (bp)</th>
<th>Read length (bp)</th>
<th>Raw coverage (X)</th>
<th>Coverage after QC (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE</td>
<td>225</td>
<td>126</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>PE</td>
<td>500</td>
<td>300</td>
<td>33</td>
<td>23</td>
</tr>
<tr>
<td>MP</td>
<td>3,000</td>
<td>126</td>
<td>35</td>
<td>7</td>
</tr>
<tr>
<td>MP</td>
<td>5,000</td>
<td>126</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>MP</td>
<td>10,000</td>
<td>126</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>PacBio-CCS</td>
<td>-</td>
<td>8556*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PacBio-CLR</td>
<td>-</td>
<td>7271*</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>-</strong></td>
<td><strong>190</strong></td>
<td><strong>86</strong></td>
<td></td>
</tr>
</tbody>
</table>

To understand which sequencing technology would produce the best assembly, the data was assembled by three different approaches: Pacbio-only using canu, Illumina-only and hybrid assembly using SPAdes (table 3). The PacBio-only approach produced a much superior assembly, which was evident by the much higher N50 (373 kbp), largest contig (2,329 kbp), and much smaller number of scaffolds (1036). The GC content of the assemblies, 36.4-36.8%, was
comparable to *C. sativa*’s GC content of 34.0% (Kagale et al., 2014). Since the PacBio assembly was superior, scaffolding and gap closing were performed on this assembly. The assembly was scaffolded using the PacBio data first, and then using the Illumina data in SSPACE. Each round of scaffolding was followed by gap closing and genome finishing using PBjelly. The final assembly covered 72% of the genome, in 269 scaffolds with an N50 of 1,231kbp and the largest scaffold was 4,810kbp long (table 3). This final assembly was used as the draft assembly of *C. microcarpa* and was further evaluated for contamination, the completion of the gene space, and compared with the *C. sativa* genome.

Table 3. Metrics obtained with QUAST on the three best assemblies using Illumina, Pacbio or both (hybrid) sequencing data for a whole genome assembly of *C. microcarpa* (genome size estimated at 264Mb, table 1). SPAdes (Illumina and hybrid) and canu (PacBio) were used for the assembly, scaffolds smaller than 500bp were filtered out. The final assembly was obtained after scaffolding the PacBio assembly with SSPACE and gap closing with PBjelly

<table>
<thead>
<tr>
<th>Sequencing</th>
<th>Illumina</th>
<th>Hybrid</th>
<th>PacBio</th>
<th>Final Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>#scaffolds</td>
<td>27,748</td>
<td>21,427</td>
<td>1036</td>
<td>269</td>
</tr>
<tr>
<td>Coverage (%)</td>
<td>73</td>
<td>74</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>Largest scaffold (kbp)</td>
<td>858</td>
<td>1,111</td>
<td>2,329</td>
<td>4,810</td>
</tr>
<tr>
<td>N50 (kbp)</td>
<td>81</td>
<td>118</td>
<td>373</td>
<td>1,231</td>
</tr>
<tr>
<td>L50</td>
<td>561</td>
<td>403</td>
<td>159</td>
<td>47</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>36.8</td>
<td>36.7</td>
<td>36.4</td>
<td>36.4</td>
</tr>
<tr>
<td>#N’s per 100kbp</td>
<td>242</td>
<td>145</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td><em>C. sativa</em> orthologous genes (%)</td>
<td>97.4</td>
<td>97.5</td>
<td>96.9</td>
<td>97.0</td>
</tr>
<tr>
<td>CEGMA (% out of 248)</td>
<td>97.6</td>
<td>99.2</td>
<td>95.6</td>
<td>96.8</td>
</tr>
<tr>
<td>CEGMA (% out of 458)</td>
<td>99.3</td>
<td>99.3</td>
<td>98.7</td>
<td>98.7</td>
</tr>
</tbody>
</table>

N50: the contig length such that the sum of all the contigs of that size or larger add to 50% of the total length of the assembly

L50: the number of contigs that are of the N50 length or larger
CEGMA: Core Eukaryotic Genes Mapping Approach

The completeness of the assemblies was evaluated by estimating the proportion of core conserved eukaryotic genes and the proportion of *C. sativa* orthologous genes in the assembly. CEGMA was used to estimate the proportion of ultra-conserved and conserved Core Eukaryotic Genes (CEG) in the assembly. From a total of 248 ultra-conserved CEG, 99.1% were identified in the assembly, from which only 2.42% were incomplete. In addition 98.7% of the 458 conserved CEG were identified in the assembly. An alignment of the *C. sativa* transcriptome to the draft assembly using GMAP showed that 97% of these *C. sativa* genes were present as orthologs in the assembly (54,034 total).

To evaluate if there was contamination in the PacBio assembly, the scaffolds were plotted in a Taxon-annotated GC-coverage (TAGC) plot by blobtools (figure 3). Here, each scaffold is represented as a circle with a diameter proportional to its length, the scaffolds are plotted by their GC proportion (x-axis) and coverage (y-axis). Coverage was estimated by aligning all the reads used for the assembly to the draft assembly. Each scaffold was annotated by species according to a BLASTn search to the non-redundant nucleotide (nr) database. Most of the reads mapped to the assembly (95.55%, figure 3B), and were annotated to close relative plant species (*A. thaliana, C. sativa* and *A. lyrata*). According to the GC proportion and coverage, there appears to be no exogenous DNA contamination in the draft assembly.
Figure 3. Taxon-annotated GC-coverage (TAGC) plot of the *C. microcarpa* draft assembly. A) Each scaffold is plotted as a circle according to its GC proportion and coverage, the diameter is proportional to the length of the scaffold. The scaffolds were annotated by species by a BLASTn search against the nr database, the colours are explained in the legend, which has the scaffold-count, scaffold span and scaffold N50. The top and right histograms show the sum GC proportion and coverage (respectively). B) The proportion of mapped reads to the assembly, used to estimate coverage (left) and the proportion of scaffolds by species (right).
2. Genetic Distance between *C. sativa* Subgenomes and *C. microcarpa*

The *C. microcarpa* assembly was aligned with the *A. lyrata* and *C. sativa* reference genomes to identify collinear regions (figures 4 and 5). The scaffolds were ordered by mauve to follow the same order as the reference genome. The whole genome plots show that *C. microcarpa* aligns once to *A. lyrata*, but three times to the hexaploid *C. sativa* as expected. When evaluating more closely the largest scaffolds, large collinear regions were observed with *C. sativa* (figure 5, bottom). In addition, *C. microcarpa* was found to be more collinear with most of subgenome 2, except for chromosomes 6, 10 and 18, where *C. microcarpa* was more collinear to chromosomes 11 and 4 from subgenome 1.
Figure 4. Collinearity between the draft *C. microcarpa* assembly (269 scaffolds) and the 8 chromosomes of *A. lyrata* by MUMmer. The lines represent collinear regions between the two genomes, and the colours indicate whether the match is found in the forward (red) or reverse (blue) direction. Scaffolds were ordered with mauve, using *A. lyrata* as a reference genome.
Figure 5. Collinearity between the draft *C. microcarpa* assembly and the 20 chromosomes of *C. sativa*, all 269 scaffolds (A) and the 10 longest scaffolds (B) were aligned against the three subgenomes of *C. sativa* (CsatG1-3) and plotted using by MUMmer. The lines represent collinear regions between the two genomes, and the colours indicate whether the match is found in the forward (red) or reverse (blue) direction. Scaffolds were ordered with mauve, *C. sativa* as a reference genome. The most collinear regions between the two genomes are circled.
Using the draft assembly of *C. microcarpa* the phylogenetic relationship between *C. microcarpa* and the chromosomes of *C. sativa* was evaluated. First, collinear regions between each ACK block of *A. lyrata* and the *C. sativa* genomes (CsatG1-G3) were identified. Only regions that overlapped between all of the genomes were kept. From all the fragments identified, the aim was to extract 30 fragments at random per block, align and concatenate them to get approximately 50,000bp long blocks to construct a phylogeny. However, for some ACK blocks very few overlapping fragments were identified (O) and for some blocks (I, K) no fragments were identified (appendix table A3).

The concatenated alignments were used to construct phylogenetic trees for each block by Bayesian Inference (figure 6). Within *C. sativa*, chromosomes assigned to CsatG1 were found to be most closely related to chromosomes assigned to CsatG2, except for the blocks from chromosome 17 (A, B, C). For these blocks chromosome 14 (assigned to CsatG2) was most closely related to chromosome 3 (assigned to CsatG3), instead being more closely related to chromosome 17 (assigned to CsatG1). This suggests that chromosome 3 was erroneously assigned to CsatG3 and should rather be assigned to CsatG1.

The *C. sativa* chromosomes that were most collinear with *C. microcarpa* were also found to be the most closely related in the phylogenetic analysis (figure 5, 6). In addition, *C. microcarpa* was found to be most closely related to chromosomes assigned to *C. sativa* subgenome 2 (CsatG2), except for the following cases. **Case 1:** chromosome 4 from CsatG1 is
more closely related to \textit{C. microcarpa} than chromosome 6 from CsatG2. This was evident from the phylogenies produced from blocks L, M, and N. \textbf{Case 2}, chromosome 11 from CsatG1 is more closely related to \textit{C. microcarpa} than chromosomes 10 and 18 from CsatG2. This was observed in the trees from T, U, W, and X. However, within \textbf{case 2}, blocks V and S provided contradictory evidence as represented by phylogenies with different topologies. Specifically, \textit{C. microcarpa}'s block V, with only 4 fragments and a short alignment, was more closely related to subgenome 3; and block S the phylogeny indicated that subgenomes 1 and 2 were more closely related to each other than to \textit{C. microcarpa} D, J, O, P, S, V

Contrary to the observations described above, blocks D, G, J, O, P, S and V showed phylogenetic relationships that were different to the rest of the chromosome. However, from these, D, J, O, P and V/ were produced from shorter alignments with 6,112, 12,470, 817, and 6,342 bp, respectively.
Figure 6. Phylogenetic analysis of *C. microcarpa* (Cmic), the chromosomes of *C. sativa* (Csat1-20) from each subgenome (G1-3) and *A. lyrata* (Alyr) as an outgroup. The consensus Bayesian tree for each conserved ancestral block (A-X) was constructed from an alignment of the length shown in the scale. Trees were constructed under the GTR+gamma model. Posterior probabilities for each of the nodes was 100%, except for one node that received 89% support (*). The colors correspond to the chromosomes from the Ancestral Crucifer Karyotype (ACK), which are divided into 24 conserved ancestral blocks (A-X, bottom right).
DISCUSSION

Whole genome sequencing is allowing a deeper understanding of the processes that have shaped evolution, such as polyploidy, hybridization and genome rearrangements. Polyploidy is recognized as a fundamental evolutionary mechanism in eukaryotes (Soltis et al., 2014; Husband et al., 2013). Although largely unexplored, polyploidy is known to occur in fungi and to a lesser extent polyploidy is found in animals (Albertin and Marullo, 2012). All flowering plants are recognized to be polyploids and many have undergone several rounds of polyploidy (De Bodt et al., 2005), thus studying polyploidy is at the core of understanding plant evolution. The sequencing of plant genomes has unveiled some of the mechanisms that follow polyploidy and the effects it has on genome structure and gene function (Renny-Byfield and Wendel, 2014). Particularly, the Brassicaceae has become a platform to study polyploidy given the large number of allopolyploids and the sequencing effort for the crops in this family. Further, the description of the ancestral crucifer karyotype and genomic blocks have facilitated whole genome comparative studies (Schranz et al., 2006; Murat et al., 2015; Lysak et al., 2016).

The Brassicas from lineage II (Cheng et al., 2013), and Capsella from lineage I are well studied polyploids (Douglas et al., 2015). In lineage I the hexaploid C. sativa, an emerging oilseed crop, has a sequenced genome (Kagale et al., 2014). Analysis of the hexaploid genome of C. sativa can allow us to expand our knowledge beyond current models in the Brassicaceae (Lysak et al. 2016). However, the parental diploid species that gave rise to C. sativa are
unknown. The three known extant diploid species in the genus, *C. laxa*, *C. hispida* and *C. microcarpa* are potential descendants of *C. sativa*’s ancestors, but the genomic data to investigate this possibility is lacking. From these three candidates *C. microcarpa* provides an excellent study system for polyploidy, as there are diploid, tetraploid and hexaploid cytotypes of this species. The first step to understand polyploidy in the *Camelina* genus is to sequence diploid *C. microcarpa* and evaluate its evolutionary relationship with *C. sativa*.

1. **Draft Assembly of Diploid *C. microcarpa***

   Plant genomes can be very complex, they range from small, 61 Mb in *Genlisea tuberosa* Rivadavia, Gonella & A.Fleischm., to over 150,000 Mb in *Paris japonica* Franch. (Michael and VanBuren, 2015). They can also have a complex history of polyploidy and hybridization and can be highly repetitive. For example, *Saccharum officinarum* L. (sugarcane) is a decaploid and has a 12 Gb genome, which is 80% repetitive (Michael and VanBuren, 2015). Advances in sequencing technology have made sequencing plant genomes possible. However, there isn’t a single approach to sequence all plant genomes, since different plant species present different challenges. The *Camelina microcarpa* accession sequenced here is diploid, has a relatively small genome (275Mb), and was inbred for five generations. Sequencing was done with Illumina and PacBio technologies (table 2), which require different considerations in terms of quality of the raw data and the final assembly.
Illumina data and PacBio data have different quality limitations. PacBio is known to have low base call quality with a 13% expected error, compared to 0.1% error with Illumina HiSeq (Rhoads and Au, 2015). In turn, Illumina is known to suffer from bias and duplication errors (Zhou and Rokas, 2014). There are a great number of quality control tools and approaches used to process raw Illumina data and deal with adapters, low quality bases, duplication and library contamination. For the quality control of the Illumina data, I followed a conservative approach based on the considerations from Zhou and Rokas (2014). Quality control tools for PacBio data are less common. These use Illumina data to correct reads, or use self-correction modules that make use of the high coverage to correct bases. After comparing the assemblies that resulted from correcting with proovread, a standalone error correction tool, and the error correction module from canu, I chose to correct the PacBio reads using the module from the assembler with an error rate of 2.5%. Sequencing of *C. microcarpa* was done to 190X raw coverage, which reduced to 86X after quality control (table 2). The 104X lost during QC was uniquely from the Illumina data, since error correction used in PacBio does not filter out data.

2. **PacBio-based Assembly is Superior to Hybrid-based Assembly**

The PacBio-only approach produced the best assembly, which was in 1036 scaffolds and had a 373 MB N50, while the Illumina-only assembly was fragmented into 27,748 scaffolds and the hybrid assembly was in 21,427 scaffolds (table 3). Given the lower error rate of the Illumina sequencer, it is possible that the Illumina assembly had a more accurate base call. However, the
PacBio assembly was chosen for the rest of the study because a more contiguous assembly was preferred for whole genome comparative analyses. Future work with this data should include assessing the quality of the assembly using the Illumina data.

The PacBio assembly was scaffolded and gap-closed for a final assembly of 269 scaffolds, 1,231 Mb N50 and total length of 197 Mb (table 3). This final assembly covered a total of 72% of the estimated genome length, had 97% of C. sativa’s gene orthologs, and 99% of the CEGMA ultra conserved eukaryotic genes (out of 248), out of which 97% were found to be complete genes. The assembly was also found to be contamination free and sequences were found to be homologous to C. sativa and Arabidopsis spp. (figure 3). In addition, 95.55% of the raw reads were successfully aligned to the draft assembly, the rest of the reads may have been excluded by the assembler due to low quality (figure 3B). The draft assembly of C. microcarpa presented here represents a highly contiguous assembly that likely contains most of the coding regions of the genome.

Other than the gene space, duplications and transposable elements are key genomic regions that should be evaluated after assembling a genome. Repetitive sequences were not annotated in this C. microcarpa assembly. However, it is expected to be near 30% of the genome, since C. sativa was found to be 28% repetitive which is comparable to A. thaliana (24%). Segmental duplications, large duplication across the genome, occur in plant genomes and have been observed in A. thaliana (The Arabidopsis Genome Initiative, 2000). Some
segmental duplications were observed in the small collinear regions between *C. microcarpa* and the *C. sativa* genomes. Other duplications might not be visible due to low percent identity given that sequences diverge after duplication (figure 4, 5). Transposable elements play a role in mediating genome rearrangements and epigenetic changes (Lysak et al., 2016; Renny-Byfield and Wendel, 2014). Future annotation of the transposable elements in the *C. microcarpa* genome may reveal the role of transposable elements in the evolution of this genome.

Long reads are key to assemble plant genomes, as they can span long repetitive regions that would otherwise cause fragmentation of the assembly (Todd and VanBuren, 2015). The advantage of PacBio sequencing is that the reads are much longer than those obtained from Illumina sequencing. PacBio reads were in average 7,000 bases long, while Illumina reads were 126-300 bases long. In exchange for long reads, PacBio sequencing can be more than 10-fold more expensive (per base) than Illumina HiSeq sequencing (Rhoads and Au, 2015). As an alternative to long reads, Illumina developed mate pairs, which are long fragments of DNA that get circularized and the ends are sequenced so that the distance between the reads is known (Illumina, 2012). However, in this study the mate pair reads (3, 5, 10 kb inserts) were found to be low quality, with almost 90% of the data filtered out during the quality control step. Low data quality could be explained by low quality DNA used for sequencing. However, PacBio sequencing was successful using DNA that was extracted using the same protocol. Even though Illumina had a lower cost than PacBio sequencing, the loss of data during quality control and
the poor assembly that resulted from using these data made Illumina sequencing not cost-effective for the assembly of diploid *C. microcarpa*.

The lack of long reads might explain why the Illumina-only assembly was highly fragmented, suggesting that PacBio reads may have been able to help assemble repetitive regions. Surprisingly the hybrid assembly which included the PacBio reads was also highly fragmented (table 3). An important consideration is that two different assemblers were used; SPAdes was used for the Illumina-only and hybrid assemblies and canu was used for the PacBio-only assembly. It is possible that assemblers are not yet equipped to deal with heterogeneous data. Programs like SPAdes, which were developed specifically for Illumina data, have then incorporated the use of PacBio data (Bankevich et al., 2012). Canu, on the other hand, was developed for PacBio data and appears to be better at assembling the long reads (Koren et al. 2016). The assembly of the *C. microcarpa* genome shows that more data does not necessarily result in a better assembly. Further, the cost of the assembly is not only in the price of the sequencing, but also the time spent doing bioinformatics should be considered. The PacBio data required simple quality control and the canu assembler was very easy to use, significantly reducing the bioinformatics time compared to an Illumina-based assembly. Thus, the higher price of PacBio sequencing might be more cost effective given the lower bioinformatics cost.
3. The Evolutionary Relationship between *C. microcarpa* and *C. sativa*'s Subgenomes

*Camelina microcarpa* was confirmed to have a diploid genome as the draft assembly aligned once to the *A. thaliana* genome and once to each subgenome of *C. sativa* (figure 4, 5). Diploid *C. microcarpa*, appears to be more closely related to most of subgenome 2 of *C. sativa*. This was evident in the collinearity plots and phylogenies between the two species (figure 5, 6). However, the phylogenies constructed here suggest that the assignment of *C. sativa*'s chromosomes to subgenome requires reassessment.

**Genome Changes after Polyploidization**

Following genome doubling the newly formed polyploid undergoes a series of genomic changes to return to a diploid state, which may include genome rearrangements and gene loss or fractionation (Soltis et al., 2016). During the process of gene loss, one of the subgenomes may become preferentially fractionated, leading to subgenome dominance (i.e. one of the genomes retains more genes; Renny-Byfield and Wendel, 2014). Based on the assumption of genome dominance, the subgenomes of maize and *B. rapa* were reconstructed (Schnable et al. 2011; Cheng et al., 2012). Fractionation and extensive genome rearrangements are predominant mechanisms that have been identified in the *Brassicas, Arabidopsis*, maize and cotton (Cheng et al. 2013; Thomas et al., 2006; Schnable et al. 2011; Rapp et al. 2009). However, these mechanisms are far from being considered to occur across all plant species (Soltis et al., 2016). In fact, subgenome dominance was not found on all plant species, and it is
thought to be associated with allopolyploidy, but not with autopolyploidy (Garsmeur et al., 2014).

Division of Chromosomes into Subgenomes

The chromosomes were assigned to the subgenomes by Kagale et. al (2014) based on a visual evaluation of the level of synteny with *A. lyrata*. The (less syntenic) most fractionated chromosomes were considered the most diverged from *A. lyrata* and assigned to CsatG3 while the most syntenic were assigned to CsatG1 (personal com. Kagale, Parkin, and Sharpe June 13, 2017). However, genome rearrangement and fractionation are not predominant mechanisms in *C. sativa*. In fact, the *C. sativa* chromosomes show very little reshuffling (Lysak et al., 2016). Extensive collinearity was found between *C. sativa* and *Arabidopsis* species, and ancestral genomic blocks were found to be almost intact in all three subgenomes (Kagale et al., 2014). In addition, the subgenomes have retained almost all copies of their genes CsatG1-3 encode 28,274, 27,218 and 29,207 genes, respectively (Kagale et al., 2014). Reconstructing the *C. sativa* subgenomes is important, because few rearrangements in the genome implies that each subgenome should be a close representation of *C. sativa* diploid ancestors’ genomes. However, assigning the chromosomes to subgenomes in *C. sativa* based on fractionation and without knowing the evolutionary history (allopolyploidy or autopolyploidy) may be problematic (Garsmeur et al. 2014).
Reevaluation of C. sativa Subgenomes

The relationship between the C. sativa’s chromosomes and homologous genomic regions of C. microcarpa were evaluated using a Bayesian phylogenetic analysis within each ancestral karyotype block (figure 6). This method takes advantage of the phylogenetic signal in the sequence data and the fact that C. sativa’s genome has been anchored to chromosomes. This approach produces two lines of evidence that can be used to assign C. sativa’s chromosomes to subgenome. First, the phylogenetic trees contain information on how C. sativa’s individual chromosomes are related separating chromosomes from Csat-3 from the more closely related chromosomes in Csat-1 and Csat-2. Second, the inclusion of regions from C. microcarpa provide additional context for interpreting the relationship among C. sativa’s chromosomes and refines the ability to assign chromosomes to CsatG1 or CsatG2.

The phylogenies show that, in general, chromosomes assigned to CsatG1 and CsatG2 are, as expected, more closely related to each other than to chromosomes assigned to CsatG3. However, there were a few cases where a different phylogenetic relationship was observed, specifically in sequences from blocks A, B, C, G, O, and P. For blocks belonging to ancestral karyotype chromosome 1 (A, B, C) the chromosomes assigned to CsatG2 (chromosome 14) and CsatG3 (chromosome 3) were found to be more closely related than chromosome 17 (CsatG1). Since this relationship was observed consistently throughout the blocks on these chromosomes and the phylogenies were constructed from long alignments (appendix table A3), I propose that
chromosome 3 should be assigned to CsatG1 and chromosome 17 should be assigned to CsatG3 (figure 7). Genomic block G, indicates that chromosome 19 from CsatG2 is more closely related to chromosome 1 from CsatG3 instead of chromosome 15 CsatG1. However, the two other blocks on this chromosome, which were constructed from larger alignments, support the expected relationship of CsatG1 and CsatG2 being more closely related. There are three potential explanations for this contradiction: 1) at an early stage during the formation of the polyploid there was recombination between the subgenomes, 2) recombination occurred in block G between CsatG1 and CsatG3; or, more probably, 3) the chosen regions for G may have not been truly homologous. Therefore this evidence, from block G, is insufficient to suggest chromosomes 15, 19, and 1 have been misassigned. As, blocks O and P were constructed from short alignments they also are weak evidence for altering the current assignments and will be discussed below.

Recombination between subgenomes or non-reciprocal recombination (gene conversion) between duplicated sequences is expected to result in conflicting phylogenetic signals in polyploids (Wendel, 2000). However, the rate of these would have to be high or the alignments would have to be small to have an impact on the overall phylogeny. For example, in Escherichia coli it was shown that a high rate of conversion, higher than the observed value, would be necessary to change the tree topology (Touchon et al., 2009). Thus, we do not expect that these processes will undermine the conclusions presented here.
The inclusion of genomic regions from *C. microcarpa* in phylogenies with *C. sativa*’s chromosomes allows for a better assessment of the relationship between chromosomes assigned to CsatG1 or CsatG2. The more contiguous and darker lines in figure 5 indicate that *C. microcarpa* is more collinear with most of CsatG2, except the chromosomes 6, 10 and 18, which represent the ancestral chromosomes 5, 6 and 8. In this case, *C. microcarpa* is more collinear to chromosomes 4 and 11 from CsatG1. Most phylogenies reflected these observations (figure 6). Nine phylogenies supported the conclusion that *C. microcarpa* is most closely related to the chromosomes assigned to CsatG2 (blocks A, B, C, E, F, G, H, Q, R). In contrast, most genomic blocks from ancestral chromosomes 5, 6 and 8 were more closely related to *C. sativa* CsatG1 (blocks L, M, N, T, U, W, X). I propose that these *C. sativa* chromosomes were erroneously assigned to these subgenomes. Based on the relationship with *C. microcarpa*, chromosomes 6, 10 and 18 should be placed in CsatG1, and 4 and 11 should be placed in CsatG2 (figure 7). This indicates that CsatG2 is a 6 chromosome genome, which is the same number of chromosomes found in diploid *C. microcarpa*.

While some of the phylogenetic relationships of the genomic blocks (D, J, O, P, S, V), show signals that conflict with the general observation that *C. microcarpa* is most closely related to CsatG2, most of these trees were constructed from shorter alignments than the rest of the phylogenies (e.g. O with 817 bp; figure 6, appendix table A3). The reason why we retrieved a small number of fragments for some blocks may be due to a lack of collinearity
between the genomes within those blocks. To construct the phylogenies, each genome (or subgenome) was aligned to *A. lyrata*. The alignments were used to find the regions of each genome that were collinear to each of the ACK blocks in *A. lyrata*. Out of all the fragments found per block, only the fragments that were found to overlap between all of the genomes were selected. Then, based on how the ACK blocks are described in the *C. sativa* karyotype (Kagale et al. 2014), the only groups of overlapping fragments accepted had the *C. sativa* blocks in the expected chromosomes. For example, for block A only the groups of fragments that were collinear with *A. lyrata* from chr1:47-8,419,004, and the collinear *C. sativa* fragments were found on chromosomes 17, 14 and 3 were used for the phylogeny of this block (appendix table A1, A2). Only one possible *C. sativa* combination was evaluated, for example collinear fragments that included *C. sativa* chromosomes 16, 7 and 5 for block E were evaluated, but fragments with other valid combinations such as 16, 7 and 9 were not evaluated. The small number of fragments found for certain blocks suggests that the regions evaluated, for that block, were not the most collinear and may lack homology. This could be because we selected a group of *C. sativa* chromosomes that were not homologous, or rearrangements within and between blocks has reduced collinearity in these regions. In fact, this appears to be the case with the exceptions found (D, J, O, P, S, V); D, J have undergone several splits and rearrangements, O split in CsatG2 and CsatG3, P and S split in CsatG3, and V rearranged in CsatG3. Furthermore, we did not recover any fragments from blocks I and K. Genomic block I,
has been rearranged and split in *C. sativa*, and could indicate that the sections with homology fall on different chromosomes than expected or that collinearity has been reduced after rearrangements making it difficult to extract a phylogenetic signal. Conversely, the lack of fragments for Block K may be because it is one of the smallest blocks. Further, Lysak et al. (2016) merged K-L in their new ancestral karyotype block’s designation into one block, thus the information from block L should be consistent with block K. As a result, these conflicting signals appear to be most likely attributable to a lack of homology in these particular blocks and do not change the interpretation based on the more complete information available in the majority of the blocks.
Figure 7. *Camelina sativa* subgenome structure as proposed by Kagale et al. (2014, left) and the structure proposed here (right). The new chromosome assignments are based on the evolutionary relationship of each conserved block (A-X) with *C. microcarpa*. The arrows indicate which chromosomes have been reassigned. Figure edited from Kagale et al. (2014)
Summary and Future Work on Subgenome Structure

This study shows how the ACK blocks can be used to infer phylogenetic relationships between plant species and inform the genome reconstruction process. Using the draft assembly of diploid *C. microcarpa*, based on PacBio data, the evidence presented here suggests that the genome is most closely related to CsatG2. I also show that the subgenome structure proposed by Kagale et al. (2014) conflicts with the results obtained here and propose a new subgenome structure for *C. sativa*. Given that biased fractionation and genome rearrangements are not predominant in *C. sativa* (Kagale et al., 2014; Lysak et al., 2016), it may be more accurate to assign the chromosomes based on phylogenetic data.

Future work should seek to confirm the changes proposed here with alternative methods. For example, whether the genomic block associations in CsatG2 are found in diploid *C. microcarpa* could be evaluated through genome in situ hybridization (GISH) using bacterial artificial chromosomes (BACs) constructed for that purpose for each of the ACK blocks (Lysak et al., 2016). This could, for example, be used to evaluate if ancestral chromosome 7 and 8 are joined in *C. microcarpa*. Alternatively, a genome-wide approach could be taken if the genome is finished to the chromosome level. Then, the genomic blocks would be identified using microscopy or gene synteny, the latter would also require transcriptome sequencing. Anchoring the scaffolds to chromosomes would also allow for the evaluation of misassembly, for example *C. microcarpa*’s scaffold 13 was found to span two different chromosomes of *C. sativa* (figure
With the current assembly, it is not possible to ascertain if this joining is a misassembly by canu, or a rearrangement that occurred in *C. microcarpa* after the two species diverged. The draft genome of *C. microcarpa* will also facilitate the sequencing of tetraploid and hexaploid *C. microcarpa*, which would provide a parallel system for understanding polyploidy and genome evolution when compared to *C. sativa*.

4. **Insights into Plant Evolution can aid Crop Development**

With the sequencing of diploid *C. microcarpa* we can start to unravel the evolutionary history of the *Camelina* genus. More broadly, we can increase our understanding of how hybridization, polyploidization and genome structure have changed and shaped the genus and give us an insight into how these processes have shaped plant evolution. The description of the ACK blocks propelled our understanding of crucifer evolution (Lysak et al., 2016). The genome of *C. microcarpa* once finished to the chromosome level will add to the diploid karyotypes described for the Brassicaceae (figure 2) and allow further comparisons between diploid *C. microcarpa* and hexaploid *C. sativa*.

As a next step in this process, the other known diploids in the genus: *C. laxa* and *C. hispida*, should be evaluated with a repetition of the analysis presented here. In the long term, obtaining a *C. microcarpa* reference genome and those of the other two known diploid species will allow us to better understand the evolutionary history of this genus. If the progenitors of *C.
sativa’s are identified, it may be possible to resynthesize hexaploid C. sativa to develop higher producing C. sativa lines, an approach that has been successful canola (Girke et al., 2011). The draft genome of diploid C. microcarpa is also valuable in terms of the development of the C. sativa crop.

Years of domestication have led to decreased genetic variability in C. sativa populations (Manca et al., 2012), which will mean that to improve C. sativa, other species of the genus might need be explored. One major challenge that will be faced in attempts to improve the oil profile is that classical breeding is hindered by the complex nature of the hexaploid genome of C. sativa (Bansal and Durrett, 2016). The overwhelming majority of plant research is focused on crop species and model plants, but wild crop relatives, which are key in understanding the evolutionary history of crops, are often neglected. Wild relatives can also be a source of desirable traits for the improvement of crop species (Ford-Lloyd et al., 2011) something that is facilitated by the availability of genomic resources. Some potential traits of interest for C. sativa that have been identified in C. microcarpa are fungal disease resistance, shatter resistance and insect resistance (greenhouse observations). Further, these genomic resources are increasingly being recognized for their value in developing pest management strategies for weeds (Stewart et al. 2009).

DNA re-sequencing and RNA sequencing will facilitate the characterization of the genetic variability in different C. microcarpa accessions collected from North America and
Eurasia. After characterization, if traits of interest for breeding such disease resistance are found, the *C. microcarpa* genome will allow for the identification of target genes for breeding. Additionally, this diploid genome is a step towards the reconstruction of these larger polyploid genomes. Importantly, since we have found that hexaploid *C. microcarpa* and *C. sativa* are reproductively compatible, *C. microcarpa* reference genomes could allow for *C. sativa* transgenes to be designed to prevent movement into *C. microcarpa* wild populations. For example, transgenes in *C. sativa* could be placed in one of the chromosomes from a subgenome that is not shared with hexaploid *C. microcarpa*. Furthermore, sequencing *C. microcarpa* provides an opportunity to compare the genomes of a domesticated and a weedy species.

The study of the processes that are involved in plant evolution are not only important to understand the natural history of plants, but can be applied in the development of better crops (Renny-Byfield and Wendel, 2014). Not only should natural polyploids be studied to develop superior crops, but synthetic polyploids were proposed very early to develop better crop varieties (Randolph 1941). Polyploid plants often show increased vigour (e.g. biomass, stature, growth rate and fertility) that can allow them to thrive in new environments (Chen, 2010, Sattler et al., 2016). In fact, many of our current crops are polyploids and it is estimated that 40% of all cultivated plants are polyploids (Sattler et al., 2016). Eigsti and Dustin suggested that colchicine-induced synthetic polyploids would be the new era of polyploid research and promise great opportunities for crop improvement (1955). One successful example of a
synthetic polyploid crop is triticale, a man-made wheat and rye hybrid that has been developed in different ploidy levels: tetraploid, hexaploid and octoploid. The triticale crop was made to incorporate the yield and nutritious qualities from wheat and disease resistance of rye (Chen, 2010). The hexaploid variety is the most successful in terms of meiotic stability and fertility (Kang et al., 2016), and has also been a model to study allopolyploid at an early stage (Ma and Gustafson, 2008). The triticale crop is successfully grown in Poland, Australia, Germany, France, China and in Canada, for animal feed (Chen, 2010; Mergoum and Gómez-Macpherson, 2004). However, to date, there has been limited success with induced polyploids (Sattler et al., 2016) because of a lack of ability to predict or control the characteristics of induced polyploids prior to synthesis. The knowledge extracted from diploid-polyploid pairs might be the step necessary to make synthetic polyploids for crop development more successful.

5. Next Steps
The draft genome of *C. microcarpa* presented here is in 269 scaffolds, to obtain a reference genome the contigs would have to be anchored to the 6 chromosomes of *C. microcarpa*. This could be done by the development of a physical map (Fierst, 2015) or through methods such as Hi-C (Burton et al., 2013). To generate a physical map, mapping populations need to be established and genetic markers have to be identified by sequencing a large number of individuals, which can be time consuming and expensive (Fierst, 2015). A recent approach involves the HiC technique, which induces chromatin cross-links and uses high-throughput
sequencing to identify all interactions between chromatin segments. Based on the frequency of the interactions between two segments, the orientation and distance between two sequences (e.g. scaffolds) can be estimated (Fierst, 2015). To obtain a chromosome-level assembly through the Hi-C approach, the sequencing facility would require DNA extracted from the plant, the draft assembly and the raw data used for the draft assembly. A chromosome-level assembly would allow for a more complete understanding of diploid *C. microcarpa*’s genome. For instance, the collinearity plots presented here would be easier to interpret at the chromosome level than visualizing 269 pieces of DNA. Furthermore, a chromosome-level assembly could be used to identify the ACK blocks *in silico* and evaluate some of the block associations that are characteristic of CsatG2 to confirm the results presented here, as mentioned above.

An important part of scientific research is to make data available to the scientific community, to allow for collaborations and accelerate scientific advances. In the case of this draft genome, it is essential to make it available in databases such as the National Center for Biotechnology Information (NCBI). However, many researchers will be looking for specific genes to make comparisons between genomes, perform blast searches or identify gene targets. To allow for this type of research, annotation of the genome and the development of a searchable database would be necessary. Genome annotation could be done either by sequencing the transcriptome or a predictive *in silico* annotation using *C. sativa*’s known genes (Ekblom and Wolf, 2014). Both annotation approaches would be easier once a chromosome-level genome is
obtained. To have a searchable database, a genome browser could be established (Want et al., 2013). Genome browsers have the advantage of providing a user interface that make browsing and extracting data easier than working directly with sequence files. A genome browser has already been developed for *C. sativa* (Kagale et al., 2014).

Once the genome is publicly available, some *C. microcarpa* genes will be of interest, for example genes involved in vernalization and stratification. To promote the growth of *C. microcarpa* plants, the seeds need to be treated with KNO₃ to overcome dormancy (stratification) and the potted plants need to be exposed to cold temperatures to induce the production of seeds (vernalization). These processes are characteristic of weeds (Wagenvoort and Van Opstal, 1979), and it would be of interest to understand the genetic basis of these mechanisms in *C. microcarpa*.

Moreover, future work will aim to improve breeding strategies for *C. sativa*. Increasing oil content and seed size are main targets for successful oilseed crops (Francis and Warwick, 2009). Currently, the oil content of *C. sativa* seed is between 30-43% (Francis and Warwick, 2009), which needs to be considerably increased to compete with canola where oil content as high as 64.8% has been reported (Hu et al., 2013). It would be important to describe the genes involved in seed oil content and size in *C. microcarpa* and to compare to the *C. sativa* genes. Furthermore, knowing these genes will become important if *C. sativa* was to be resynthesized.
Similarly, the oil biosynthesis pathway in *C. microcarpa* should be described. The strategy would be to identify genes involved in seed oil biosynthesis and analyze the copy number and if they are functional in *C. microcarpa*.
REFERENCES


Davis, P.H. (1965) Flora of Turkey and the East Aegean Islands Vol. 1 Univ. Pr.


APPENDICES

Appendix A: Flower picture references for figure 1 and 2. All images are under creative commons license: https://creativecommons.org/licenses/


Appendix B: Script used to run Mr. Bayes, the parameters used here followed recommendations from: Hall, 2011

begin mrbayes;
log start replace filename = nameof.log;
set autoclose = no nowarn=no;
lset applyto = (all) nst=6 rates =invgamma;
   [this line sets parameters for the GTR+G model]
prset applyto = (all);
unlink revmat=(all) shape=(all) pinvar=(all) statefreq=(all) tratio=(all);
showmodel;
mcmc ngen=10000000 printfreq=1000 samplefreq=100 nchains=4 temp=0.2 checkfreq=50000
diagnfreq = 1000 stopval=0.01 stoprule=yes;
sumt relburnin = yes burninfrac = 0.25 contype = halfcompat;
sump relburnin =yes burninfrac = 0.25;
log stop;
end;
Table A1. Coordinates used to define the 24 Ancestral Crucifer Karyotype (ACK) blocks in this study, based on Murat et al. (2015)

<table>
<thead>
<tr>
<th>Block</th>
<th>A. lyrata chromosome</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>47</td>
<td>8,419,004</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>8,451,241</td>
<td>17,907,944</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>17,927,677</td>
<td>29,795,265</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>47,120</td>
<td>4,870,167</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>7,479,474</td>
<td>19,302,151</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>1,739</td>
<td>13,046,955</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>13,064,549</td>
<td>14,793,089</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>17,459,514</td>
<td>24,143,508</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
<td>92,641</td>
<td>6,759,443</td>
</tr>
<tr>
<td>J</td>
<td>4</td>
<td>10,397,564</td>
<td>23,326,981</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>82,874</td>
<td>2,204,116</td>
</tr>
<tr>
<td>L</td>
<td>5</td>
<td>2,230,599</td>
<td>7,079,440</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
<td>11,139,946</td>
<td>14,828,303</td>
</tr>
<tr>
<td>N</td>
<td>5</td>
<td>14,828,445</td>
<td>21,206,319</td>
</tr>
<tr>
<td>O</td>
<td>6</td>
<td>22,225,787</td>
<td>25,073,253</td>
</tr>
<tr>
<td>P</td>
<td>6</td>
<td>17,238,895</td>
<td>22,194,597</td>
</tr>
<tr>
<td>Q</td>
<td>6</td>
<td>9,794,390</td>
<td>13,607,357</td>
</tr>
<tr>
<td>R</td>
<td>6</td>
<td>1,708</td>
<td>9,714,052</td>
</tr>
<tr>
<td>S</td>
<td>7</td>
<td>18,070,607</td>
<td>24,468,494</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>12,288,881</td>
<td>14,981,448</td>
</tr>
<tr>
<td>U</td>
<td>7</td>
<td>2,747</td>
<td>12,257,902</td>
</tr>
<tr>
<td>V</td>
<td>8</td>
<td>59,585</td>
<td>6,055,532</td>
</tr>
<tr>
<td>W</td>
<td>8</td>
<td>11,545,323</td>
<td>19,711,563</td>
</tr>
<tr>
<td>X</td>
<td>8</td>
<td>19,735,868</td>
<td>22,764,519</td>
</tr>
</tbody>
</table>
Table A2. The 24 ACK blocks and their placement on the chromosomes of *A. lyrata* (Alyr) and the three subgenomes (G1-3) of *C. sativa* (Csat)

<table>
<thead>
<tr>
<th>ACK Block</th>
<th>Chromosome Placement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alyr</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
</tr>
<tr>
<td>J</td>
<td>4</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
</tr>
<tr>
<td>L</td>
<td>5</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
</tr>
<tr>
<td>N</td>
<td>5</td>
</tr>
<tr>
<td>O</td>
<td>6</td>
</tr>
<tr>
<td>P</td>
<td>6</td>
</tr>
<tr>
<td>Q</td>
<td>6</td>
</tr>
<tr>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td>S</td>
<td>7</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
</tr>
<tr>
<td>U</td>
<td>7</td>
</tr>
<tr>
<td>V</td>
<td>8</td>
</tr>
<tr>
<td>W</td>
<td>8</td>
</tr>
<tr>
<td>X</td>
<td>8</td>
</tr>
</tbody>
</table>
Table A3. Concatenated sequence length for each Ancestral Crucifer Karyotype (ACK) block used to construct phylogenies of *C. microcarpa*, the three subgenomes of *C. sativa*, and *A. lyrata* (outgroup)

<table>
<thead>
<tr>
<th>ACK block</th>
<th>Sequence length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57293</td>
</tr>
<tr>
<td>B</td>
<td>58676</td>
</tr>
<tr>
<td>C</td>
<td>61570</td>
</tr>
<tr>
<td>D</td>
<td>61112</td>
</tr>
<tr>
<td>E</td>
<td>52092</td>
</tr>
<tr>
<td>F</td>
<td>60645</td>
</tr>
<tr>
<td>G</td>
<td>30165</td>
</tr>
<tr>
<td>H</td>
<td>55309</td>
</tr>
<tr>
<td>I</td>
<td>-</td>
</tr>
<tr>
<td>J</td>
<td>12470</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
</tr>
<tr>
<td>L</td>
<td>49806</td>
</tr>
<tr>
<td>M</td>
<td>53054</td>
</tr>
<tr>
<td>N</td>
<td>50599</td>
</tr>
<tr>
<td>O</td>
<td>817</td>
</tr>
<tr>
<td>P</td>
<td>6342</td>
</tr>
<tr>
<td>Q</td>
<td>27886</td>
</tr>
<tr>
<td>R</td>
<td>54305</td>
</tr>
<tr>
<td>S</td>
<td>28845</td>
</tr>
<tr>
<td>T</td>
<td>14489</td>
</tr>
<tr>
<td>U</td>
<td>50167</td>
</tr>
<tr>
<td>V</td>
<td>4604</td>
</tr>
<tr>
<td>W</td>
<td>50679</td>
</tr>
<tr>
<td>X</td>
<td>52288</td>
</tr>
</tbody>
</table>