High-quality genome assemblies of plant species using third generation genomic technologies and comprehensive genotypic characterization of *Arabis* inter-species introgression lines

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Wenbiao Jiao

aus Xianning, Hubei, China

Köln, 2017

Berichterstatter: Prof. Dr. George Coupland

Prof. Dr. Martin Hülskamp

Tag der mündlichen Prüfung: Januar 10 2018

Abstract

As genome sequences are fundamental to many molecular studies, researchers have been aggressively pursuing a cost-and-time efficient solution to dissect the complexity of genome sequences with the progressively advanced DNA sequencing technologies. Over the past decades these have experienced developments from traditional Sanger-based sequencing to so-called second generation sequencing and recently to third generation genomic technologies. Genome assembly is time-consuming and expensive for Sanger-based sequencing, or leads to fragmented genome sequences when it is based on second generational sequencing. The newly emerging third generation genomic technologies including long-read sequencing and long-range mapping however promise fast, cost-effective and chromosome-level genome assemblies. Long-read sequencings such as PacBio Single Molecular Real-Time (SMRT) and Oxford Nanopore sequencing currently generate reads with average length of around 10 kb. While long-range mapping technologies like optical mapping and chromatin contact sequencing can produce linking information even spanning multiple hundred kb or several Mb.

In this thesis, I will firstly summarize current third generation genomic technologies and their applications on plant genome assembly, and discuss how they might overcome the assembly challenges of heterozygous and polyploid plant genomes in the introduction section. In my first project, to compare the performance on genome assembly of these different technologies, I used PacBio long reads, BioNano optical mapping and Dovetail chromatin contact data to obtain high-quality assemblies of three relatives of the model plant *Arabidopsis thaliana*. Both of the two long-range scaffolding technologies introduced similar improvements to a long-read based sequence assembly – not only the assembly contiguity but also the assembly accuracy. I also developed workflows to independently integrate optical mapping or chromatin contact data into assembly scaffolding. Further integration of these two long-range

data showed they were complementary in assembly scaffolding.

In the second project, I present the assembly and annotation of *Arabis montbretiana*, an annual sister of the perennial plant *A. alpina* (which assembly was introduced in the first project). The two high-quality *Arabis* genome assemblies allowed us to investigate the genomic changes underlying the annuality-perenniality evolution transition, and comprehensively characterize introgression lines constructed from the two *Arabis*. Interestingly, comparative genomic analysis between *A. montbretiana* and *A. alpina* revealed substantial genomic rearrangements including over 1,200 translocated genes, which introduced heterogeneous copy number variations among different introgression lines.

Zusammenfassung

Genomsequenzen sind für viele molekulare Studien von grundlegender Bedeutung. In den letzten Jahren wurden immer wieder neue kosten- und zeitsparende Lösungen gesucht, um Genomsequenzierung zu effizient und korrekt wie möglich zu gestalten. Die Entwicklung der DNA-Sequenzierungsverfahren ging dabei von der traditionellen Sanger-basierten Sequenzierung über die sogenannten Sequenzierungen der zweiten Generation bis zu den vor kurzem eingeführten Technologien der dritten Generation. Genom-Assemblierung mit der ursprünglichen Sanger-basierten Methode war zeitaufwändig und teuer, bzw führte basierend auf der zweiten Generation der Technologien zu fragmentierten Genomsequenzen. Die neuesten Technologien der dritten Generation, also Technologien, die lange DNA Moleküle auslesen können, versprechen schnelle, kostengünstige Assemblierung von nahezu vollständigen Genomsequenzen. Sequenzierungsmethoden wie PacBio's Single Molecular Real-Time (SMRT) und Oxford Nanopore's MINION Sequenzierungen generieren dabei DNA-Sequenzen mit einer durchschnittlichen Lese-Länge von etwa 10 kb. Darüber hinaus können Genomkartierungs-Technologien wie Optische Kartierung und Chromatin-Kontakt-Sequenzierung Strukturinformation über mehrere hundert kb oder mehrere Mb erstellen.

In dieser Arbeit werde ich zunächst die genomischen Technologien der dritten Generation und ihre Anwendungen auf die Pflanzengenom-Assemblierung zusammenfassen und diskutieren, wie sie die besonderen Herausforderungen von heterozygoten und polyploiden Pflanzengenomen in Einleitung überwinden können. In meinem ersten Projekt, beschreibe ich, wie ich zwei der neuen Genomekartierungs-Technologien im Kontext von der Genomassemblierung von drei Verwandten der Modell-Pflanze *Arabidopsis thaliana* verglichen habe. Beide Technologien führten zu ähnlichen Verbesserungen der Assemblierungsqualität. Darüber hinaus habe ich Workflows entwickelt, um beide Technologien zu integrieren, umso die Assemblierung nochmals zu verbessern.

Im zweiten Projekt stelle ich die Genomassemblierung und Annotation von *Arabis montbretiana* vor. *A. montbretiana* ist eine einjährige Schwester der mehrjährigen Pflanze *A. alpina* (deren Genomassemblierung in meinem ersten Projekt beschrieben ist). Die beiden hochwertigen Arabis-Genom-Assemblierungen ermöglichten es uns, die genomische Basis zu untersuchen, die der unterschiedlichen Evolutionen der beiden Pflanzen zugrunde liegt. Interessanterweise zeigte die vergleichende genomische Analyse zwischen *A. montbretiana* und *A. alpina* wesentliche genomische Umlagerungen, darunter über 1.200 translozierte Gene, die heterogene Variationen in den Nachkommen einer Kreuzung aus den beiden Arten erzeugt.

Acknowledgements

First of all, I would like to extend my sincerest appreciation to my supervisor, Dr. Korbinian Schneeberger. Thanks to your consistent support, thoughtful advices and continuous encouragement, which help me finish my PhD study.

I would also like to thank the referees, Professor George Coupland and Professor Martin Hülskamp. And I would like to appreciate George for supporting my PhD study. Many thanks to my collaborators: Christiane Kiefer, Edouard Severing, Eva Madrid-Herrero, Stefan Woetzel, Bruno Huettel, Richard Reinhard, Gonzalo Garcia Accinelli, Bernardo Clavijo, Daniel Swan.

Also thanks to my group mates: Benjamin Hartwig, Eva-Maria Willing, Hequan Sun Jonas Klasen, Manish Goel, Maria Cartolano, Mathieu Piednoel, Ulrike Hümann, Vidya Oruganti, Vimal Rawat, Vipul Patel, for your help and discussions.

Finally, special thanks to my family and my girlfriend for their love and support.

Contents

List of Figures

List of Tables

List of Abbreviations

ACK	Ancestral Crucief Karyotype
CENT	centromere
CMAP	consensus optical map
CNV	copy number variation
CTX	inter-chromosome translocation
InDel	insertion and deletion
INV	inversion
ITX	intra-chromosome translocation
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism
SYN	synteny

1 Introduction

Since the DNA structure was firstly discovered in 1953 (Watson and Crick 1953), numerous marvelous technologies have been developed to dissect the complexity of genome sequences. The first generation sequencing technology appeared around forty years ago, and was improved to perform automatically at large-scale (Sanger et al. 1977; Swerdlow et al. 1990; Hunkapiller et al. 1991), which paved the start of the era of genomics. However, Sanger sequencing is time-consuming and expensive. Later, cheap high-throughput, so-call second (or next) generation sequencing (NGS) were introduced including 454, Illumina and SOLiD technologies, and soon applied for various omics-based studies such as genomics, transcriptomics, or epigenomics (Shendure and Ji 2008). The read length of NGS is typically up to multiple hundred bp, which affects the performance of some applications such as the high-quality assembly of genome sequences and the identification of large structural variation (SV). In the recent years, long read sequencing technologies like PacBio Single Molecular Real-Time (SMRT) sequencing and Oxford Nanopore sequencing have been developed to overcome the limitations of NGS, promising one assembly contig per chromosome. Besides, other long-range mapping technologies including optical mapping, dilution-based lined-read sequencing and proximity ligation-based sequencing, also emerge to support chromosome-level genome assembly, haplotyping of diploid genomes and detection of large SV. These long-read sequencing and long-range mapping methods are collectively referred to as third generation genomic technologies.

1.1 The impact of third generation genomic technologies on plant genome assembly

This chapter section 1.1 was the basis of a manuscript that was published as an overview article in Current Opinion in Plant Biology (Jiao and Schneeberger 2017), which lists me as first author.

As the expense of genome sequencing decreases substantially in the past ten years, around 200 plant genome reference sequences have been assembled until now (as of March 2017; www.plabipd.de), and *de novo* assemblies of multiple individuals within the same species have been reported in recent years (Zapata et al. 2016; Schatz et

al. 2014; Zhang et al. 2016; Li et al. 2014; Hirsch et al. 2016). However, only a few of them are assembled on chromosome-level (e.g. (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Schnable et al. 2009)). Most of them are assembled by using short reads and greatly fragmented.

These poor assembles are typically caused by the intrinsic features of plant genomes (Schatz et al. 2012), which often present greatly repetitive sequences, huge genome size (Neale et al. 2014; Nystedt et al. 2013), and complex polyploid nature (Salman-Minkov et al. 2016). These features are also reflected by the repetitiveness analysis, which obviously indicates that the assembly of plant genome is challenging and even more difficult comparing to vertebrate genome assembly (Fig. 1). Recently, increasing number of high-quality or even chromosome-level genome assembly are reached (Koren and Phillippy 2015), by using the third generation genomic technologies including long-read sequencing (Eid et al. 2009; Deamer et al. 2016; McCoy et al. 2014) and long-range scaffolding methods (Amini et al. 2014; Burton et al. 2013; Schwartz et al. 1993; Zheng et al. 2016). In this first section of the introduction, I will briefly introduce several third generation genomic technologies and focus on their applications for plant genome assembly.

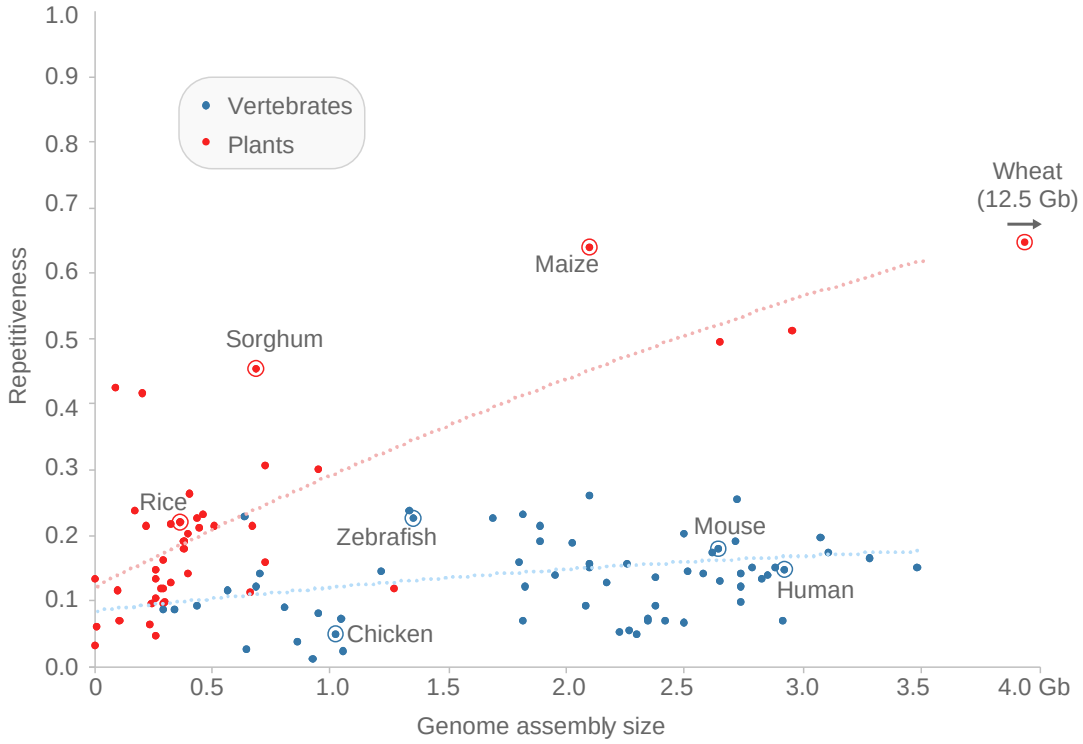


Figure 1 Comparison of size and repetitiveness of plant and vertebrate genomes. The figure is from (Jiao and Schneeberger 2017). In total, repeat and genome size of 44 plant and 68 vertebrate genome assemblies are analyzed. Plant genomes generally has higher repetitiveness than vertebrates and the repetitiveness is more correlated to genome

size. The assembly of plant genome is challenged not only by genome size but also by higher levels of repetitiveness. (All genome sequences are from the database of Ensembl Genomes release 32 (Kersey et al. 2016); The repetitiveness in each assembly was calculated based on the percentage of unique 31-mers of all 31-mers; dashed lines indicate the data follows a polynomial regression)

1.1.1 Long-read sequencing technologies

Several long-read sequencing technologies have been developed and applied commercially. Among of them, Pacific Biosciences' (www.pacb.com) Single Molecule Real-Time (SMRT) sequencing is the most-widely used, with an average read length of nearly 20kb and a maximum length of more than 60kb currently (Eid et al. 2009; Kim et al. 2014; VanBuren et al. 2015). It generates raw reads with error rates of up to 15% which is much less accurately than Sanger or second generation sequencing. However, these long error-prone reads can be corrected either by short sequencing reads (Koren et al. 2012; Bashir et al. 2012) or self-correction with sufficient sequencing coverage (Chin et al. 2013; Koren et al. 2013), enabling genome assemblies with a sequence accuracy more than 99.999%, simply by running assembly tools, such as FALCON or PBcR(MHAP) (Chin et al. 2016; Berlin et al. 2015). As the PacBio sequencing are still expensive especially for assembling large genomes, lower sequencing depth of long reads are often applied for assembly improvement by gap closure or scaffolding (Zapata et al. 2016; Bombarely et al. 2016; Yang et al. 2016). Not like *de novo* assemblies, typically single software tools are used to integrate such data (English et al. 2012; Bashir et al. 2012).

So far, a few of genomes have been assembled from PacBio data alone including several plants such as *Arabidopsis thaliana* (Berlin et al. 2015) *, Oropetium thomaeum* (VanBuren et al. 2015). The *Arabidopsis* genome of accession Ler-0 was assembled at nearly chromosome-arm level, while the 245Mb *O. thomaeum* genome presented a contig N50 of 2.4Mb. Such assembly contiguities were never approximated only by short-read data without scaffolding using long-range read pairs before (Hoshino et al. 2016). Apart from contiguity, PacBio assemblies have less gaps and cover more genomic space. For example, the PacBio reads based assembly of *Arabis alpina*, was 337Mb long, 30Mb longer than an earlier Illumina short-read assembly, while the gaps percent decreased from 9.2% to 3.3% (Willing et al. 2015; Jiao et al. 2017). However, the PacBio contigs could have some misassemblies, where two or more regions were mis-joined due to substantially high transposon elements (Jiao et al. 2017). Besides, a few of residual errors of single nucleotide or short InDels still exist in the PacBio

assembly, which can be further corrected with modest depth of short-read sequencing (Jiao et al. 2017).

In addition, two other long-read sequencing technologies are commercially available. One is called nanopore sequencing developed by Oxford Nanopore Technologies (www.nanoporetech.com), which released their first sequencing system in 2014 (Deamer et al. 2016; Quick et al. 2014). In nanopore sequencing, single DNA molecules pass through voltage-biased nano-scale holes, made from proteins puncturing membranes or solid materials, where the change of ironic current created by individual nucleotides are measured to identify that molecule. In theory, the nanopore sequencing read length is only limited by the DNA fragment length as the system can process DNA molecules with any length. While Nanopore reads reported recently have similar average read length and nucleotide accuracy compared to PacBio, and longest length over 200kb. First whole-genome assemblies using Oxford Nanopore data were from fungal by hybrid correction using short sequencing data or bacterial by self-correction with sequencing depth nearly 30x , which had contig N50 of 678 kb for fugal genomes, and a single contig for an *E. coli* genome with nucleotide accuracy more than 99.9% (Loman et al. 2015; Goodwin et al. 2015).

Another long-read technology is Illumina's Synthetic Long-Reads (SLR) system (www.illumina.com), which specifically assemble long reads from Illumina short reads (Voskoboynik et al. 2013; McCoy et al. 2014). In this technology, the large DNA fragments of ~10 kb are distributed into 384-well plates such that each well only contains a very small fraction of genome. Thus overlapping fragments in one well are unlikely from the same region. Next, these fragments are amplified by PCR, connected with a unique barcode for each well, and sequenced on an Illumina platform. Finally, the SLRs can be assembled from these short reads originating from the same well with same barcode. Recent reports showed SLR had length ranging from 2 to 18 kb and nucleotide accuracy of over 99.9%, much higher than that of PacBio or nanopore reads (Li et al. 2015b; McCoy et al. 2014). However, substantially higher amounts of short reads with multiple hundred-fold sequencing depth are required to generate SLRs for genome *de novo* assembly. And it is not clear whether reads of 10 kb level can be assembled in highly repetitive regions or genomes. Besides, a recent research revealed considerable misassemblies in *C. elegans* genome assembly from SLRs (Li et al. 2015b). However, the shortcomings of assembly tools might also result in misassemblies, implying more advanced algorithms especially for SLRs *de novo* assembly are needed, as shown in a recently developed method (Bankevich and Pevzner 2016). All these might limit it to be widely applied for whole genome assembly. Until now, this technology has only been used to assemble several hundred-Mb

eukaryotic genomes (McCoy et al. 2014; Voskoboynik et al. 2013) and improve the assembly of a maize genome (Hirsch et al. 2016).

1.1.2 Long-range scaffolding technologies

Even though the long-read DNA sequencing exhibits remarkable progresses and advantages, it is still hard to obtain chromosome-level genome assembly only using long reads. Traditionally, the initial assembly contigs are further scaffolded to improve assembly contiguity, which can be typically conducted by using paired reads with fragment insert size ranging from several hundred bp to ~100 kb (Roach et al. 1995). Longer range of read pairs from BAC or forsmid ends are more powerful to overpass long repeats (Nagarajan and Pop 2013). In addition, genetic or physical maps are often used to generate a chromosome-level assembly. However, several new long-range genomic technologies are recently introduced to improve assembly scaffolding impressively, promising to eventually replace genetic and physical maps.

The first one is optical mapping which can generate ordered restriction maps of DNA molecules with length up to several hundred kb (Schwartz et al. 1993; Lam et al. 2012). Although it was already invented in 1993, the high-throughput platforms including Irys system (BioNano Genomics, bionanogenomics.com) and Argus system from (OptGen: www.opgen.com), emerged only a few years ago. This technology can measure the physical distances of restriction enzyme cut sites on DNA molecules using fluorescently marked enzymes. The individual optical maps can be assembled into consensus maps to scaffold primary contigs or identify large and complex structural variations (Kawakatsu et al. 2016; Chamala et al. 2013; Tang et al. 2014; Nagarajan et al. 2008; Yang et al. 2016; VanBuren et al. 2015). Another review has summarized the applications of optical mapping to plant genomic studies in detail (Tang et al. 2014).

Unlike sequencing reads assembly typically fragmented by repeats, optical map assemblies are biased to break at regions with two restriction sites closely located on opposite strand (Pendleton et al. 2015), suggesting sequencing data and optical maps can be efficiently combined to improve assembly contiguity. While this performance of assembly improvement relies on the contiguity of prior sequencing assembly and also the optical consensus maps themselves (Tang et al. 2014; Zhou et al. 2007). One powerful combination is to use optical maps to scaffold long-read assembly contigs, as applied on several plant genomes (Jiao et al. 2017). Besides, optical maps can also help to find and correct misassemblies (Yang et al. 2016; Hastie et al. 2013). By breaking misassembled contigs or scaffolds, it even generated longer scaffolds (Jiao et al. 2017). Intriguingly, the errors in optical consensus maps could be also identified

by sequence contigs. However, the current combining strategies seldom integrate *de novo* assembly of sequencing reads while simultaneously include optical maps to disentangle the assembly graph (Lin et al. 2012).

Another method to enable chromosome-scale assembly is using chromosome conformation capture sequencing (Hi-C). Hi-C is initially used to detangle the three-dimensional architecture of chromosomes, where spatially close DNA regions are ligated and sequenced using paired-end sequencing (Lieberman-aiden et al. 2009). Most of Hi-C read pairs are from two closely linked regions, although not all of them are close in linear chromosomes. Besides, the contact frequency of distinct regions generally decreases in proportion to the linear distance and intra-chromosome regions interact more frequently than that from different chromosomes. Therefore, the Hi-C read pairs can provide from mid-range to long-range, even centromere-spanning information of linear distance of interacting regions, which can be utilized to do assembly scaffolding (Burton et al. 2013; Selvaraj et al. 2013). Recently Dovetail Genomics (www.dovetailgenomics.com) introduced a modified Hi-C protocol called Chicago, which is based on *in vitro* reconstituted chromatin to remove confounding biological signals (Putnam et al. 2016). The Chicago sequencing data integration for scaffolding also contains two steps. The misassembled regions of initial assembly are identified and broken firstly, and the resulting contigs are scaffolded. A recent study showed the Dovetail read pairs could generate similar improvement of assembly contiguity compared to optical consensus maps (Jiao et al. 2017). Moreover, they can be combined to further increase the assembly contiguity as they help connecting different complex regions of the genomes (Jiao et al. 2017).

The last remarkable technology is a microfluidics-based sequencing developed by 10X Genomics (www.10xgenomics.com) based on the GemCode technology. In their system, DNA fragments with size 50kb or longer are dispersed into over 10,000 droplet partitions, then barcoded and pooled together to conduct Illumina short-read sequencing (Zheng et al. 2016). The newest system called “Chromium” utilizes an updated partitioning system with approximately one million partitions (Weisenfeld et al. 2016). This technology is conceptually similar to Illumina’s SLRs and long fragment read (LFR) technology introduced by Complete Genomics (www.completegenomics.com) (Peters et al. 2012). However, it does not attempt to assemble individual fragment since each fragment is sequenced at shallow coverage, but generates Linked-Reads amplified from the same DNA fragment with the same barcode. These Linked-Reads can be used to scaffold initial assembly sequences, identify large structural variations and do haplotype phasing. It is noteworthy that the 10X Genomics recently developed a software called Supernova to utilize Linked-

Reads for whole genome *de novo* assembly of diploids (Weisenfeld et al. 2016). With this new method, they assembled seven human genomes with contig N50 over 100kb and scaffold N50 close to 20 Mb. These assemblies showed significantly higher scaffold contiguity compared to previous PacBio assemblies, while only requiring modest coverage (~56x) on cheaper Illumina sequencing platforms. Notably, current 10X Linked-Reads data has only been successfully applied to human genome *de novo* assembly. It is still not clear whether it can effectively assemble other non-human genomes.

1.1.3 Assembly of heterozygous and polyploid genomes

Most of already sequenced plant genomes are performed on inbred or homozygous lines. In theory, the long read sequencing should be effective to assemble both haplotypes of a heterozygous diploid genome as long reads can span the repetitive and polymorphic regions. One recent study released a diploid-aware assembler FALCON and an associated haplotyping tool FALCON-Unzip, to assemble haplotype contigs (Chin et al. 2016). By using this method, an Arabidopsis F1 hybrid and a heterozygous grapevine accession were assembled with haplotype contig N50 of 6.9 and 0.8 Mb, respectively (Chin et al. 2016). Besides, long-range scaffolding technologies have also been used for inferring haplotypes including chromatin contact sequencing (Burton et al. 2013; Selvaraj et al. 2013; Putnam et al. 2016) or dilution-based haplotype methods (Kaper et al. 2013; Kitzman et al. 2011; Amini et al. 2014; Snyder et al. 2015). For example, 10X Genomics' Linked-Reads were utilized to assemble seven human genomes with the algorithms Supernova, reached phased scaffold N50 values of up to 9 Mb (Weisenfeld et al. 2016).

Additionally, substantial plant genomes are polyploid which can lead to similar challenges as heterozygous genomes. Similarly, algorithms dedicated for diploid haplotype assembly may be adjusted to assemble homeologous chromosomes of polyploidy genomes. So far, there are no algorithms tailored to assemble allopolyploids or autopolyploids. However, the assembly of several allopolyploids showed surprisingly high assembly contiguity even using the same assembly methods as for diploid genomes (Chalhoub et al. 2014; Yang et al. 2016; Zhang et al. 2015; Li et al. 2015a; Jarvis et al. 2017). For example, the recent assembly of *Chenopodium quinoa*, a 1.5 Gb allotetraploid, was assembled with scaffold N50 of 3.8 Mb by using PacBio SMRT sequencing, BioNano Optical Mapping and Dovetail Hi-C data (Jarvis et al. 2017). This suggests the homeologous chromosome in allopolyploid have sufficiently high sequence divergence to be distinguished and assembled separately by normal

assembly methods. Even the 17 Gb hexaploid wheat genome was sequenced on Illumina short-read sequencing platforms to cover the most of gene space.

Unlike allopolyploid, the homeologous chromosomes of autopolyploid plant genomes have high sequence similarity, making them difficult to be reconstructed separately. Previously, such plant genomes like potato were sequenced on diploid or haploid lines (Potato and Sequencing 2011). In theory, the algorithms designed for assembly haplotypes of diploid could be modified for autopolyploid genomes to phase polymorphic sites while the efficiency may depend on the level of sequence similarity of the homeologous chromosomes. There is no doubting that the long-read sequencing or long-range mapping will help bridging adjacent polymorphic sites and eventually reconstructing individual homeologs.

1.1.4 Final remarks

In conclusions, new third generation genomic technologies enable high-quality genome assembly. Integrating multiple long-read sequencing or long-range scaffolding technologies promise chromosome-level and near complete assembly for small genomes or even some large and highly repetitive genomes. There may be a trend towards generating multiple genome assemblies within the same species using third generation genomic technologies, which will influence current researches mainly based on one reference assembly and also challenge the computational analysis when multiple reference sequences are included (Schneeberger et al. 2009; Iqbal et al. 2012).

In the future, population-scale *de novo* assembly of each individual may emerge as resequencing based analysis often miss complex polymorphic genomic regions. Considering the current assembly costs, the 10X Genomics Chromium system is potentially to be applied on such population-scale of assembly. As the genomic technologies advance, the appearance of large panels of assemblies may end the resequencing era and enable direct whole-genome comparisons.

1.2 High-quality plant genome assembly using long-read sequencing and long-range scaffolding technologies

This section (1.2) was the basis a manuscript which was published into a research article in Genome Research (Jiao et al. 2017), which lists me as first author. All the analysis results not generated by me were not shown here. Data from my colleagues or my collaborators was clearly pointed out as described in the corresponding subsections of 3.1.

Authors list (Jiao et al. 2017):

Wen-Biao Jiao[1], Gonzalo Garcia Accinelli[2], Benjamin Hartwig[1], Christiane Kiefer[1], David Baker[2], Edouard Severing[1], Eva-Maria Willing[1], Mathieu Piednoel[1], Stefan Woetzel[1], Eva Madrid-Herrero[1], Bruno Huettel[3], Ulrike Hümann[1], Richard Reinhard[3], Marcus A. Koch[4], Daniel Swan[2], Bernardo Clavijo[2], George Coupland[1], Korbinian Schneeberger[1]

Author affiliations:

[1] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2] Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK. [3] Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. [4] Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, 69120 Heidelberg, Germany.

Authors contributions (Jiao et al. 2017):

Conceived and designed the project: KS, WBJ. Sample preparation: BHartwig, CK, SW, UH, MAK. PacBio sequencing: BHuettel, RR. Optical mapping: DB, DS. Genetic map: ES, EMH, GC. Data analysis: WBJ, GGA, EMW, MP, BC. Wrote the paper: KS, WBJ with help of all other authors.

In the past decade, the low-cost and short-read next generation sequencing technologies have allowed for the generation of thousands of eukaryotic genome assemblies, but frequently with low contiguity, which greatly limits genome-based downstream analyses. Recently two long-read sequencing technologies, namely Single Molecular Real Time (SMRT) sequencing from Pacific Biosciences (PacBio) (Eid et al. 2009) and nanopore sequencing from Oxford Nanopore, have been introduced to reach high-contiguity genome assembly. These two different technologies can generate reads with average length of up to 20kb, though the reads have error rates of nearly 15%, the final assembly accuracy is similar with the gold standard reference assembled from Sanger sequencing reads (Quick et al. 2014; Koren and Phillippy 2015; Berlin et al. 2015).

Long reads can span most of repetitive regions, which typically result in thousands of fragmented contigs in short read assemblies. Particularly, plant genomes frequently have high percent of transposon elements and present more repetitive k-mers compared to mammalian genomes (Nordström et al. 2013). Recently a few of high-quality genome assemblies of plants were achieved only by PacBio sequencing. These

included *Arabdopsis thaliana* (Landsberg erecta) with 38 contigs and an N50 of 11.2 Mb (Berlin et al. 2015) and *Oropetium thomaeum*, with 625 contigs and an N50 of 2.4 Mb (VanBuren et al. 2015).

To reach chromosome-level assembly, mapping information that order and orientate the contigs into the correct positions of chromosome is often necessary. In the past, this can be done using time-consuming genetic maps, but many contigs from heterochromatic regions cannot be anchored, as these regions lack meiotic recombination. Moreover, cytogenetic methods such as comparative chromosome painting are also applied to link contigs to their genome regions (Schranz et al. 2006); while they still require many experiments and have low resolution (Willing et al. 2015). In addition, reads pairs sequenced from the two ends of DNA fragment with roughly known size can also help anchoring contigs (Roach et al. 1995). Particularly, reads pairs from long DNA fragments such as BAC ends, can greatly improve the assembly contiguity.

Fortunately, two recently developed high throughput genomic technologies show remarkable improvements on assembly scaffolding and possibly reach chromosome-level genome assembly. One is optical mapping, which was invented more than twenty years ago (Schwartz et al. 1993) and was improved to high-throughput automatic platforms, like the Irys system introduced by BioNano Genomics (Tang et al. 2015). In this technology, restriction enzyme cutting site on long DNA fragments of several hundred kb can be imaged. The physical distances of adjacent cutting sites can be evaluated based on the pixel distance on image (Lam et al. 2012). Such order and distance information of restriction sites can be further assembled into genome-wide consensus maps, which can be then used to scaffold primary assembly contigs. Another technology developed by Dovetail Genomics, is called the *Chicago*, which is similar with the Hi-C technology, but simplified using *in vitro* reconstituted chromatin. Such sequencing read pairs can bridge genome regions separated by several kb to several hundred kb, which enables long-range scaffolding. It has been reported the assembly scaffold N50 of a human genome was improved from 33kb to 43 Mb. (Putnam et al. 2016).

However, it is still not clear whether the optical mapping and the Hi-C data can generate the similar improvement of assembly quality, as no direct comparison between assemblies on the same species is available. Besides, there are no comparisons of the same technology applied on the assembly of genomes with different structures or organizations. Furthermore, accurate chromosome-level assemblies are very important for comparative genomics studies like chromosome karyotype evolution. Therefore, in the first project of my thesis I will focus on these to

compare the performance of different long-range scaffolding technologies in different plant genomes and show how to approximate chromosome-level assemblies by integrating these technologies.

1.3 Genomic study on closely related plant species

As the cost of *de novo* assembly of genomes rapidly reduces, the community attempt to simultaneously or gradually sequence closely related plant species, shedding genomic insights on the evolution of important traits, such as genome size changes like it was shown in the comparison of *A. thaliana* and *A. lyrata* (Hu et al. 2011) or for mating system divergence in *Capsella* (Slotte et al. 2013). These closely related plant species frequently present substantial divergence in their phenotypes as well as genome sequences even after relatively short evolutionary time-scales. For example, *A. thaliana* and *A. lyrata*, which diverges only 10 million years ago, have considerable differences on genome size, chromosome karyotype, reproduction system and life cycle.

As the closely related sister species have different trait expressions, inter-specific experimental populations like recombinant inbreed lines (RILs) or introgression lines (ILs) are frequently constructed for breeding studies and quantitative trait locus (QTL) analyses. Inter-specific hybridization has been widely applied to plant species including some important crops (Zamir 2001; Baack and Rieseberg 2007), such as rice (Moncada et al. 2001), soybean (Wang et al. 2004) and tomato (Eshed and Zamir 1995). Typically, the genome of only of the parents is considered as genome assembly is still tedious. In the consequence not all of genetic information of the parents are covered. Particularly, when the two parental genomes have high level of genome sequences diverge such as sequence insertions, deletions, inversions or translocations, it will affect the normal sequencing-read-based analysis due to incorrect alignment of reads from genome rearranged regions (Qi et al. 2014). Furthermore, comprehensive characterization of genome sequence divergence requires high-quality chromosome-scale genome assembly of both parents. To achieve chromosome-scale assembly, researchers traditionally use methods including time-consuming genetic maps from RFLP markers or high throughput SNP markers (Lander and Botstein 1989; Baird et al. 2008; Elshire et al. 2011; Fierst 2015), cytogenetic maps from comparative chromosome painting (CCP) (Scherthan et al. 1994; Schranz et al. 2006; Slotte et al. 2013; Willing et al. 2015) or synteny with closed species genomes (Kim et al. 2013b; Tamazian et al. 2016; Damas et al. 2016). As a result of extraordinary advancement of third generation genomic technologies, chromosome-level assembly of plant genomes

will appear without using above traditional methods in the near future (Koren and Phillippy 2015) , although they still cannot assemble each chromosome sequence alone.

One of the most striking traits in plants is life cycle divergence between annuality and perenniality which occurs frequently among angiosperm species. Annual plants complete their life history within one year while perennial plants can live for many years. Interestingly, one species can exhibit annual and perennial ecotypes depending on the local environment, which has been observed among several species such as monkey flower (Hall and Willis 2006). Besides, closely related species can also exhibit differentiation of life-cycle (Datson et al. 2008; Karl and Koch 2013). For example, numerous genera in the family Brassicaceae contain both annual and perennial species, like *Arabidopsis*, *Brassica* and *Arabis*. However, the transition direction of annuality-perenniality is still controversial (Friedman and Rubin 2015), as both transition directions were indicated (Datson et al. 2008; Bena et al. 1998; Tank and Olmstead 2008; Stebbins 1957). A large tribe called Arabideae in the Brassicaceae encompasses wide and independent evolutionary transitions between annuality and perenniality (Karl and Koch 2013), which provides powerful resources for investigating the causes, consequences and directions of life cycle transitions within relatively evolutionary short time scales.

One of the perennial Arabideae species *Arabis alpina* (2n=2x=16) is used as a model to study genetic basis of perenniality and was the only assembled genome among the Arabideae (Willing et al. 2015; Jiao et al. 2017). Intriguingly, phylogenetic analysis has revealed a sister species *Arabis montbretiana* (2n=2x=16), which is annual and approximately diverged 3-5 Mya ago, while one of their common sisters *Arabis nordmanniana* is perennial (Karl and Koch 2013). Besides the life-cycle divergence, great genome structure differentiation might also occur during evolution, as the estimated genome size of *A. montbretiana* is 275 Mb (Hoffmann et al. 2010), and thereby much smaller than the 372 Mb as estimated for *A. alpina* (Lysak et al. 2009). However, comprehensive elucidation of the evolutionary life-cycle transition requires more genome assemblies and population genetic resources as highlighted in other studies on the ongoing transition from outcrossing to selfing as observed in *Capsella* (Slotte et al. 2013).

Therefore, in the second project “Comparative genomic and genotypic characterization of annual-perennial interspecific introgression lines constructed from two divergent *Arabis* species”, I will work on the comparative genomics of two divergent *Arabis* species, *A. alpina* and *A. montbretiana* and genotypic characterization of their inter-specific introgression population.

1.4 Thesis aims

My thesis includes two projects. The first project is mainly about how to obtain high quality genome assemblies of plant species using third generation genomic technologies of long-read sequencing, optical mapping and chromatin conformation capture sequencing. And the second project is “Comparative genomic and genotypic characterization of annual-perennial interspecific introgression lines constructed from two divergent *Arabis* species”.

In the first project, I will select three Brassicaceae genomes with different genome size and repetitive sequence percent for my research. I will present how to get initial high-quality genome assemblies using PacBio sequencing and also compare the assembly scaffolding performance of Bionano optical mapping and Dovetail Genomics’ chromatin capture data in different integrating strategies. And I will also try to integrate optical mapping and Dovetail scaffolding data to further increase assembly contiguity and accuracy. Apart from improvements of assembly contiguity, I will also focus on the assembly correctness. The accuracy of each intermediate assembly during initial assembling and scaffolding will be carefully evaluated and improved by utilizing Illumina short reads, Illumina mate-pair reads and a genetic map.

In the second project, I mainly characterize the genomic divergence and between two *Arabis* species and genotypic patterns among their interspecific introgression lines. To reveal the complex genetic mechanism underlying annuality-perenniality life-cycle divergence, our collaborators have constructed an introgression population by introgressing the *A. montbretiana* genome into the genome of *A. alpina*. To help to detangle genotypes correlating with the diversity on flowering related phenotypes, Genotyping by sequencing (GBS) was performed on an introgression population. Besides, as one reference genome sequence of parents cannot completely represent the parental genomic information and will affect downstream genomic data analysis, I will also attempt a chromosome-level assembly of *A. montbretiana* genome based on cytogenetic maps and GBS analysis. With these chromosome-level assembly of two *Arabis*, I can comprehensively characterize the genomic rearrangements emerging during evolutionary life-cycle transition and the genotypic patterns of population genotyping, which will be helpful for understanding the underlying mechanism of annuality-perenniality life-cycle transition.

2 Results

2.1 High-quality plant genome assembly using long-read sequencing and long-range scaffolding technologies

This results section (2.1) was the basis a manuscript which was published into a research article in Genome Research (Jiao et al. 2017), which lists me as first author. All the analysis results not generated by me were not shown here. Data from my colleagues or my collaborators was clearly pointed out as described in the corresponding subsections of 3.1.

Authors list (Jiao et al. 2017):
Wen-Biao Jiao[1], Gonzalo Garcia Accinelli[2], Benjamin Hartwig[1], Christiane Kiefer[1], David Baker[2], Edouard Severing[1], Eva-Maria Willing[1], Mathieu Piednoel[1], Stefan Woetzel[1], Eva Madrid-Herrero[1], Bruno Huettel[3], Ulrike Hümann[1], Richard Reinhard[3], Marcus A. Koch[4], Daniel Swan[2], Bernardo Clavijo[2], George Coupland[1], Korbinian Schneeberger[1]

Author affiliations:
[1] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2] Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK. [3] Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. [4] Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, 69120 Heidelberg, Germany.

Authors contributions (Jiao et al. 2017):
Conceived and designed the project: KS, WBJ. Sample preparation: BHartwig, CK, SW, UH, MAK. PacBio sequencing: BHuettel, RR. Optical mapping: DB, DS. Genetic map: ES, EMH, GC. Data analysis: WBJ, GGA, EMW, MP, BC. Wrote the paper: KS, WBJ with help of all other authors.

2.1.1 Long-read assembly of three plant genomes

Three diploid inbred plants from the Brassicaceae family including *A. alpina*, *Euclidium syriacum,* and *Conringia planisiliqua* were selected for genome *de novo* assembly. *A.*

alpina has been used as a model for studying perennial flowering. The other two species represent different phylogenetic lineages of Brassicaceae. For all of them, PacBio sequencing data were produced on PacBio RSII system using P6-C4 chemistry. The filtered subreads had an average length of 8.5 kb, 6.9 kb and 7.9 kb, and N50 length of 11.3 kb, 10.8 kb and 11.1 kb (Table 1, Fig. 2). The sequence coverage of them were about 86x, 47x and 54x, calculated based on the previously estimated genome sizes of 370 Mb, 262 Mb and 224 Mb (Hohmann et al. 2015).

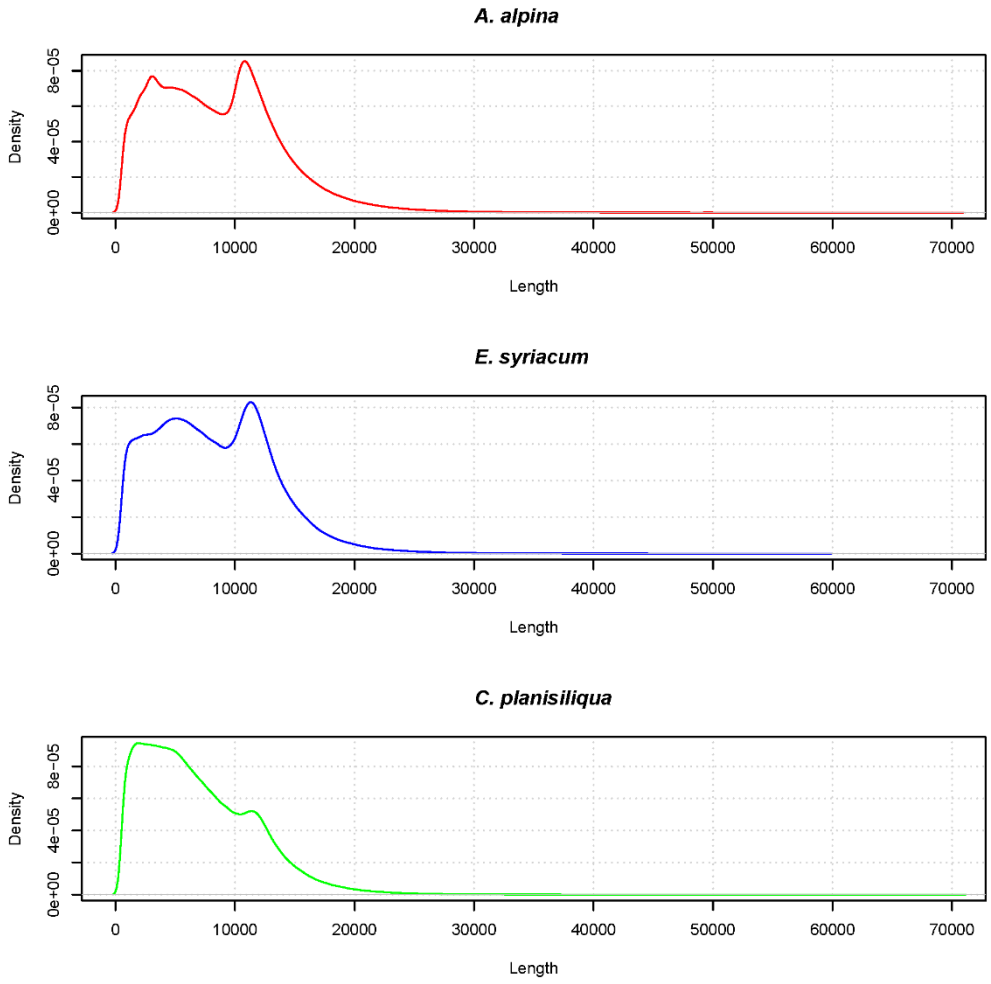


Figure 2 Length distribution of PacBio filtered subreads for the three genomes. (from Jiao et al. 2017)

Table 1 PacBio raw polymerase reads and filtered subreads statistics. (from Jiao et al. 2017)

	A. alpina		*E. syriacum*		*C. planisiliqua*	
	raw reads	subreads	raw reads	subreads	raw reads	subreads
SMRT Cells	35	35	30	30	18	18
Total bases (Gb)	38.4	32.1	14.7	12.3	12.7	12
Total number (M)	5.3	3.8	4.5	1.8	2.7	1.5
Length N50 (kb)	18.7	11.3	10.4	10.8	14.5	11.1
Length mean (kb)	7.3	8.5	3.3	6.9	4.7	7.9
Coverage	102.3	85.5	55.6	46.6	56.6	53.5

To do *de novo* assembly, two different tools Falcon (Chin et al. 2016) and PBcR (Berlin et al. 2015) were used, and followed by two polishing steps with raw PacBio subreads and Illumina paired-end reads, respectively (Table 2). For all three genomes, Falcon generated fewer contigs compared to PBcR, especially for *E. syriacum*, where the number of contigs assembled by Falcon was only about 25% of that assembled by PBcR. The total lengths of contigs from Falcon and PBcR were very similar in the assembly of *E. syriacum* and *C. planisiliqua*, while for *A. alpina* the assembly generated by Falcon revealed 19 Mb fewer sequences than PBcR. For the three genomes of *A. alpina*, *E. syriacum,* and *C. planisiliqua,* Falcon assembly had N50 values of 770kb, 3.3 Mb and 3.6 Mb (L50: 121, 14 and 14), while PBcR had N50 values of 914 kb, 975 kb and 1.5 Mb (L50: 99, 51 and 23) (Table 2, Fig. 3 A-C).

The assembly contiguity is normally characterized as contig or scaffolds N50 and L50. However, for genomes with multiple chromosomes, these statistics cannot come at their theoretical optimum. For example, L50 should be 1 in such case. In addition, they cannot reflect the assembly contiguity of individual chromosomes. To improve this, *chromosome-N50* (CN50) and *chromosome-L50* (L50) were introduced to estimate the median assembly contiguity (N50) of each chromosome assuming chromosomes have equal length and assembly quality. I found the CN50 of contigs was close to N50 of contigs, however, the CL50 could reach the optimum value of 1. For instance, the L50 of *C. planisiliqua* Falcon assembly was 14, while the CL50 was only 2. This CL50 number suggested that half of chromosomes were assembled into not more than two contigs, which could not be indicated from the L50 value alone.

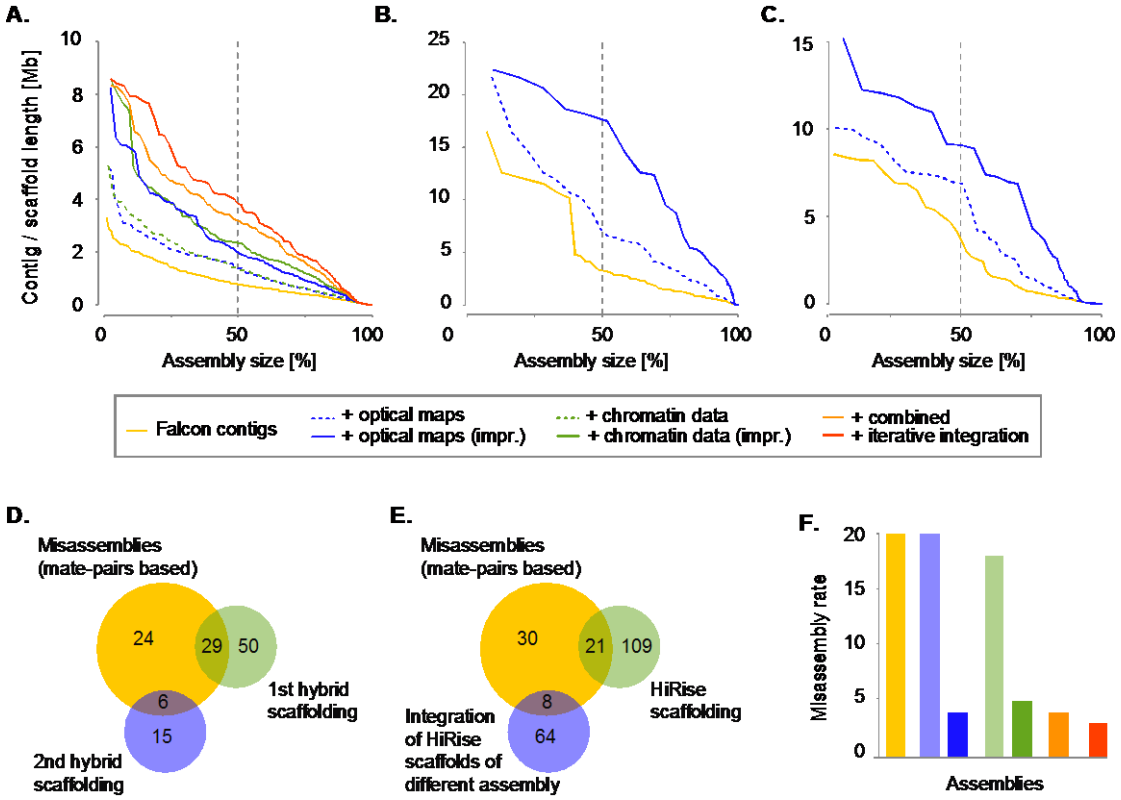


Figure 3 Assembly results and strategies. (from Jiao et al. 2017)

(A-C) Assembly contiguity of the assemblies of three species: *A. alpina*, *E. syriacum*, *C. planisiliqua*. The x-axis indicates the cumulative length of contigs sorted by length (expressed as percent of the entire assembly). The y-axis shows individual contig or scaffold length. The dashed line indicates the N50/L50 values. **(D)** Misassemblies identified by my improved scaffolding workflow for integrating optical maps (in two steps show in green and blue), and their overlap with Illumina mate-pairs (yellow) **(E)** Misassemblies identified by my improved scaffolding workflow for integrating Dovetail chromatin conformation capture data (also in two steps show in green and blue), and their overlap with Illumina mate-pairs (yellow). **(F)** Inter-chromosome misassemblies detected by a F2 genetic map in the seven assemblies (as shown in **(A)**).

Table 2 Assembly statistics (from Jiao et al. 2017)

	PacBio assembly		Optical mapping		Chromatin capture		Combined integration	
	Falcon	*PBcR*	*IrysSolve*	My workflow	*HiRise*	My workflow	Optical mapping + Chromatin capture	Iterative integration
A. alpina								
Assembly length [Mb]	328.2	347.1	332.6	336.3	328.2	329.6	337.0	337.0
Ambiguous bases [Mb]	0	0	4.9	9.4	0.03	2.1	10.4	10.6
Contig number (>10 kb)	1,204	2,074	1,044	900	995	901	841	817
N50 [Mb] / L50	0.8 / 121	0.9 / 99	1.4 / 75	2.3 / 46	1.3 / 72	2.0/ 47	3.2 / 36	3.8 / 31
CN50 [Mb] / CL50	0.8 / 16	0.9 / 13	1.4 / 10	2.4 / 6	1.4 / 9	2.1 / 6	3.2 / 5	4.0 / 4
Longest contig/scaffold [Mb]	3.3	6.2	5.2	8.6	5.3	8.3	8.3	8.6
Nucleotide error rate [%]*	0.0012	0.0008	-	-	-	-	-	-
Errors (mate-pairs)**	59	60	59	24	38	30	21	20
Errors (genetic map)	20	20	20	4	11	5	4	3
E. syriacum								
Assembly length	226.4	231.8	227.4	229.4	-	-	-	-
Ambiguous bases [Mb]	0	0	1.1	3.2				
Contig number (>10 kb)	228	944	168	119	-	-	-	-
N50 [Mb] / L50	3.3 / 14	1.0 / 51	6.5 / 10	17.5 / 6	-	-	-	-
CN50 [Mb] / CL50	3.7 / 2	0.9 / 7	6.5 / 2	18.7 / 1	-	-	-	-
Longest contig/scaffold [Mb]	16.5	7.7	21.6	22.4	-	-	-	-
Nucleotide error rate [%]*	0.0042	0.0045	-	-	-	-	-	-
C. planisiliqua								
Assembly length	177.7	175.2	179.2	184.3	-	-	-	-
Ambiguous bases [Mb]	0	0	1.8	6.9				
Contig number (>10 kb)	557	917	507	464	-	-	-	-
N50 [Mb] / L50	3.6 / 14	1.5 / 23	6.9 / 11	8.9 / 9	-	-	-	-
CN50 [Mb] / CL50	5.0 / 2	1.4 / 4	6.9 / 2	7.4 / 2	-	-	-	-
Longest contig/scaffold [Mb]	8.6	12.1	10.1	15.2	-	-	-	-
Nucleotide error rate [%]*	0.0065	0.0031	-	-	-	-	-	-

* Nucleotide errors were estimated with short read alignments. Errors were corrected after estimation of error rates. Nucleotide error rates of final assemblies are thus expected to be even lower as shown.

** Only mate-pair patterns in regions without thoroughly aligned optical consensus maps are shown.

2.1.2 Assembly quality and contiguity control

The PacBio assembly error rates at single nucleotide level were estimated using the Illumina short reads which were also used for genome assembly polishing. This estimation showed the error rates of all six assemblies before polishing were quite low (<0.01%, Table 3). Most of the errors found in both Falcon and PBcR assemblies were short InDels, which could be mostly caused by InDel-biased sequencing errors in PacBio raw reads and very few residual heterozygosity (*A. alpina*: 0.086%; *C. planisiliqua*: 0.061%; *E. syriacum*: 0.045%).

Table 3 PacBio assembly nucleotide-level accuracy estimation. (from Jiao et al. 2017)

	A. alpina		*E. syriacum*		*C. planisiliqua*	
	Falcon	PBcR	Falcon	PBcR	Falcon	PBcR
Mismatch	580	468	275	603	1,429	624
Indel	3,479	2,312	9,274	9,631	9,945	4,640
Error rate	0.0012%	0.0008%	0.0042%	0.0045%	0.0065%	0.0031%

Apart from single-nucleotide errors, I also used three Illumina mate-pair libraries with different average fragment sizes of 5 (Lib.1), 7 (Lib.2), and 10 kb (Lib.3) from *A. alpina* to estimate large-scale of misassemblies which linked two or multiple unlinked regions together as the example shown in Fig. 5 A . In total, 50.8 (Lib.1), 50.1 (Lib.2) and 26.5 (Lib.3) million reads were produced, 87.0%~94.2% of these reads could be mapped to Falcon or PBcR assembly contigs, 77.3%~89.7% of these mate-pairs could be both mapped to the assemblies and 19.6%~24.5% of aligned pairs were mapped on different contigs (Table 4). These both-mapped read pairs were checked to select those where could be mapped two different contigs and at least one of the paired reads was mapped to the contig's internal region (Table 4). Based on such read pairs, I found 59 and 60 large-scale misassembled regions in Falcon and PBcR contigs, respectively (Fig. 3 D, E). As incorrect mapping of mate pairs might introduce wrong distribution patterns and resulted in false identification of misassemblies, I also used a genetic map with 734 markers from *A. alpina*. These marker sequences were aligned to contigs of both Falcon and PBcR assembly, to identify inter-chromosome misassemblies by searching for contigs with multiple markers but from different linkage groups. This alignment analysis showed both Falcon and PBcR assembly had 20 inter-chromosome misassembled regions (on 19 Falcon contigs and 15 PBcR contigs), whereas they had no common misassembled regions (Fig. 3 F).

Table 4 Mate-pair library read statistics. (from Jiao et al. 2017)

	Reads	mapped reads		mapped pairs		inter-contig pairs	
		Falcon	PBcR	Falcon	PBcR	Falcon	PBcR
Lib. 1 (5 kb)	50,804,106	91.7%	94.2%	86.3%	89.7%	19.6%	23.9%
Lib. 2 (7 kb)	50,138,688	90.3%	92.7%	83.6%	86.9%	20.4%	24.3%
Lib. 3 (10kb)	26,492,772	87.0%	89.5%	77.3%	81.0%	20.5%	24.5%

Inter-contig pairs: read pairs mapped on different contigs

2.1.3 Optical mapping data integration

To scaffold the assembly contigs for each species, optical mapping data was generated on BioNano Iyrs system. In total, around 1.7, 0.8 and 0.5 million single molecule maps were produced for *A. alpina*, *E. syriacum,* and *C. planisiliqua* (Table 5, Fig. 4). These maps had an average length of 157, 145 and 200kb and represented 722, 446 and 410x physical coverage. These maps were further assembled by BioNano Genomic's IrysSolve software, resulting in consensus maps with N50 values of 625 kb, 924 kb and 1.5 Mb. These consensus maps were further aligned to the Falcon and PBcR contigs (Table 6). Although most of these maps could be confidently aligned, 79, 10 and 23 conflicts were observed in the alignments with the Falcon assembly contigs of *A. alpina*, *E. syriacum,* and *C. planisiliqua.* Similarly, 69, 41 and 25 conflicts were found in the alignments with the PBcR assembly contigs.

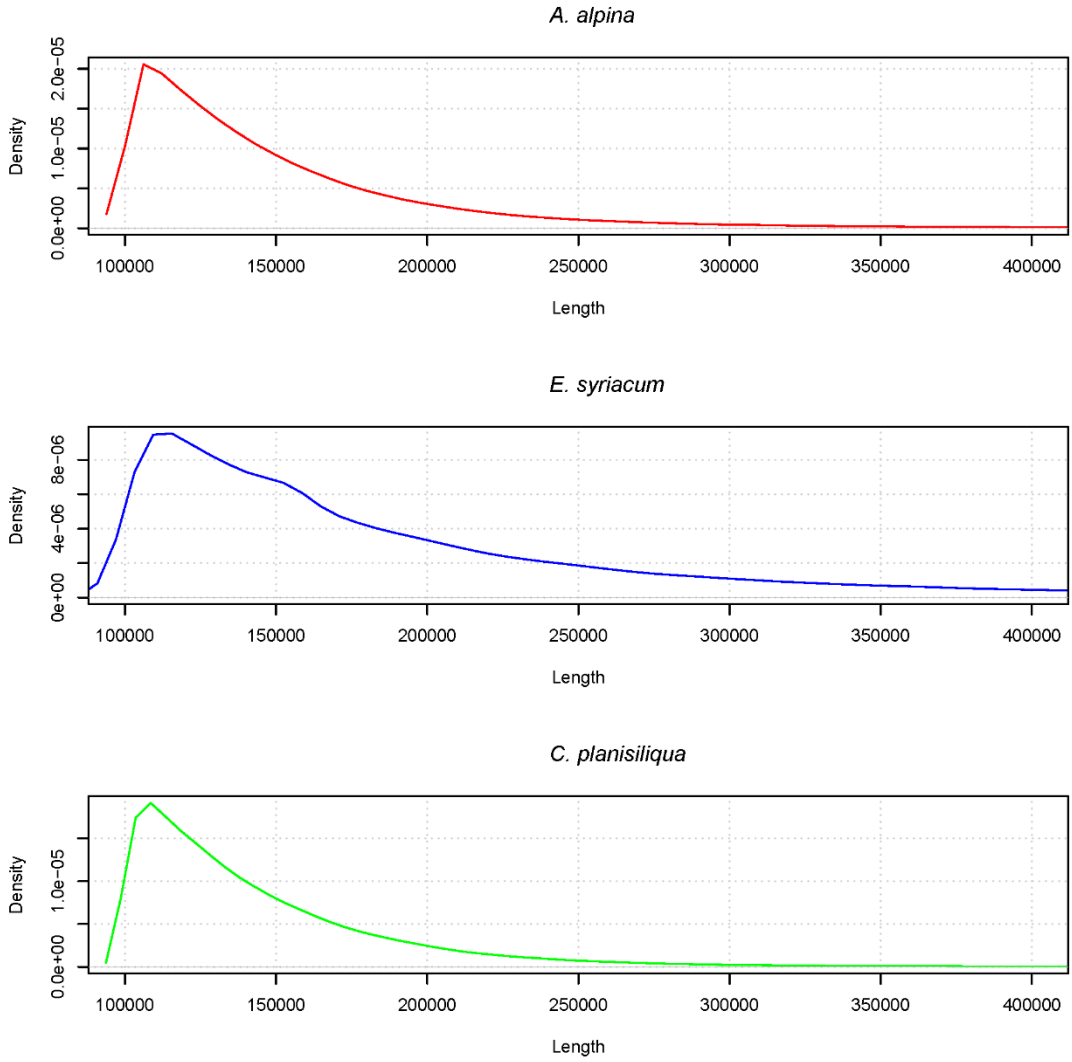


Figure 4 Length distribution of optical mapping molecules for the three genomes. (from Jiao et al. 2017)

Table 5 Optical mapping data and consensus map statistics. (from Jiao et al. 2017)

	Number of maps	Avg. map length(kb)	Coverage	Assembly size(Mb)	N50(kb)	L50	Nick sites / 100kb
A. alpina	1,729,537	157	722	322.8	624.6	166	9.6
E. syriacum	810,303	145	446	233.8	924.3	77	11.2
C. planisiliqua	461,383	200	410	199.7	1,474.2	41	12.3

Table 6 Consensus map (c-map) alignment statistics. (from Jiao et al. 2017)

	A. alpina		*E. syriacum*		*C. planisiliqua*	
	Falcon	PBcR	Falcon	PBcR	Falcon	PBcR
Aligned c-map number	601	604	318	315	156	152
Aligned c-map length (%)	97.3	97.7	97.5	97.1	90.1	89.2
Covered c-map length (%)	85.0	87.6	94.0	89.3	82.3	77.3
Aligned contig number	495	446	140	430	151	262
Aligned contig length (%)	91.2	87.2	98.8	93.8	92.8	89.8
Covered contig length (%)	77.9	75.8	94.1	87.5	89.7	85.7

Aligned c-map/contig length: the total length of consensus maps/contigs, which can be aligned by contigs/consensus maps. Covered c-maps/contigs length: the total length of consensus map/contig regions, which were covered by contigs/consensus maps.

In the workflow of BioNano's Irys software, all the conflicting consensus maps and assembly sequences are not considered for hybrid scaffolding. This scaffolding workflow could merge 253, 80 and 67 contigs and improved CN50 to 1.4 Mb, 6.5 Mb and 6.9 Mb in the three Falcon assemblies (Table 2). However, this workflow did not solve and utilize the conflicting alignments where potential misassemblies exist.

To improve this, I tried to determine whether the consensus maps or the contigs sequence in the conflicting alignments were misassembled (Fig. 5). First, I checked if two maps showed conflicting alignments with one contig at the same region, which indicated a misassembly of the contig (Fig. 5 C). Moreover, I checked whether the conflicting consensus map was well aligned with the contig from another assembler at the conflicting region, which also indicated the contig had a misassembly in this region (Fig. 5 D). However, when the contig from other assembly also presented the same conflict, the consensus map was considered to be misassembled. In the other cases that I could not determine which was misassembled, I just flagged both the consensus maps and contigs as potentially misassembled. Finally, I found 93% of the conflicts could be assigned as misassembly of sequences, while only 7% of conflicts were caused by misassembly of maps. These misassembled regions had significantly higher content of transposon elements, which suggested the repeats still challenged the long-read assembly and were not correctly assembled by current assembly tools. Similarly, across the three PBcR assemblies, 47, 23 and 35 misassembled regions in contigs were found and also enriched transposon elements (Fig. 6 and Table 7).

In my improved scaffolding workflow, these potentially misassembled sequences or maps were split at the misassembled regions. This workflow improved the scaffold CN50 of three Falcon assemblies to 1.6 Mb (*A. alpina*), 8.9 Mb (*E. syriacum*) and 7.4 Mb (*C. planisiliqua*). Similar assembly improvements were obtained by performing this workflow on PBcR assemblies.

Beside the increase of contiguity, it corrected 19 (95%) inter-chromosome misassemblies detected by the genetic map and 29 (49%) of large-scale misassemblies identified by mate-pair reads. Interestingly, this integration could be further improved. As hybrid consensus maps or called consensus scaffold maps from the hybrid scaffolding with one assembly might contain the connection information that were not included from the hybrid consensus maps scaffolding with another assembly. Therefore, I did a second round of hybrid scaffolding by integrating the PBcR-based hybrid consensus maps into the Falcon-based scaffold sequences (Fig. 5 B). After this final integration, the CN50 values reached 2.4 Mb (*A. alpina*), 18.7 Mb (*E. syriacum*) and 7.4 Mb (*C. planisiliqua*) (Table2, Fig. 3 A-C). The CL50 values were 6, 1 and 2 meaning that some of chromosome arms might be completely covered. In addition, 19 (95%) inter-chromosome misassemblies and 35 (59%) large-scale misassemblies were removed, suggesting my workflow could greatly increase assembly contiguity as well as assembly accuracy.

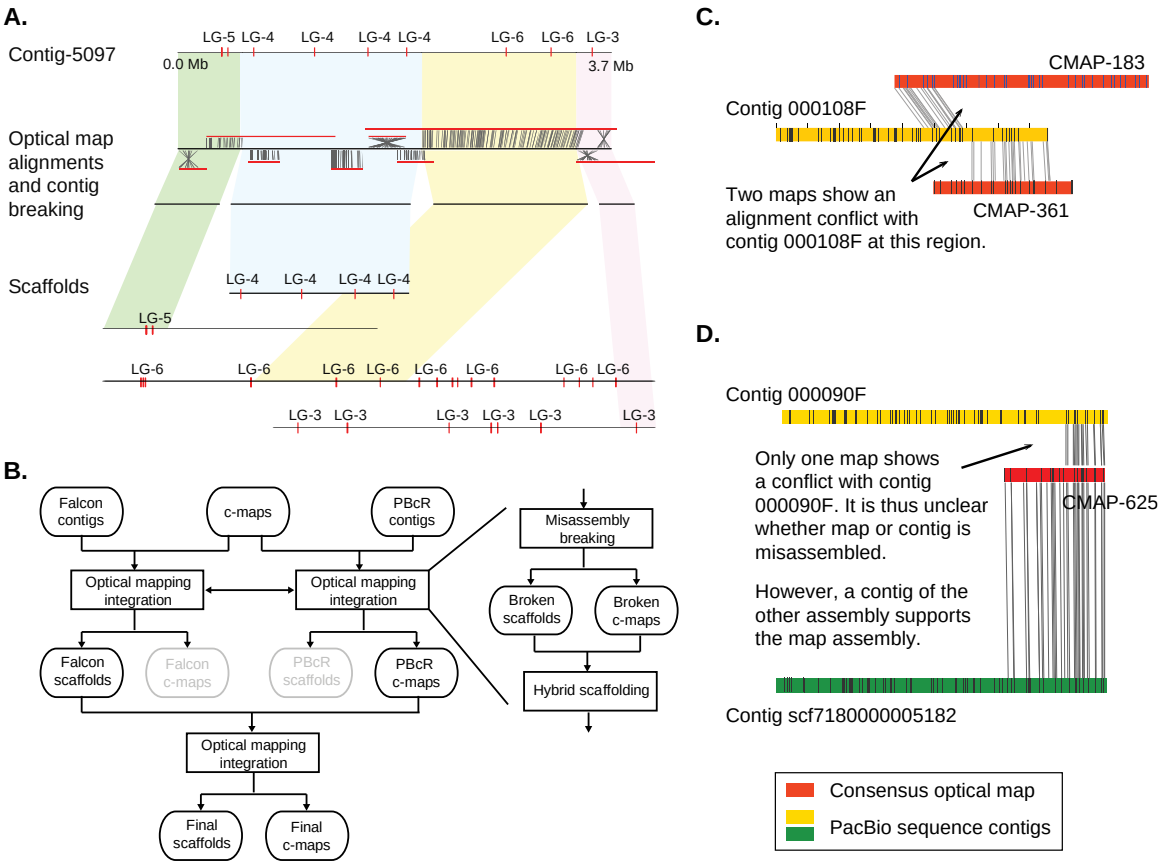


Figure 5 Optical mapping based assembly correction and scaffolding. (from Jiao et al. 2017) **(A)** Example of misassembly breakage and new scaffolding using optical mapping data. Three misassemblies in contig-5097 were detected based on the optical map alignments (and also validated by the genetic maps, markers shown with red ticks). The original contig was broken, and the subsequent scaffolding of the four contigs, which resulted from breaking the original contig at the misassemblies, assembled them into larger scaffolds, which were validated by the genetic map. LG:

linkage group. **(B)** Improved scaffolding workflow for integrating optical maps. The integration includes breaking of misassembled contigs and consensus maps (c-maps) and subsequent hybrid scaffolding between broken contigs and c-maps. **(C)** An example shows that the conflicts exist in alignments of two c-maps (CMAP-183 and CMAP-361) against Falcon contig 000108F, indicating a misassembled sequence. **(D)** The origin of the underlying misassembly is hard to be determined when a conflict only exists between one sequence (e.g. Falcon contig 000090F) and one c-map (CMAP-625). However, such a c-map might be fully aligned by a sequence from another assembly (e.g. PBcR contig scf7180000005182), suggesting the c-map (CMAP-625) is correctly assembled and thereby the sequence (000090F) is misassembled.

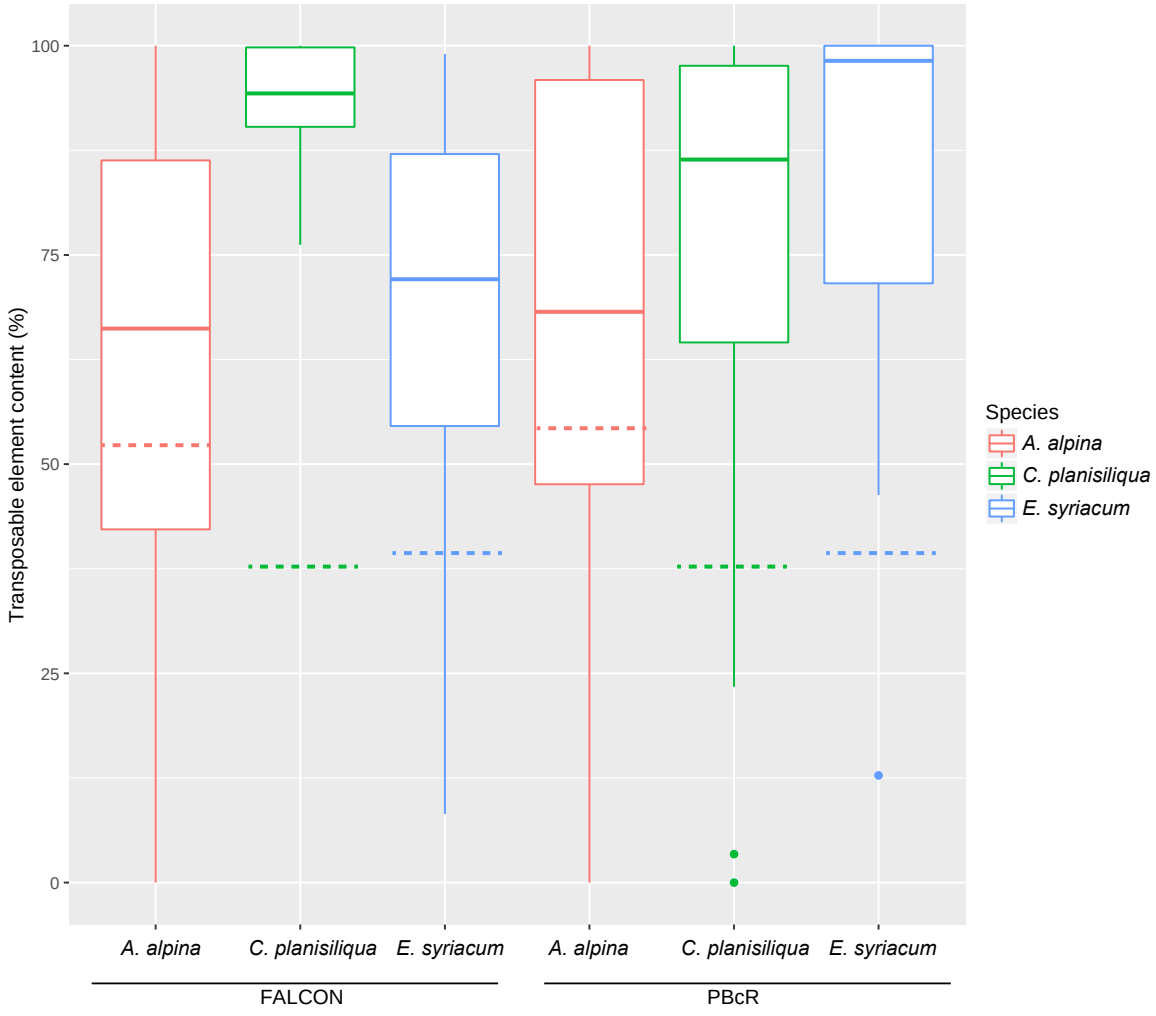


Figure 6 Distribution of transposable element content (%) in misassembled regions in the six initial assemblies (three species, two assembly tools) as compared to the average TE content density across the assembly (indicated with dashed lines)**.** (from Jiao et al. 2017)

Table 7 Misassembled regions are enriched for transposable elements (TEs). (from Jiao et al. 2017)

Species	Assembler	Misassemblies	TE-rich
A. alpina	Falcon	63	43 (68%)
A. alpina	PBcR	47	36 (77%)
C. planisiliqua	Falcon	15	15 (100%)
C. planisiliqua	PBcR	23	20 (87%)
E. syriacum	Falcon	7	6 (86%)
E. syriacum	PBcR	35	34 (97%)
Sum	**Falcon**	**85**	**64 (75%)**
Sum	**PBcR**	**105**	**90 (86%)**

Note: Misassemblies include all conflicting regions between optical mapping data and sequence contigs. TE-rich column describes how many of the misassembled regions harbor more TEs than the genome average.

2.1.4 Chromatin capture data integration

In addition to optical mapping data, chromatin conformation capture sequencing data for *A. alpina* genome was also generated. This data was produced in *in vitro* chromatin (Chicago library) from Dovetail Genomics. Read pairs of this sequencing are similar with reads from Hi-C sequencing, which have been applied for assembly scaffolding (Putnam et al. 2016). In total, 155.8 million read pairs were produced and nearly 39% and 40% of them could be mapped to the Falcon and PBcR assembly contigs. As expected, most of the aligned pairs in the same contigs were separated by less than 25kb. However, still some of them had distance of up to multiple hundred kb, including 1.3% of them with a distance over 25 kb (Fig. 7)

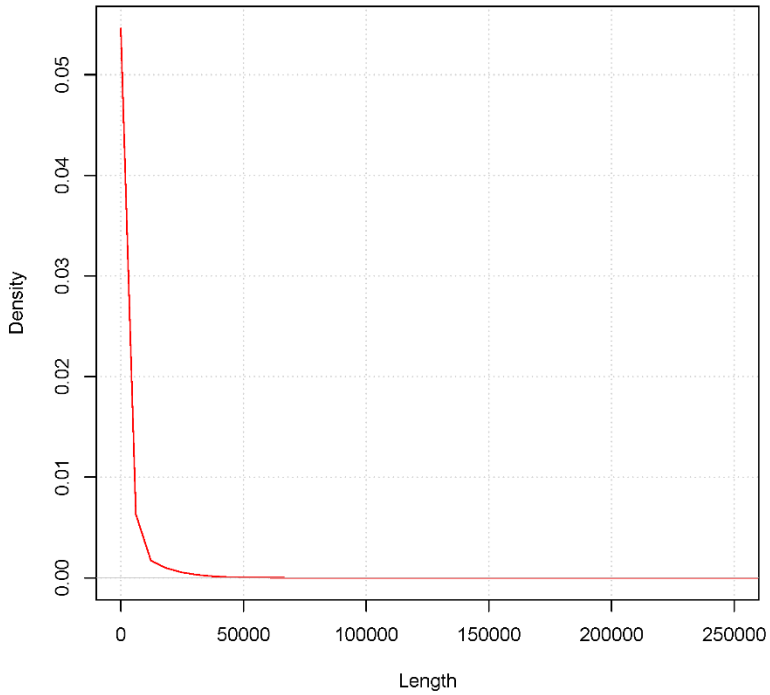


Figure 7 Insert size distribution of the Dovetail Genomics data of *A. alpina*. (from Jiao et al. 2017)

To do genome scaffolding, the software HiRise was firstly run. This scaffolding method can identify and split potentially misassembled sequences, and scaffold the error-corrected contigs iteratively using read pairs mapped to different contigs. After running this scaffolding, the CN50 of *A. alpina* Falcon assembly was improved from 771 kb to 1.4 Mb (Table 2). Again based on the genetic map, however, only four of the 20 inter-chromosome misassemblies were corrected and two additional such misassemblies were generated. While the number of misassemblies identified with mate-pair data decreased from 59 to 38.

Similar with above integration workflow using optical mapping data where two sets of hybrid maps from different assembly integrations were combined, I also attempted to improve the scaffolding based on chromatin conformation capture data. To do this, the HiRise PBcR scaffolds were transformed into artificial in-silico optical maps (Fig. 8) for doing hybrid scaffolding between these in-silico maps and HiRise Falcon-based scaffold sequences. After doing such improved scaffolding, the scaffolds CN50 reached 2.1 Mb. Furthermore, 19 of 20 those inter-chromosome misassemblies and 20 of the 59 those large-scale misassemblies were corrected during this integration (Fig. 3 E). However, four additional misassembles were also introduced due to partially complementary misassemblies between Falcon-bassed scaffolds and PBcR-based scaffolds.

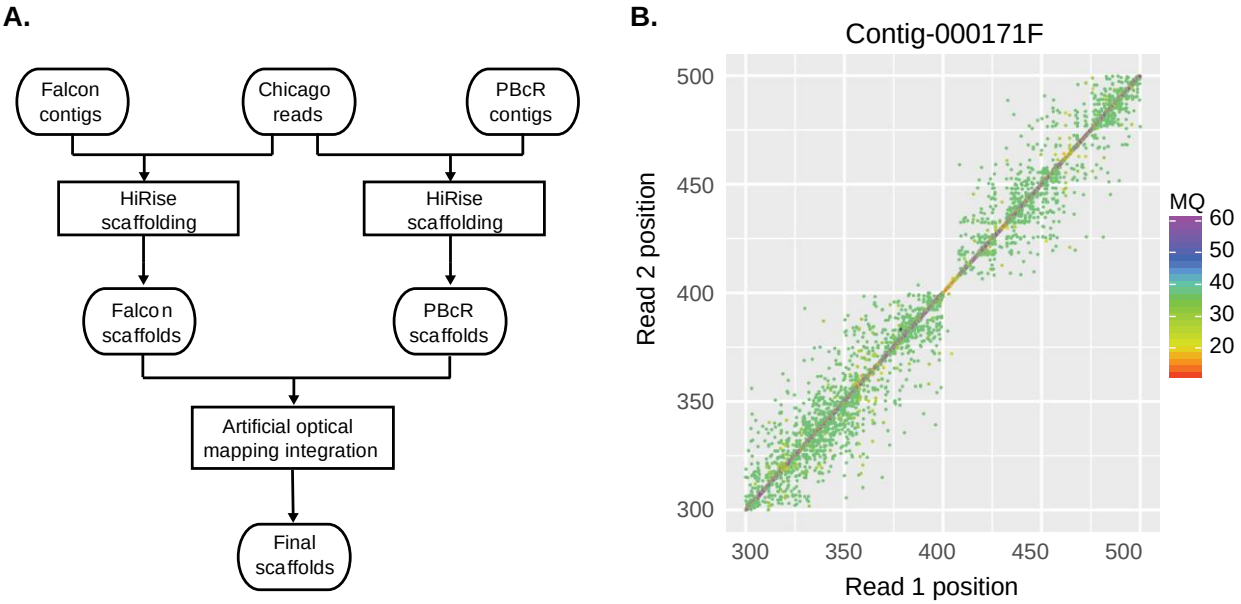


Figure 8 Assembly scaffolding using chromatin capture data. (from Jiao et al. 2017) **(A)** Improved scaffolding workflow for integrating chromatin conformation capture data. **(B)** Misassembly detection by mapping positions of read pairs of Dovetail chromatin capture sequencing. In the region 300-500 kb of Falcon contig 000171F, a region with absence of spanning read pairs was shown at around 410 kb. This misassembled region was identified by HiRise. MQ: mapping quality.

2.1.5 Comparing and combining optical mapping and chromatin capture data

As shown above, the independent integrations of optical mapping and chromatin conformation capture sequencing data had similar improvements on both Falcon and PBcR assemblies (Fig. 3 A). However, it does not mean they provide almost redundant scaffolding information. Considering these two technologies have different challenges in scaffolding, they may improve scaffolding even better when combined together. On the one hand, Dovetail Hi-C scaffolding cannot perform well in tandem repeat regions due to the inaccurate alignments of short Illumina short reads, while optical maps may span these regions. On the other hand, optical maps are often broken in those regions with closely linked restriction sites, but Dovetail Hi-C data is not affected by these sites (Pendleton et al. 2015).

Therefore, I further scaffolded the scaffolds from the optical mapping-based integration, based on the Dovetail Hi-C data using the HiRise tool. The CN50 value increased to 3.2 Mb while CL50 decreased to 5, implying the connecting information from these two technologies was partially complementary (Table 2). By examining the broken regions introduced by HiRise, I found HiRise was very conservative when finding potential misassemblies. Actually, some of misassembled regions found by HiRise were not true as the connections of these regions were

perfectly supported by optical consensus maps. Then I additionally integrated the hybrid consensus maps generated from the integration of optical mapping data, to re-scaffold falsely broken scaffolds during the previous step of HiRise scaffolding. After this iterative integration, the final assembly of *A. alpina* had CN50 value of 4.0 Mb, CL50 value of 4, and a longest scaffold of 8.6 Mb (Fig. 3 A, Table 2). Besides, all the 20 inter-chromosome misassemblies in the original Falcon contigs were corrected even though three new ones were also introduced during scaffolding. Simultaneously, 39 (66%) of large-scale misassemblies in these contigs identified by mate-pair data were also resolved.

2.1.6 Assembly of chromosomes

Chromosomes of plant genomes often have very long repetitive regions, especially centromere, telomere and rDNA cluster, which substantially challenge the assembly of whole chromosomes. Centromeres frequently contain tandem repeat arrays of short sequence with size up to hundred kb (Henikoff et al. 2001). To check whether my assembly scaffolds covered the centromeric regions, I firstly searched for tandem repeats that were highly abundant since such repeats are generally regarded as the candidates of centromeric repeats (Melters et al. 2013). Furthermore, as the *Arabidopsis lyrata* genome represents the ancestral karyotype of Brassicaceae family and has some centromeres conserved with other Brassicaceae species (Schranz et al. 2006), I also aligned each of three assemblies against *A. lyrata* genome to find syntenic regions flanking its centromere. In addition, as a typically descriptive pattern, the repeat and gene density along each scaffold were calculated. This is analyzed to check whether they match the typical feature that Brassicaceae chromosome has high repeat and low gene density around the centromere, while low repeat and high gene density at the ends of euchromatic regions.

The tandem repeat analysis revealed clear candidates for centromeric repeats in *A. alpina* and *C. planisiliqua*. Their tandem repeat monomers were 496 bp and 221 bp, respectively (Table 8). These monomers frequently arranged into higher order in the repeat arrays as observed previously (Melters et al. 2013). The centromeric repeat arrays in most of the scaffolds were closely linked into a cluster at the end of scaffolds or across the entire sequence, like one extreme example where one 690 kb scaffold of *A. alpina* had 23 centromeric repeat arrays distributed across its whole sequence. This again implied that centromeric region were hard to be assembled.

.

Table 8 Location of rDNA and centromeric repeat arrays. (from Jiao et al. 2017)

species	Scaffold	scaffold length	array start	array end	unit number	unit length	type
A. alpina	scaffold_113	318,510	3,526	69,721	114	119	5S
A. alpina	scaffold_397	24,131	45	23,774	49	119	5S
A. alpina	scaffold_443	21,612	206	21,553	42	119	5S
A. alpina	scaffold_648	15,052	168	14,847	31	119	5S
A. alpina	scaffold_364	26,159	395	25,761	22	119	5S
A. alpina	scaffold_358	26,695	338	26,560	19	119	5S
A. alpina	scaffold_838	9,551	495	9,402	19	119	5S
A. alpina	scaffold_867	8,656	373	8,463	18	119	5S
A. alpina	scaffold_935	7,005	139	6,807	14	119	5S
A. alpina	scaffold_958	6,069	31	5,611	12	119	5S
A. alpina	scaffold_986	5,253	172	5,200	11	119	5S
A. alpina	scaffold_1023	4,275	29	4,066	9	119	5S
A. alpina	scaffold_15_1	1,417,828	1,960	56,842	6	5,350	NOR
A. alpina	scaffold_310	31,895	5,871	19,752	2.3	5,350	NOR
A. alpina	scaffold_474	20,252	2,662	19,196	2	5,350	NOR
A. alpina	scaffold_740	12,257	2,269	12,244	2	5,350	NOR
C. planisiliqua	scaffold_309	16,420	230	16,255	32	119	5S
C. planisiliqua	scaffold_18	3,088,089	3,078,359	3,087,683	19	119	5S
C. planisiliqua	scaffold_55	94,367	7,779	93,989	10	5,353	NOR
C. planisiliqua	scaffold_107	45,594	1,302	44,719	5.7	5,353	NOR

C. planisiliqua	scaffold_186	26,997	746	26,982	3.3	5,353	NOR
C. planisiliqua	scaffold_276	18,011	213	17,658	2.7	5,353	NOR
C. planisiliqua	scaffold_319	15,920	1,422	15,260	2	5,353	NOR
E. syriacum	scaffold_89	15,746	25	15,694	34	119	5S
E. syriacum	scaffold_92	15,505	170	15,132	32	119	5S
E. syriacum	scaffold_10	9,524,166	4,937,560	4,998,160	7.3	5,352	NOR
A. alpina	scaffold_5	8,313,247	8,311,056	8,313,247	4.4	495	CENT
A. alpina	scaffold_9	7,192,857	7,119,231	7,121,434	2.2	992	CENT
A. alpina	scaffold_9	7,192,857	7,165,320	7,192,857	55.3	509	CENT
A. alpina	scaffold_9	7,192,857	7,172,465	7,192,857	41.3	495	CENT
A. alpina	scaffold_38	3,081,905	3,019,109	3,020,774	3.4	495	CENT
A. alpina	scaffold_45	2,631,477	1	11,579	11.7	992	CENT
A. alpina	scaffold_56	1,882,626	1	1,253	2.5	495	CENT
A. alpina	scaffold_56	1,882,626	17,725	19,905	4.4	496	CENT
A. alpina	scaffold_56	1,882,626	45,692	48,362	2.7	990	CENT
A. alpina	scaffold_76	977,799	958,286	960,453	4.4	496	CENT
A. alpina	scaffold_91	632,066	26,864	31,481	4.7	990	CENT
A. alpina	scaffold_91	632,066	67,829	69,937	2.1	992	CENT
A. alpina	scaffold_91	632,066	89,451	91,919	4.9	496	CENT
A. alpina	scaffold_91	632,066	137,083	139,689	2.6	990	CENT
A. alpina	scaffold_91	632,066	148,864	149,954	2.2	494	CENT
A. alpina	scaffold_91	632,066	198,453	200,567	4.3	496	CENT

A. alpina	scaffold_91	632,066	206,431	209,096	5.4	496	CENT
A. alpina	scaffold_91	632,066	214,409	217,369	6.0	496	CENT
A. alpina	scaffold_91	632,066	228,424	229,525	2.2	495	CENT
A. alpina	scaffold_91	632,066	275,256	276,293	2.1	495	CENT
A. alpina	scaffold_91	632,066	300,707	301,921	2.5	495	CENT
A. alpina	scaffold_91	632,066	317,899	320,453	2.6	992	CENT
A. alpina	scaffold_91	632,066	325,857	327,338	3.0	495	CENT
A. alpina	scaffold_91	632,066	350,460	352,447	4.0	496	CENT
A. alpina	scaffold_91	632,066	369,166	371,648	5.0	496	CENT
A. alpina	scaffold_91	632,066	400,514	407,036	6.6	992	CENT
A. alpina	scaffold_91	632,066	441,588	448,065	13.1	496	CENT
A. alpina	scaffold_91	632,066	487,579	490,436	5.8	495	CENT
A. alpina	scaffold_91	632,066	527,132	529,024	3.8	496	CENT
A. alpina	scaffold_91	632,066	540,161	541,320	2.3	495	CENT
A. alpina	scaffold_91	632,066	552,453	553,501	2.1	495	CENT
A. alpina	scaffold_91	632,066	564,545	565,927	2.8	496	CENT
A. alpina	scaffold_91	632,066	604,422	608,152	3.8	990	CENT
A. alpina	scaffold_95	568,052	547,850	550,320	5.0	496	CENT
A. alpina	scaffold_95	568,052	556,005	557,812	3.6	496	CENT
A. alpina	scaffold_95	568,052	558,740	566,921	16.6	494	CENT
A. alpina	scaffold_104	448,397	345,740	346,823	2.2	494	CENT
A. alpina	scaffold_104	448,397	372,076	374,246	4.4	497	CENT

A. alpina	scaffold_104	448,397	399,500	400,687	2.4	497	CENT
A. alpina	scaffold_104	448,397	442,174	443,663	3.0	493	CENT
A. alpina	scaffold_109	359,092	420	3,045	5.3	493	CENT
A. alpina	scaffold_109	359,092	8,217	13,055	4.9	989	CENT
C. planisiliqua	scaffold_1	15,208,799	10,089,605	10,090,827	5.5	221	CENT
C. planisiliqua	scaffold_1	15,208,799	10,113,773	10,116,480	2.6	1,059	CENT
C. planisiliqua	scaffold_1	15,208,799	10,182,906	10,183,676	3.5	221	CENT
C. planisiliqua	scaffold_1	15,208,799	10,202,604	10,204,237	3.7	441	CENT
C. planisiliqua	scaffold_1	15,208,799	10,215,506	10,218,151	6.0	442	CENT
C. planisiliqua	scaffold_1	15,208,799	10,234,820	10,236,456	7.4	221	CENT
C. planisiliqua	scaffold_1	15,208,799	10,813,225	10,814,534	6.0	221	CENT
C. planisiliqua	scaffold_1	15,208,799	15,178,432	15,182,366	17.8	221	CENT
C. planisiliqua	scaffold_3	12,074,320	12,072,687	12,074,320	7.4	221	CENT
C. planisiliqua	scaffold_4	11,819,561	10,876,803	10,879,849	6.9	442	CENT
C. planisiliqua	scaffold_4	11,819,561	10,898,921	10,901,008	9.5	221	CENT
C. planisiliqua	scaffold_4	11,819,561	11,786,835	11,788,801	8.9	221	CENT
C. planisiliqua	scaffold_4	11,819,561	11,795,725	11,798,056	3.5	663	CENT
C. planisiliqua	scaffold_4	11,819,561	11,816,473	11,817,406	4.2	221	CENT
C. planisiliqua	scaffold_5	11,279,646	10,044,908	10,048,569	16.6	221	CENT
C. planisiliqua	scaffold_6	10,977,244	6,616,706	6,618,339	3.7	444	CENT
C. planisiliqua	scaffold_6	10,977,244	6,637,799	6,652,393	65.8	223	CENT
C. planisiliqua	scaffold_6	10,977,244	6,662,541	6,665,249	12.3	221	CENT

C. planisiliqua	scaffold_6	10,977,244	7,771,294	7,772,850	2.3	665	CENT
C. planisiliqua	scaffold_6	10,977,244	10,951,673	10,977,244	116.4	221	CENT
C. planisiliqua	scaffold_12	6,996,948	6,985,637	6,989,964	19.5	221	CENT
C. planisiliqua	scaffold_12	6,996,948	6,994,638	6,996,948	3.5	662	CENT
C. planisiliqua	scaffold_14	5,442,335	22,323	25,214	13.1	221	CENT
C. planisiliqua	scaffold_14	5,442,335	50,727	51,662	2.1	441	CENT
C. planisiliqua	scaffold_14	5,442,335	3,457,147	3,458,221	4.9	220	CENT
C. planisiliqua	scaffold_17	3,665,691	3,662,919	3,665,691	4.2	663	CENT
C. planisiliqua	scaffold_18	3,088,089	545	2,783	5.1	439	CENT
C. planisiliqua	scaffold_18	3,088,089	3,127	7,527	20.1	220	CENT
C. planisiliqua	scaffold_18	3,088,089	11,140	12,990	8.4	221	CENT
C. planisiliqua	scaffold_21	1,970,407	1	25,718	38.9	662	CENT
C. planisiliqua	scaffold_21	1,970,407	125,391	143,691	10.4	1,765	CENT
C. planisiliqua	scaffold_21	1,970,407	290,182	291,688	3.4	442	CENT
C. planisiliqua	scaffold_21	1,970,407	317,023	317,943	4.2	221	CENT
C. planisiliqua	scaffold_21	1,970,407	499,090	499,748	3.0	220	CENT
C. planisiliqua	scaffold_23	1,380,778	2	1,418	6.4	221	CENT
C. planisiliqua	scaffold_23	1,380,778	7,537	35,224	41.7	659	CENT
C. planisiliqua	scaffold_23	1,380,778	276,925	280,180	3.7	882	CENT
C. planisiliqua	scaffold_23	1,380,778	292,068	294,546	5.7	442	CENT
C. planisiliqua	scaffold_37	230,006	21	1,790	2.7	656	CENT
C. planisiliqua	scaffold_37	230,006	13,495	18,680	23.5	221	CENT

C. planisiliqua	scaffold_37	230,006	131,068	143,114	54.6	221	CENT
C. planisiliqua	scaffold_37	230,006	144,218	152,347	7.4	1,104	CENT

5S: 5S rDNA arrays.

NOR: nucleolus organizer region, including 18S, 5.8S and 25S rDNA, only those with at least two units were shown.

CENT: putative centromeric repeat arrays. Only scaffolds more than 200 kb were shown.

Table 9 Location of telomeric repeat arrays. (from Jiao et al. 2017)

species	scaffold	scaffold length	array start	array end	unit number	unit sequence
A. alpina	scaffold_161	91,596	1	1,980	283.3	AAACCCT
A. alpina	scaffold_31	3,882,864	3,880,881	3,882,863	284.7	TAGGGTT
C. planisiliqua	scaffold_8	9,101,398	9,097,573	9,101,397	550	AGGGTTT
C. planisiliqua	scaffold_7	9,129,079	9,125,697	9,129,079	485.6	GTTTAGG
C. planisiliqua	scaffold_11	7,283,533	7,279,710	7,283,533	532.3	TAGGGTT
C. planisiliqua	scaffold_1	15,208,799	1	3,224	462.9	AAACCCT
C. planisiliqua	scaffold_2	12,270,481	1	2,816	402.3	AACCCTA
E. syriacum	scaffold_11	8,766,530	8,748,992	8,766,530	2509	TTTAGGG
E. syriacum	scaffold_12	6,520,592	6,510,080	6,520,592	1507.3	GTTTAGG
E. syriacum	scaffold_9	12,372,032	1	6,880	991.4	CCCTAAA
E. syriacum	scaffold_6	17,487,894	17,481,731	17,487,894	877.7	TTTAGGG
E. syriacum	scaffold_3	20,634,497	20,628,542	20,634,497	851.7	GGTTTAG
E. syriacum	scaffold_4	18,658,056	18,652,321	18,658,056	810.9	TTAGGGT
E. syriacum	scaffold_7	14,560,423	14,555,641	14,560,423	688.1	TTTAGGG
E. syriacum	scaffold_16	4,329,799	1	2,111	304.4	ACCCTAA
E. syriacum	scaffold_2	21,647,715	1	700	102.4	AAACCCT

In *E. syriacum*, I could not find obvious candidates of centromeric repeats as no tandem repeats had significantly high abundance. According to the sequence alignment (Fig. 9 A), I found that three scaffolds presented homology spanning the entire centromeric region of *A. lyrata*. For example, scaffold-2 could be aligned across the whole chromosome 2 and partial chromosome 1 of *A. lyrata*. However, the repeat density of scaffold-2 gradually increased from one end to another, but gene density decreased along the same direction. This did not match a typical pattern of the chromosome structure. Although this scaffold spanned an entire ancestral chromosome, this pattern much resembled a chromosome arm where chromosome rearrangements and centromere loss occurred during the evolution. Other two scaffolds had homology to whole ancestral centromeres (CEN3 and CEN4), but they had no typical distribution of gene and repeat densities for the chromosome organization.

Although there were no scaffolds with homology across complete ancestral centromeres in *A. alpina* and *C. planisiliqua*, the scaffold sequences of *C. planisiliqua* presented four large regions (on scaffold-3, -5, -6 and -14) where the centromeric repeats were estimated, without any homology to the *A. lyrata* genome. Besides, gene density was much lower but repeat density was much higher at one side of three regions (on scaffolds-3, -5 and -6) of them. All these together suggested that *C. planisiliqua* assembly covered partial centromeric sequences that were not assembled even in the gold reference assembly of *A. lyrata* genome (Fig. 9 B).

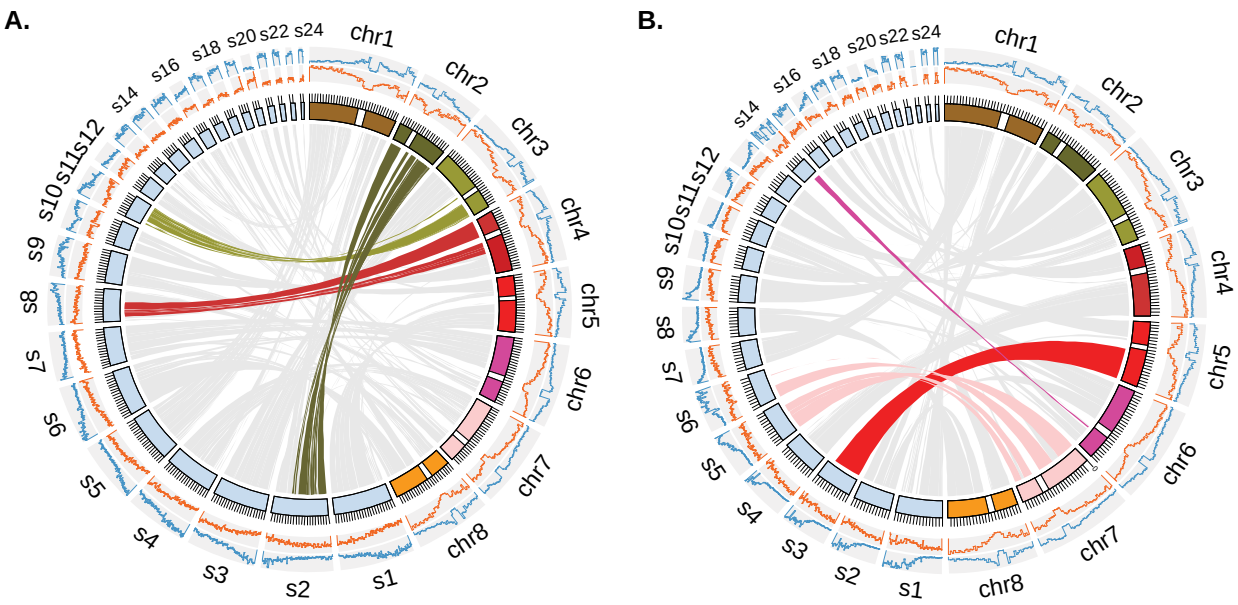


Figure 9 Comparing the assemblies of *E. syriacum* and *C. planisiliqua* to the ancestral karyotype present in the genome of *A. lyrata*. (from Jiao et al. 2017)

A. lyrata chromosomes are shown at the right side of each figure with colored blocks. White breaks in these blocks indicates their centromeres. Assembly scaffolds longer than 1 Mb are represented by light blue blocks. The two histograms outside of chromosome karyotypes show the gene (orange) and repeat (blue) densities calculated with window sizes of 1 Mb for *A. lyrata* and 200 kb for *E. syriacum* or *C. planisiliqua*. **(A)** Three scaffolds of *E. syriacum* have homology to the two flanking regions of *A. lyrata* CEN2, CEN3 and CEN4. **(B)** Scaffolds 3, 5, 6 and 14 cover up to 7 Mb of potential centromeric regions, which are not spanned in the core assembly of *A. lyrata*, as no homology parts are found in these regions.

Apart from the centromeric regions, I also searched for rDNA clusters in the three assemblies (Table 8). In *A. alpina*, *E. syriacum* and *C. planisiliqua*, I found 16, 3 and 7 rDNA clusters. The maximum unit number was 114 in a 5S rDNA cluster of *A. alpina* and the largest size was 86 kb found in a nucleolus organizer region of *C. planisiliqua*. Almost of the rDNA clusters were found across nearly the whole sequence of short scaffolds, while the others were at the end of scaffolds. This indicated that no cluster covered the whole sequence of the real rDNA cluster, also suggested that rDNA cluster repeats often resulted in fragmented assembly. In addition, I identified 2, 9 and 5

telomeric regions in the three assemblies (Table 9). The longest one was found in *E. syriacum* with more than 1,500 units of the short telomeric repeat sequence.

2.1.7 Assembly finalization and gene annotations

Finally, the *A. alpina* assembly scaffolds were ordered and orientated into eight pseudo-molecules according to the genetic map and cytogenetic maps from previous study (Willing et al. 2015). In total, 135 scaffolds with total length of 312 Mb were anchored into the chromosomes. Furthermore, protein-coding genes and transposon element were annotated for further study. Overall 29,740, 33,001, 34,766 protein-coding genes were predicated for *A. alpina*, *E. syriacum* and *C. planisiliqua* (Table 10). Additionally, 50.5%, 37.7% and 36.5% of their assembly scaffolds were annotated as transposon elements across the three genomes (Table 11).

To examine whether each assembly step affect the quality of the protein-coding gene annotation, I aligned the annotated genes of final assemblies to each intermediate assemblies to check alignment mismatches and gaps (Table 12 and Table 13). The PacBio assemblies before polishing using PacBio raw reads had considerable number of gene aligned with mismatches and InDels, while the assemblies after such polishing had only 0.6%~3.0% such imperfectly aligned genes. Such percent of these genes was negatively correlated with the depth of filtered raw subreads. Only 0.003-0.17% of genes were not perfectly aligned to the assemblies after correction using Illumina reads. These examinations suggested the importance of assembly corrections after *de novo* assembly of PacBio reads.

Table 10 Summary of protein-coding gene annotations. (from Jiao et al. 2017)

	A. alpina	*E. syriacum*	*C. planisiliqua*
Gene number	29,470	33,001	34,766
Total gene length	75,645,144	51,262,375	52,103,293
Gene region percent	23.2%	22.7%	29.4%
Coding region percent	10.5%	14.9%	19.2%

Table 11 Summary of transposable element annotations. (from Jiao et al. 2017)

	A. alpina		*E. syriacum*		*C. planisiliqua*	
	number of elements	percentage of sequence	number of elements	percentage of sequence	number of elements	percentage of sequence
SINE	5,996	0.36%	1,731	0.15%	441	0.09%
LINE	13,889	3.63%	8,872	2.46%	4,632	1.51%
LTR	78,148	29.01%	43,042	20.16%	27,599	18.53%
DNA	52,824	6.50%	27,592	5.21%	16,103	4.42%
Unclassified	80,500	10.98%	49,449	9.73%	26,625	11.98%
Total	231,357	50.48%	130,686	37.71%	75,400	36.53%

Table 12 Number and percent of perfectly aligned genes against each intermediate assembly. (from Jiao et al. 2017)

	PacBio raw	PacBio polished	Illumina corrected	1st OM scaffolded
A. alpina	13,512	29,294	29,420	29,423
	45.850%	99.403%	99.830%	99.841%
E. syriacum	11,982	32,035	33,000	33,000
	36.308%	97.073%	99.997%	99.997%
C. planisiliqua	13,809	33,956	34,765	34,765
	39.720%	97.670%	99.997%	99.997%

Table 13 Number of mismatches and alignment gaps of genes blasted against each intermediate assembly. (from Jiao et al. 2017)

		PacBio raw		PacBio polished		Illumina corrected		1st OM scaffolded	
		Mismatch	Gaps	mismatch	gaps	mismatch	Gaps	mismatch	gaps
A. a.	Genes	47,168	94,066	46	320	13	32	0	4
A. a.	Exons	15,296	31,286	5	60	0	22	0	1
E. s.	Genes	10,506	84,397	13	1,068	0	0	0	0
E. s.	Exons	6,391	45,581	11	486	0	0	0	0
C. p.	Genes	20,391	75,389	103	1,170	0	0	0	0
C. p.	Exons	14,841	43,875	85	732	0	0	0	0

A. a.: *A. alpina* E. s.: *E. syriacum*; C. p.: *C. planisiliqua*

2.2 Comparative genomic and genotypic characterization of annual-perennial interspecific introgression lines constructed from two divergent *Arabis* species

2.2.1 The assembly and annotation of *A. montbretiana*

Around 74 Gb Illumina paired-end and mate-pair reads from *A. montbretiana* were generated, representing 280 coverage of the genome (Table 15). The initial assembly sequences included 257.7 Mb assembled into 2,715 scaffolds, with contig N50 of 40 kb and scaffold N50 of 1.8 Mb (Table 14). An analysis of the k-mers in the reads approximated an estimated genome size of 265Mb, which was very close to the 275Mb estimation by flow cytometry. Our assembly spanned 94% of the estimated genome space, although it contained 67 Mb ambiguous sequences. A total of 232.2 Mb sequence organized in 173 scaffolds were further ordered and oriented into eight chromosomes based on the cytomolecular comparative maps (Fig. 10 A). To assist the gene annotation, 401 million Illumina RNA-seq reads were generated from different tissues sampled from seven growth stages (Table 16). By integrating evidences from *ab initio* prediction, homologous proteins and RNA-seq transcripts alignments, 29,917

protein-coding genes with 31,821 transcripts were annotated (Table 14). The number of annotated protein-coding genes within the *A. montbretiana* genome was around 1,000 genes less than in the perennial *A. alpina* and *A. lyrata* (Rawat et al. 2015), but nearly 3,000 more as compared to the annual *A. thaliana* (Lamesch et al. 2012), implying diploid annual species have less genes than their perennial sisters. The average gene length of *A. montbretiana* was 2.2 kb, smaller as compared to *A. alpina* (2.4 kb), but similar to *A. lyrata* (2.1 kb) and *A. thaliana* (2.2 kb). Finally, the completeness of our assembly and annotation were evaluated by the toolkit BUSCO with the plant linage specific single-copy core gene set. This assessment revealed 1,391 (96.6%) and 1,388 (96.3%) complete genes in the assembly and the annotation, respectively.

Table 14 Summary of the *A. montbretiana* genome assembly and annotation

	Number	Size
Scaffolds	2,715	257.7Mb
Contigs	13,528	191.1Mb
Scaffolds N50		1.8Mb
Contigs N50		40.0kb
Anchored scaffolds	173	232.2Mb
Gene models	29,917	64.1Mb
Transcripts	31,821	42.0Mb
Mean gene length		2.2kb

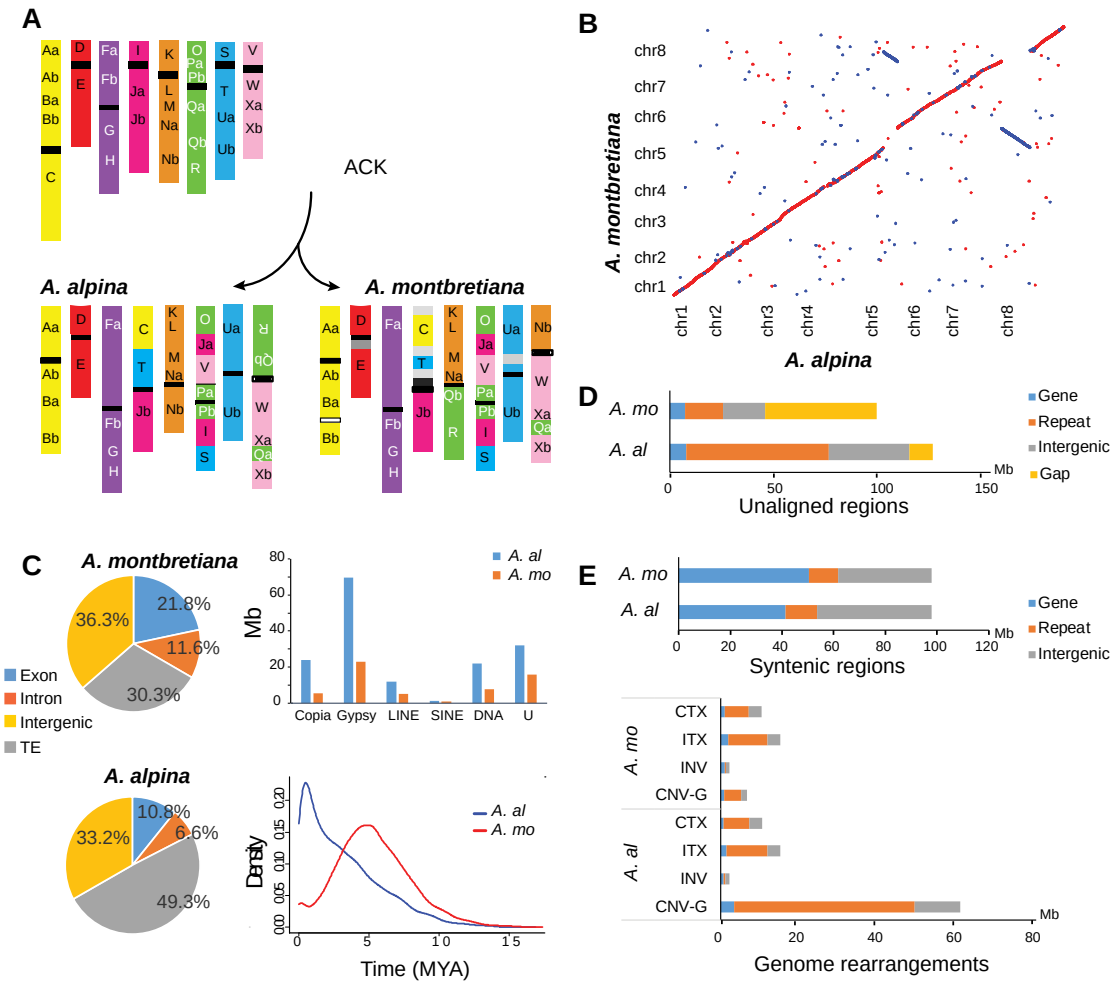


Figure 10 Genome comparisons between *A. montbretiana* and *A. alpina*.

(A) Karyotype evolution and comparison between *A. montbretiana* and *A. alpina*. ACK: Ancestral Crucifer Karyotype, defined in ref. (Schranz et al. 2006). **(B)** Dot plot of whole genome sequence alignment between *A. montbretiana* and *A. alpina*. Only alignment blocks with length more than 5 kb and identity larger than 90% are shown here. **(C)** Genome content comparison between *A. montbretiana* and *A. alpina*. The left pie charts show the genome content in *A. montbretiana* (top) and *A. alpina* (bottom). The top right histogram shows the total length of each transposon superfamily in *A. montbretiana*(red) and *A. alpina* (blue), the bottom right plot shows LTR(long terminal retrotransposon) insertion time distribution in *A. montbretiana* (red) and *A. alpina* (blue). **(D)** Genome content length of unaligned regions between *A. montbretiana* (red) and *A. alpina* (blue). LTR: long terminal retrotransposon; LINE: long interspersed nuclear element; SINE: short interspersed nuclear element; DNA: DNA transposon; OR: other repeats; N: assembly gap; INT: intergenic region. **(E)** Genome content and alignment types of aligned regions

between *A. montbretiana* (red) and *A. alpina* (blue). SYN: syntenic region; INV: inversion; ITX: intra-chromosome translocation; CTX: inter-chromosome translocations; CNV-G: gained copy number variation, namely species-specific duplications

2.2.2 Highly divergent genomes in *Arabis* annual-perennial species

In the last chapter I described an improved long-read genome assembly of *A. alpina* (Jiao et al. 2017), which allowed us to study the genome divergence of the two *Arabis* species. Like the annual *A. thaliana* and perennial *A. lyrata*, these two *Arabis* species also experienced large genome size and organization changes. *A. montbretiana* genome is nearly 100 Mb smaller than the *A. alpina* mainly due to the different transposon element (TE) richness, as TEs only make up 33.6% (64 Mb) of *A. montbretiana* genome but 53.0% (172 Mb) of the *A. alpina* genome (Fig. 10 C). Most of the TEs in *A. alpina* are from the two long terminal repeat (LTR) retrotransposon super families, Copia and Gypsy which occupy 18 Mb and 47 Mb more sequence in *A. alpina*, respectively (Fig. 10 C). This difference in these two *Arabis* is higher than that between the 125 Mb genome of *A. thaliana* and the 230 Mb genome of *A. lyrata*, which only consist of 13% and 31% TEs, respectively. Further analysis of the age distribution of intact LTRs showed *A. alpina* had significantly higher rates of recent retrotransposon activity. However, *A. montbretiana* has no such recent retrotransposon expansion with an insertion time peak around 5 Mya, close to the estimated species separation time. Despite the large difference of TE contents, their genic region sizes are close. Together, our findings suggest that *A. montbretiana* has a smaller genome because of the significantly lower TE activity as compared to *A. alpina*.

Apart from the great difference of genome size, the two *Arabis* genomes exhibit only a few large-scale karyotype changes (Fig. 10 A). The cytomolecular comparative map revealed a chromosome arm translocation between chromosome five and chromosome eight, and a centromere (CENT) repositioning on chromosome three in *A. montbretiana*, as compared with *A. alpina*. The chromosome arm translocation occurred after they separated from the common ancestors within the *A. montbretiana* lineage as *A. alpina* chromosome five was similar to the ACK. Besides, their

pericentromeric region distribution might also have dramatic differences according to the comparison of gene and repeat distribution along the chromosomes (Fig. 18, Fig. S1 - S7). Based on our previous genetic map of *A. alpina* (Jiao et al. 2017), I estimated that *A. alpina* had nearly 180 Mb pericentromeric regions, making up 58% of the assembly. As no genetic map for *A. montbretiana* exists, it is hard to its estimate pericentromeric regions in this species.

The cytomolecular comparative maps cannot identify moderate or small genome rearrangements due to the relatively low resolution of the BAC FISH experiments. To further compare the two genomes at higher resolution, I performed a whole genome sequence alignment (Fig. 10 B, D, E). In total, 186 Mb *A. alpina* and 131 MB *A. montbretiana* genome regions could be aligned with sequence identity larger than 90%. As expected, the large *A. alpina* genome presented much more copies in those duplicated alignment regions. Around 60.7 Mb (CNV-G) sequences in *A. alpina* had less copies in *A. montbretiana*, while only 4.8 Mb sequences in *A. montbretiana* originated from species-specific duplications (namely CNV-G) (Fig. 10 E). Most of the aligned sequences were from intergenic (Aa: 64 Mb, Am: 47 Mb) and genic (Aa: 48Mb, Am: 56 Mb) regions. The aligned genic regions of *A. montbretiana* and *A. alpina* made up 88% and 84% of their total genic spaces, respectively. The unaligned or highly differentiated regions were mostly repetitive (Aa: 67 Mb, Am: 18 Mb) or intergenic (Aa: 38 Mb, Am: 20 Mb) and included the assembly ambiguous nucleotides (Fig. 10 D).

Using the reference of the *A. alpina* genome, I defined syntenic blocks (SYNs) where were contiguously aligned to form a longest path of 98Mb, with the same order and orientation between *A. montbretiana* and *A. alpina*. Based on the 98 Mb syntenic path, I identified 1,279 (2.2 Mb) inversions (INVs), 13,752 (15.8 Mb) intra-chromosome translocations (ITXs) and 11,430 (10.4 Mb) inter-chromosome translocations (CTXs) (Fig. 10 E). Here, those continuously aligned regions only with small insertion and deletions were called as “contig”-blocks. Most of the SYN contig-blocks were from intergenic (44.3 Mb) or genic (41.3 Mb) regions, while most ITX and CTX contig-blocks were rich in repetitive sequences (ITX: 10.4Mb, CTX: 6.5Mb). Furthermore, those contig-blocks with same alignment type separated by unaligned or highly divergent

sequences could be connected to form a large “scaffold”-block, which probably emerged from single genome rearrangement event followed by rapid sequence divergence. However, small rearrangements could reoccur in the scaffold-blocks and resulted in separated scaffold-blocks. To examine these, the adjacent large scaffold-block separated by short scaffold-blocks were regarded as “super-scaffold”-block. Based on this definition, 40 super-scaffold genome rearrangements larger than 50 kb were identified including 22 INVs, 13 ITXs and 5 CTXs (Fig. 11). Altogether, these different levels of genomic rearrangements with nucleotide variations and short InDels contributed to the genome divergence between these two close *Arabis* sisters during life-cycle transition.

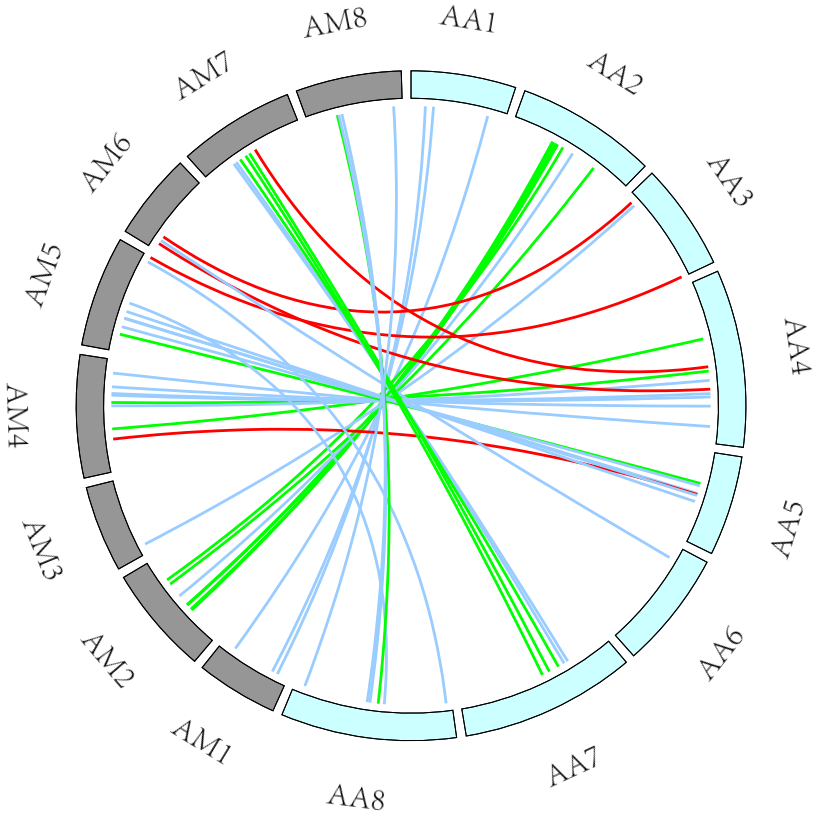
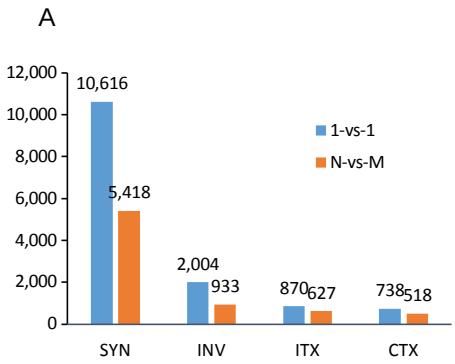


Figure 11 Large-scale genomic rearrangements between *A. montbretiana* and *A. alpina*.

Chromosomes are represented by grey (for *A. montbretiana*, AM1 - AM8) or light blue karyotypes (for *A. alpina* AA1 - AA8). Different types of rearrangements are indicated by differently colored links (INV: light blue; ITX: green; CTX: red). Only 40 ones larger than 50 kb were shown here.

To further check the variations in protein-coding genes, I firstly performed gene family clustering analysis. 17,855 gene families were clustered including 24,521 genes from *A. alpina* and 23,873 genes from *A. montbretiana*. 6,198 and 6,044 genes are potentially lineage-specific in *A. alpina* and *A. montbretiana*, respectively. These considerable number of specific genes might be overestimated due to incomplete assembly and annotation or ortholog assignment. Moreover, 1,483 gene families include different number of genes from *A. alpina* and *A. montbretiana*.

Based on these gene family clusters and syntenic regions defined by the whole genome alignments, I identified 16,304 syntenic and 5,690 non-syntenic gene pairs including 2,937 INV, 1,497 ITX, and 1,256 CTX gene pairs between *A. alpina* and *A. montbretiana*. Among of them, 10,616 syntenic and 3,672 non-syntenic gene pairs were from one-to-one orthologous families (Fig. 12. A). Syntenic gene pairs exhibited significantly higher protein sequence alignment identity and lower *Ka/Ks* ratio, as compared to ITX and CTX gene pairs (Mann–Whitney U test, p-value < 2.2e-16, Figure 12 C, D). While the INV gene pairs showed similar distribution of protein sequence alignment identity and *Ka/Ks* ratio. Interestingly, seven ortholog pairs whose orthologs in *A. thaliana* are flower control related genes locate in different homologous chromosomes between the two *Arabis* genomes. Four pairs are one-to-one orthologs, whose orthologs in *A. thaliana* are AT3G04610, AT4G20370, AT5G61850 and AT5G65070, respectively (Fig. 12 B). For example, in the ortholog pair Aa_G61540 – Am_236410, the *A. alpina* gene is on chromosome 8 while *A. montbretiana* gene is on chromosome 3.



B

Inter-chromosome translocated flower-related ortholog genes

A. alpina	*A. a* chr.	*A. montbretiana*	*A. m* chr.	*A. thaliana*
Aa_G4280	chr3	Am_395690	chr6	AT3G04610
Aa_G61540	chr8	Am_236410	chr3	AT5G61850
Aa_G585320	chr8	Am_374070	chr5	AT5G65070
Aa_G202710	chr2	Am_218830	chr3	AT1G62830; AT3G10390; AT3G13682

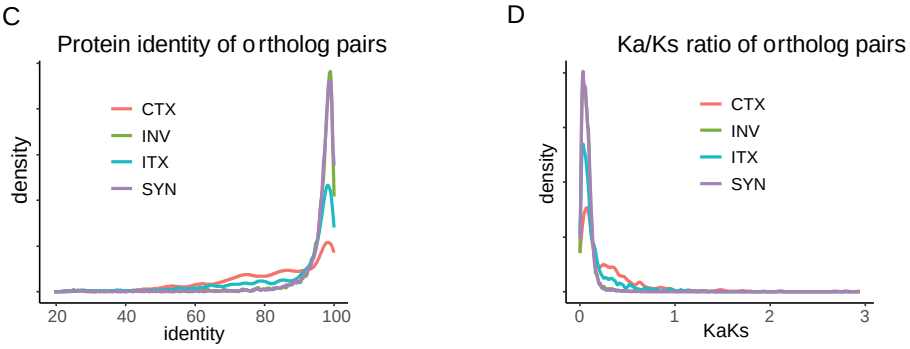


Figure 12: Gene comparison between *A. montbretiana* and *A. alpina*.

(A) Histogram of syntenic (SYN) and rearranged (INV, ITX, CTX) gene pairs from one-to-one orthologous gene families or the other not one-to-one (N-vs-M) orthologous gene families. **(B)** Examples of flower- related CTX genes. **(C)** Distribution of protein sequence alignment identity of ortholog gene pairs. **(D)** *Ka/Ks* distribution of ortholog gene pairs

2.2.3 Genotyping of *Arabis* annual-perennial introgression lines

To study the genetic basis of the annual-perennial life cycle switch, our collaborators constructed interspecific introgression lines by introgressing genomic segments of *A. montbretiana* into the recurrent parent *A. alpina* (Fig. 13). The offspring of the backcrossing (BC1) were further used to do one, two or three round of self-crossing (BC1S1, BC1S2, BC1S3). The whole population including the two parents and 460 progenies were sequenced using the genotyping-by-sequencing (GBS) method (Elshire et al. 2011), resulting in an average read number of 1.2 million per sample. After removing low quality samples, the population of 416 progenies were from nine BC1S1 families and seven BC1S2 families with 15 to 35 individuals.

By utilizing the read alignments from the parental genomes against the *A. alpina* reference genome, 6,431 markers were selected for genotyping. These markers had

relatively even distribution along the chromosomes except for the centromeric regions and chromosome terminals, where there was under and over representation respectively (Fig. 14), with two markers per 100kb on average. According to the whole genome alignment described above, 5,977 markers were found in SYN regions, while 164, 82, and 120 markers were in INV, ITX, and CTX regions, respectively. The remaining 88 markers were on the scaffolds, which were not anchored to the chromosomes.

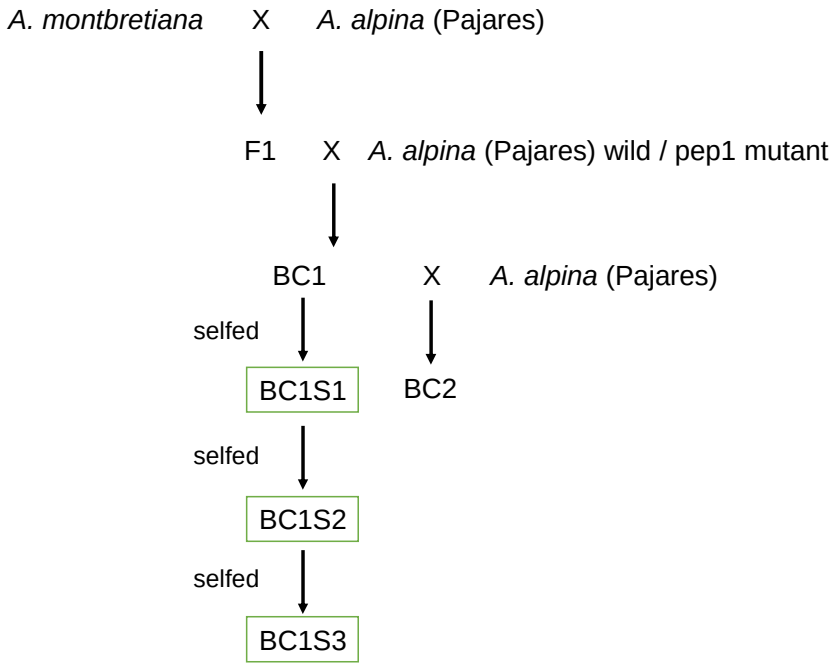


Figure 13 Schematic diagram for the introgression of *A. montbretiana* into *A. alpina* genetic background.

The parental species are *A. alpina* (accession: Pajares) and *A. montbretiana*. The progenies from the families marked by green boxes were sampled and sequenced using the GBS method. BC: backcrossing. S: selfcrossing

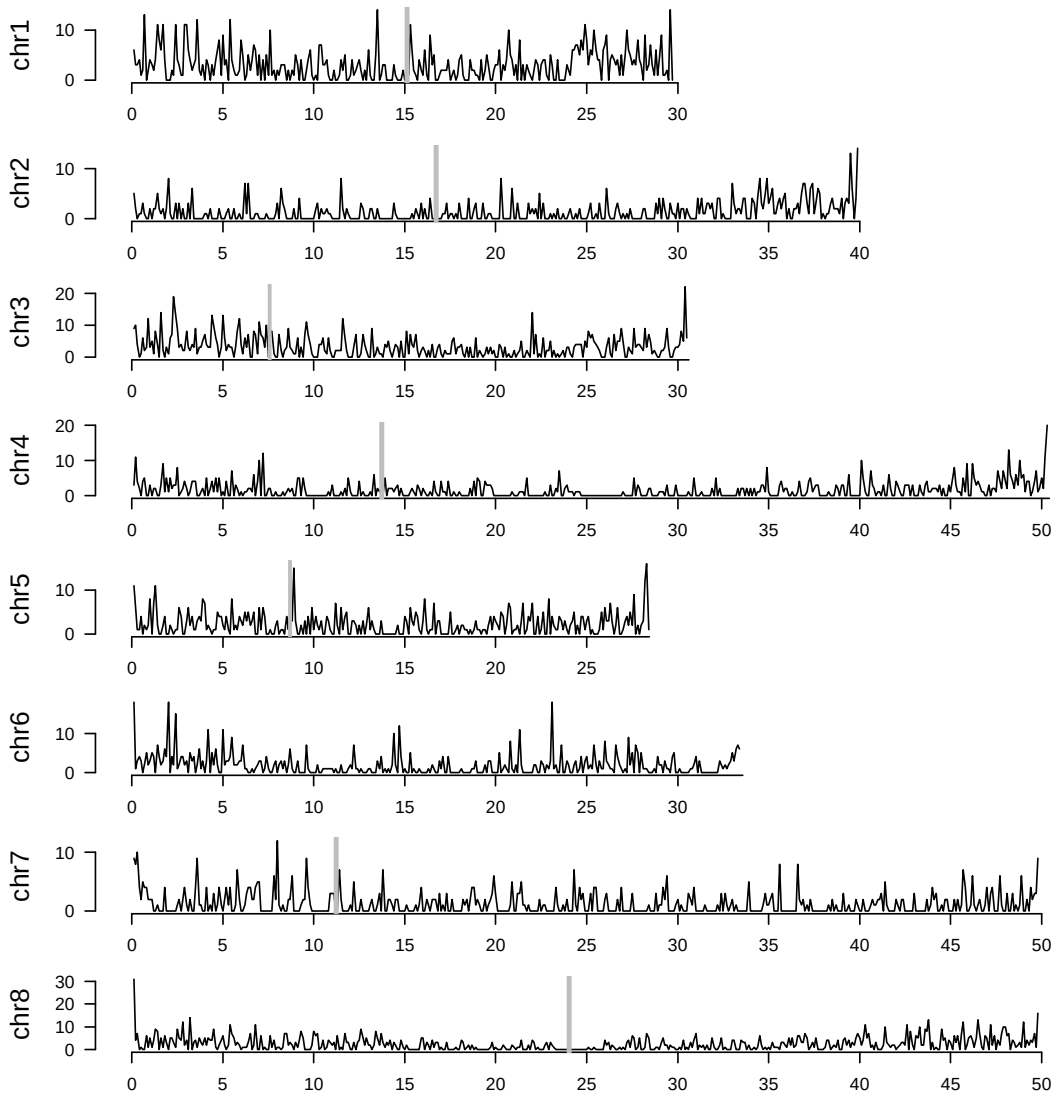


Figure 14 Marker distribution along the chromosome.

X axis corresponds to the genome position, y axis indicates the marker number. The marker number was calculated using a 1 Mb sliding window with step size of 100 kb. The grey bars show the centromere positions.

Finally, 175 Mb (56%) regions of both alleles in *A. alpina* genome were replaced. While 120 Mb (38%) of *A. alpina* genome being exchanged by a single allele of *A. montbretiana*. The introgressed regions had different sizes, including a region, which spanned nearly the entire chromosome 6, and showed nearly complete fixation for the *A. montbretiana* genome within a single line. In contrast, only 24% of chromosome 4 were introgressed into homozygous *A. montbretiana* genetic background. Interestingly, there was one family which included various individuals with extended heterozygosity

within centromeric region of chromosome 8 (CENT8), while other individuals of this family (and all other individuals of all other families) were fixed for *A. alpina* in this region, but none of the individuals show fixation of the *A. montbretiana* CENT8. Intriguingly, this pattern was perfectly linked to heterozygosity on CENT5 within this family (Fig. 15 A,B) and presumably relates to the chromosome arm translocation between the two chromosomes suggesting that it is not possible to fix either of the centromeres without fixing the other centromere for the same parental genome. Moreover, as I have not observed fixed *A. montbretiana* CENT5 and CENT8 regions it remains unclear whether both centromeres could be introgressed, or if these can only be heterozygous or fixed for *A. alpina* (Fig. 15 C). Interestingly, however, this dependence of the CENT5 and CENT8 does not imply that the chromosome arms of these chromosomes cannot recombine as parts of chromosome 5 and 8 were fixed for the *A. montbretiana* alleles.

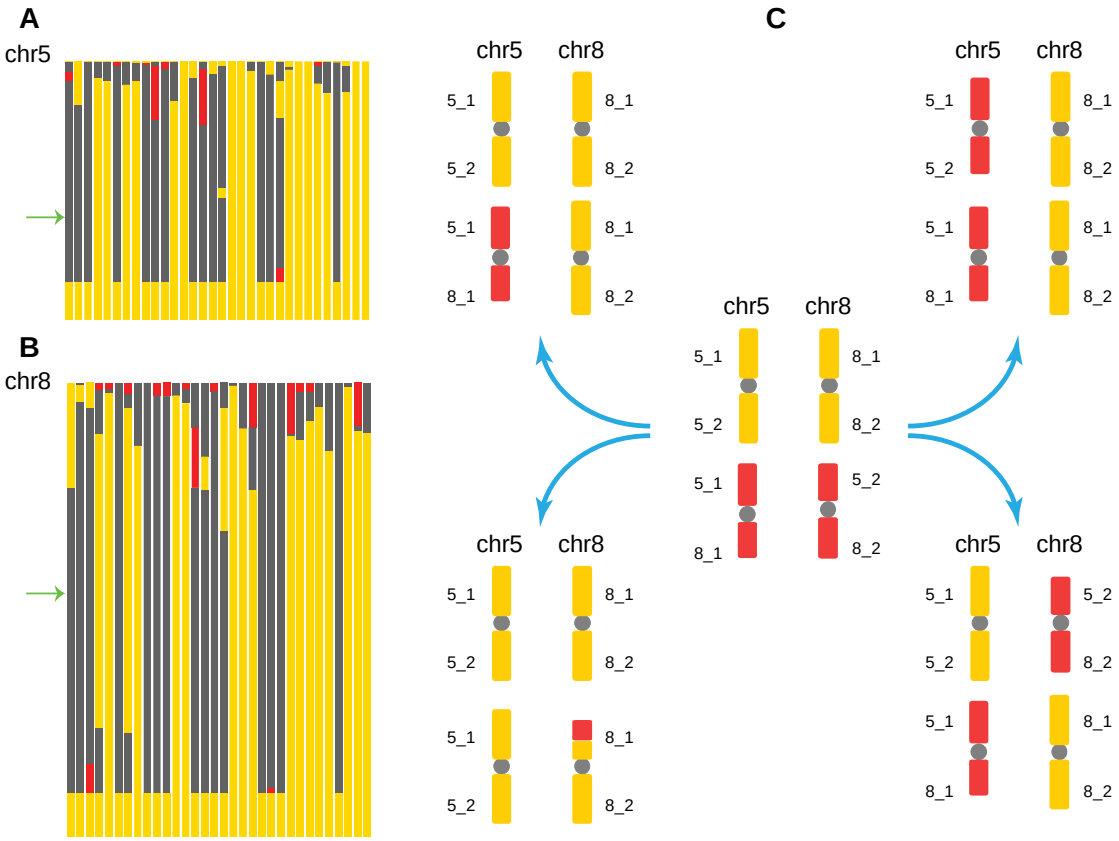


Figure 15 Recombination between *A. alpina* chromosome 5, 8 and *A. montbretiana* chromosome 5, 8.

(A, B) Recombination maps of chromosome 5 and 8 in family BC1_4_S1. Genotypes are indicated by different colors, yellow: homozygous *A. alpina*, red: homozygous *A. montbretiana*, grey: heterozygous. The centromere positions were marked by the green arrows. **(C)** The schematic diagram shows examples for unbalanced (top left and top right) or balanced (bottom left and bottom right) chromosome arm arrangements, which could lead to copy number changes of chromosome arms. The middle part represents the karyotype of chromosome 5 and 8 in *A. alpina* (yellow bars) and *A. montbretiana* (red bars). The chromosome arms were named after the ancestral karyotype. The grey circles present the positions of centromeres. CENT5: chromosome 5 centromere, CENT8: chromosome 8 centromere.

In total, I detected 3,811 recombination breakpoints across all lines (Fig. 16). As our previous study suggested *A. alpina* had very high percent of pericentromeric regions which resulted in large regions with suppressed recombination. I asked whether the *A. montbretiana* presented smaller pericentromeric regions and whether this feature would influence inter-species recombination events during meiosis. Although I had no direct evidence such as recombination maps or chromatin epigenetic marks to show where the pericentromeric regions is located in *A. montbretiana*, I estimated the extend of the peri-centromere following the typical distribution of low gene and high repeat density. I compared the gene and repeat density along the chromosome between the two *Arabis* genome. I found *A. montbretiana* obviously had smaller regions with high repeat density, which suggests that *A. montbretiana* also has smaller pericentromeric regions. Further, I found 273 recombination breakpoints in these putative pericentromeric regions (Fig. 18, Fig. S1-S7). In particular, recombination on chromosome 1 and 6 were highly increased with 30 and 19 different recombination breakpoints for both of them (Fig. 17, Fig. 18, and Fig. S5). While other chromosomes such as chromosome 2, 4, 5 and 8 had less than ten different breakpoint positions in their putative peri-centromeres (Fig. 17, Fig. S1, Fig. S3, Fig. S4, and Fig. S6). This result indicated that the heterogeneous pericentromeric structures among different chromosome influenced the recombination events in the pericentromeric

regions varyingly. As no genome-wide landscape of recombination in *A. montbretiana* F2 was available and also our introgression lines were not tailored for detecting recombination events between the two *Arabis* species, I could not conclude whether *A. montbretiana* could significantly increase the recombination frequency in pericentromeric regions within the inter-species offspring.


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
A.alpina
A.montbretiana
Heterozygous

Figure 16 Recombination map.

Yellow: homozygous *A. alpina*, Red: homozygous *A. montbretiana*, Grey: heterozygous. Each column corresponds one progeny. Progenies from the same family are drawn close. Different families are separated by a small distance. The left part includes nine BC1F2 families and the right part includes seven BC1F3 families

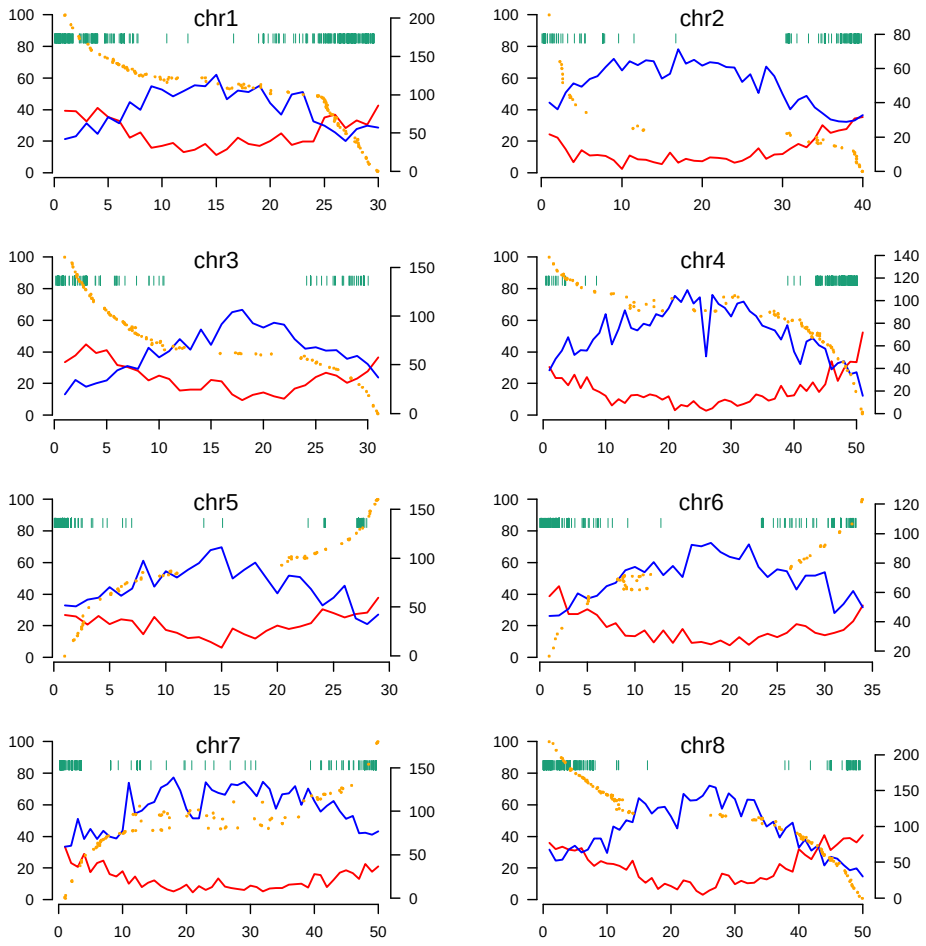


Figure 17 Recombination breakpoint distribution along the chromosomes.

Green bars indicate the recombination breakpoints. Blue and red curves show the gene and repeat density, respectively. The percentage was calculated using a 1 Mb sliding window with a step size of 100 kb. X axis corresponds to the genome position of *A. alpina*, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

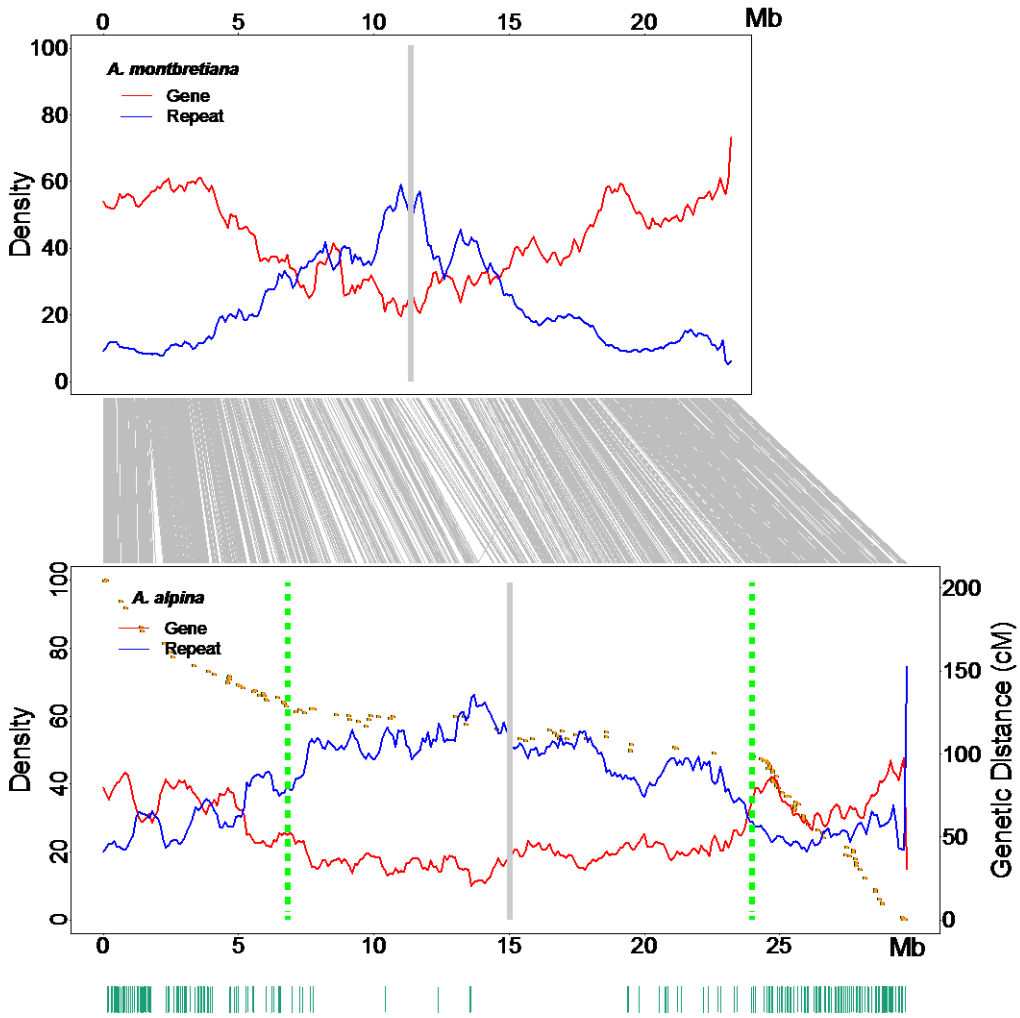


Figure 18 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* chromosome 1.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs between these chromosomes. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the physical position of the genetic marker, where their genetic positions is shown on the right Y axis. Comparisons in other chromosomes are shown in Fig. S1 – S6.

2.2.4 Recombination mediated gene copy number variation

Chromosome segregation or crossovers can introduce different combinations of the two genetic backgrounds. In general, allelic or syntenic homologous regions will not be affected by the recombination in terms of copy changes. Recombination in non-syntenic homologous regions can lead to chromosomal rearrangements and is usually suppressed, however, recombination in syntenic regions can combine different alleles of rearranged regions leading to copy number changes in the recombinant genomes (Wijnker et al. 2013). Here, as I found a substantial amount of translocated genes between the two parental genomes, I expected a high degree of copy number variations (CNVs) of translocated genes in the progenies (Fig. 22, Fig. 23). Assuming an orthologous gene, which is single-copy but resides in different genomic regions between the two *Arabis* genomes, the gene copy number in progenies can vary from zero to four. Among of the 1,256 genes in CTX regions, I found 363 which showed all the five possible copy numbers between the different recombinant lines (Fig. 19). Besides, 750 CTX genes did not show any zero copy number allele in any of the recombinant lines, suggesting these genes might be essential. For example, ortholog genes in the large CTX region between *A. alpina* chromosome 3: 30.06-30.42 Mb and *A. montbretiana* chromosome 5: 31.1-31.4 Mb (Fig. 20 A-B) had one, two, three and four copies in 1, 274, 134, and 48 progenies, respectively. However, there was no line without the translocated region. In another example of a large CTX region (*A. alpina*: chr3 2.16-2.27Mb, *A. montbretiana*: chr6 4.69-4.78Mb), even no single-copy number variation was observed in all progenies (Fig. 20 C-D). Moreover, I also observed zero copy-number variations of 127 one-to-one CTX genes in at least ten progenies.

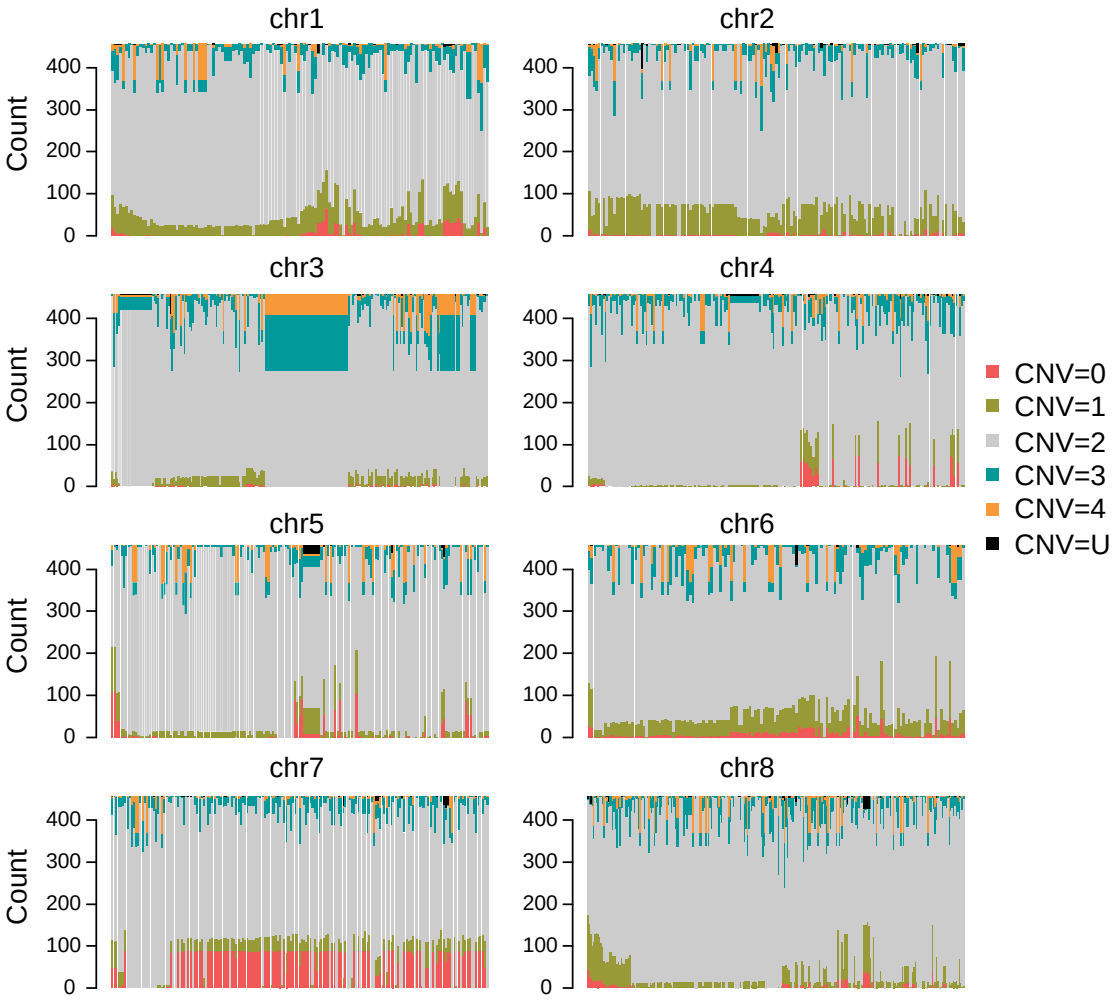


Figure 19 CNV of translocated genes in the *Arabis* introgression lines.

Each column indicate one translocated gene sorted along the x axis by their positions on *A. alpina* chromosome Each translocated gene's copy number range from 0 to 4 exists in number of progenies as the y axis indicates, CNV=U: the copy number are hard to be determined.

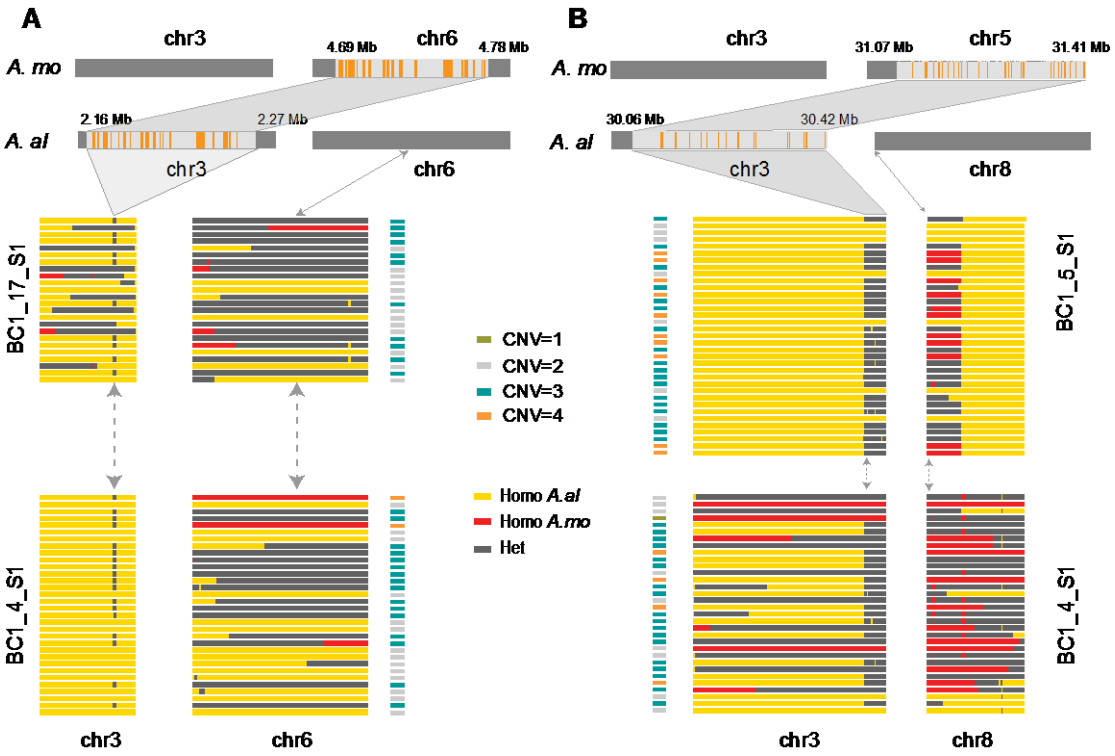


Figure 20 Two examples showing CNVs of translocated genes in *Arabis* introgression lines.

(A-B) two large translocated regions and genotyping of progenies showed no zero copy number of the orthologous genes. The top part shows two large translocations (**A**: *A. alpina* chr3: 30.06-30.42 Mb vs *A. montbretiana* chr5: 31.07-31.41Mb; **B**: *A. alpina* 2.16-2.27Mb vs *A. montbretiana* 4.69-4.78Mb) indicated by light grey color. Genes in the translocations are represented by orange bars. The flanking syntenic regions are indicated by dark grey color (Note: *A. alpina* chromosome 8 up arm and *A. montbretiana* chromosome 5 low arm are homologous). The lower part shows genotypes of translocated regions and their flanking regions. Only recombination maps of families BC1_4_S1, BC1_5_S1 and BC1_17_S1 are shown here. Three different colors indicate different genotype (yellow: homozygous *A. alpina*; red: homozygous *A. montbretiana*; grey: heterozygous). The translocated regions are indicated by the arrows. The copy number of translocated genes varies across progenies, which can be identified by the genotypes of flanking syntenic regions as shown in Fig. 22. Four different colors correspond to different copy number variations (CNVs).

3. Material and Methods

3.1 High-quality plant genome assembly using long-read sequencing and long-range scaffolding technologies

This section (3.1) was the basis a manuscript which was published into a research article in Genome Research (Jiao et al. 2017), which lists me as first author. All the analysis results not generated by me were not shown here. Data from my colleagues or my collaborators was clearly pointed out as described in each subsection of 3.1.

Authors list (Jiao et al. 2017):

Wen-Biao Jiao[1], Gonzalo Garcia Accinelli[2], Benjamin Hartwig[1], Christiane Kiefer[1], David Baker[2], Edouard Severing[1], Eva-Maria Willing[1], Mathieu Piednoel[1], Stefan Woetzel[1], Eva Madrid-Herrero[1], Bruno Huettel[3], Ulrike Hümann[1], Richard Reinhard[3], Marcus A. Koch[4], Daniel Swan[2], Bernardo Clavijo[2], George Coupland[1], Korbinian Schneeberger[1]

Author affiliations:

[1] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2] Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK. [3] Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. [4] Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, 69120 Heidelberg, Germany.

Authors contributions (Jiao et al. 2017):

Conceived and designed the project: KS, WBJ. Sample preparation: BHartwig, CK, SW, UH, MAK. PacBio sequencing: BHuettel, RR. Optical mapping: DB, DS. Genetic map: ES, EMH, GC. Data analysis: WBJ, GGA, EMW, MP, BC. Wrote the paper: KS, WBJ with help of all other authors.

3.1.1 Plant selection, sample preparation and sequencing

Three diploid and self-compatible relatives of *A. thaliana* in Brassicaceae family were selected mainly based on their phylogenetic positions in this family and genome size. *A. alpina (*2n=16), is from the tribe Arabideae (expanded lineage II), while *E. syriacum* (2n=14) is from the tribe Euclidieae (lineage III), and *C. planisiliqua* (2n=14) is from the tribe Conringieae (expanded lineage II). The seed of *C. planisiliqua was* from *BrassiBase* (http://brassibase.cos.uni-heidelberg.de; Koch et al. 2012; Kiefer et al. 2014) database with accession code HEID921022 (herbarium voucher HEID503985). The seed of *E. syriacum* was from Kew Millenium Seedbank with accession code KEW653912. The seed of *A. alpina* was from the reference accession Pajares (Willing et al. 2015).

Plants were prepared by colleagues in MPIPZ (Benjamin Hartwig, Christiane Kiefer, Ulrike Hümann). All the PacBio sequencings libraries were prepared and sequenced on a PacBio RS II sequencer using C4v1 sequencing reagents, by my colleagues (Bruno Huettel, Richard Reinhard) in the Max Planck-Genome Centre, Cologne, Germany. Illumina paired-end sequencing of *E. syriacum* was performed on a HiSeq2500 platform by Max Planck-Genome Centre, Cologne, Germany. Illumina reads of *A. alpina* and *C. planisiliqua* were reused from previous projects (Bewick et al. 2016; Willing et al. 2015). Mate-pair libraries were constructed and sequenced following the earlier method (Heavens et al. 2015) by my collaborators (Gonzalo Garcia Accinelli, Bernardo Clavijo) in Earlham Institute. Earlham Institute's Platforms.

The Illumina paired-end reads were used to evaluate heterozygosity degree based on 25-mer frequencies using the tool Jellyfish (Marçais and Kingsford 2011) and genomescope.R (https://github.com/schatzlab/genomescope).

3.1.2 Genetic map of *A. alpina*

The genetic map was generated by my colleagues in MPIPZ (Edouard Severing, Stefan Woetzel, Eva Madrid-Herrero). An F_2 mapping population including 389

individuals was constructed using three self-pollinated F_1 hybrids between two *A. alpina* accessions from the French Alps.

3.1.3 Optical mapping

All the optical mapping data was generated by my collaborators (David Baker, Daniel Swan) in Earlham Institute. Earlham Institute's Platforms and Pipelines group followed IrysPrep™ Fix'n'Blend Plant DNA extraction protocol supplied by Bionano Genomics.

3.1.4 PacBio assembly

PacBio raw reads from each species were imported into the software SMRT Analysis v2.3 to filter out subreads with length less than 500bp or bases quality (QV) smaller than 80. Two de novo assembly tools Falcon (v0.3.0) and PBcR (with Celera Assembler 8.3rc2), were used to assemble the remaining subreads. For Falcon, the minimal read length in the read correction and assembly steps was set to obtain input subreads with the total length around 25x as suggested. For PBcR, 40x filtered subreads were selected to do read self-correction with MHAP and 25x of the longest, corrected subreads were used for assembly with the Celera Assembler. Both contigs from Falcon and PBcR were polished based on the remapping of filtered subreads using the tool Quiver.

3.1.5 Estimations of assembly error rate

The assembly error rate was estimated in different ways. First, the single nucleotide errors were identified based on the alignments of Illumina paired-end reads. Reads were mapped to each Falcon and PBcR assembly of the three species using BWA (v 0.7.12) (Li and Durbin 2009). SNPs and InDels were called by the tool SAMTools (Li et al. 2009). This level of assembly error rates were then calculated by dividing the total number of homozygous SNPs and InDels by the total length of covered regions with mapping quality higher than 25 and a coverage larger than five.

Furthermore, the level of large-scale misjoins for *A. alpina* assembly was

evaluated with Illumina mate-pairs sequenced from three fragment libraries with different insert-size (5kb, 7kb, and 10kb). First, reads were mapped to each of the assemblies from different steps or data integration using BWA. Then, the reads pairs aligned to different contigs were selected to find those where both or one read was aligned to the inner part of contigs. Here the start or end of inner part of a contig was defined as the total value between average insert size of the respective library and three standard deviations of the insert size distribution. Only read alignments with mapping quality over 30 and without any mismatch or gaps were used for downstream analysis. The resulting clusters with more than five such read pairs indicated a misassembled region. Finally, a consensus set of misassembled regions was obtained by merging the analyzing results of all three libraries.

Additionally, the inter-chromosome misassemblies of *A. alpina* were identified using a genetic map with 734 markers. All marker sequences were aligned to the assembly by the tool blastn, to check whether a sequence have markers from different linkage groups.

3.1.6 Definition of CN50 and CL50

Let C represent a list of all contigs sorted by descending length. Assign them into n distinct sets of contigs, where *n* equals the chromosome number. Here the first (longest) contig is assigned to set 1, the second to set 2 and so on (Fig. 21). The *n*+1 longest contig is then assigned to set *n* again, and the *n*+2 longest contig is assigned to set *n*-1 and so on. Use S to represent such a contig set *s* sorted by descending length. For each set, find one contig $c_i \in S$ matching the below formula

$$\sum_{k=1}^{i} length(c_k) \geq \frac{\sum_{k=1}^{|S|} length(c_k)}{2}$$

where $i \in S$ and no $j < i$ exists, which matches the same criterion. Let M represent the set of contigs matching above formula. I define

$$CN50 = median\left(\bigcup_{k=1}^{|M|} length(c_k)\right)$$

where $c \in M$ and *CL50* is defined as the order number *i* of the CN50 contig $c_i \in S$.

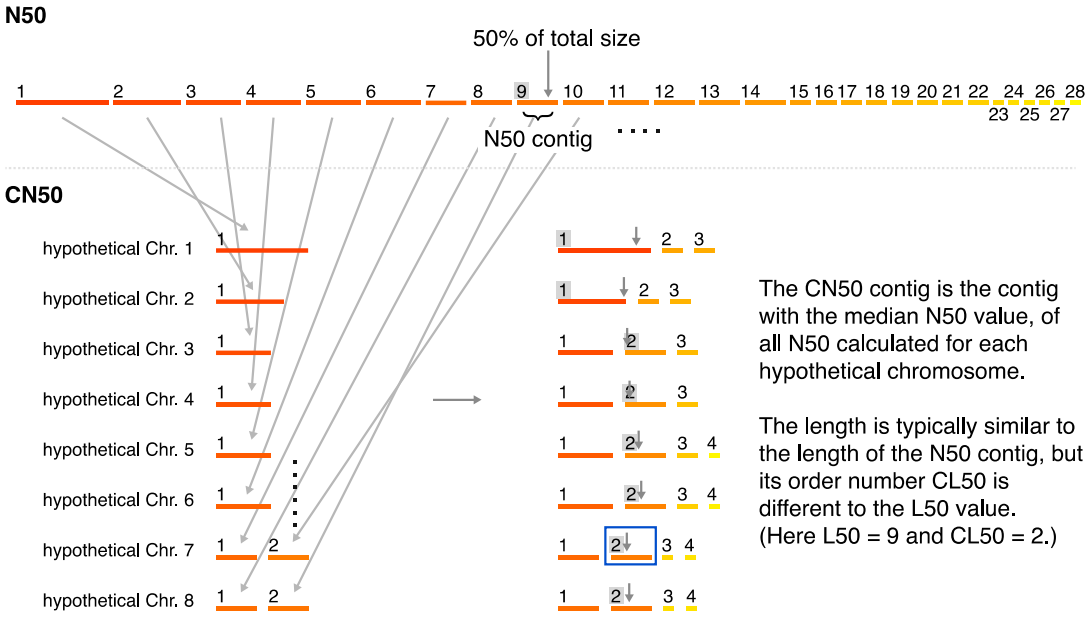


Figure 21 Definition of CN50 and CL50 statistics. (from Jiao et al. 2017)

N50 and L50 refer to one particular contig of a sequence assembly. This contig along with all longer contigs make up more than 50% of the total assembly size. The length of this contig is given by the N50 value, whereas its order number (in an ordering where contigs are sorted by length) is given as L50 (though some people prefer to annotate it vice versa). However, even in perfect assemblies, the L50 is not 1 (as it optimally would be), as the number of chromosomes limits the L50 value. This effect is marginal if an assembly consists of many contigs, however, in assemblies with high contiguity, this effect can be drastic and interfere with the interpretation of the L50 value.

CN50 and CL50 are introduced to normalize the N50 statistics for chromosome number (*n*). Contigs are assigned to hypothetical chromosomes, where the longest contig is assigned to the first chromosome, the second longest is assigned to the second chromosome and so on. The *n*+1 longest contig is then assigned to the *n*-th chromosome again (in the above example contig #9 is assigned to hypothetical chromosome 8) and the *n*+2 longest contig is assigned to chromosome *n*-2 (here

contig #10 is assigned to chromosome 7), until all contigs are assigned. For each of the *n* contig sets N50 is calculated and the median of these values describes the CN50 value. The order number (L50) of the respective CN50 contig (shown in the blue box) in the hypothetical chromosome finally represents the CL50 value.

3.1.7 Integration of Optical Mapping data

For each genome, the raw optical maps were imported and *de novo* assembled into consensus maps using the tool Assembler of the package IrysSove with cutoffs of P < 8e-8 to generate draft consensus maps, P < 8e-9 to extend draft consensus maps and P < 8e-12 to merge the extended draft consensus maps. The resulting consensus maps were used for hybrid scaffolding with above PacBio sequence assemblies generated from Falcon or PBcR. This hybrid scaffolding procedure scaffolded not only the assembly sequences but also the optical consensus maps.

To do hybrid scaffolding, two different scaffolding workflows were used. In the first workflow, the consensus maps were aligned to assembled sequences using the tool RefAligner with alignment cutoff of P < 1e-9. Only the aligned consensus maps not showing conflicting alignments were selected for hybrid scaffolding with the IrysSolve software.

In the second workflow, consensus maps showing conflicting alignments were included as these conflicts indicated the consensus maps or assembly sequences were misassembled. The misassembled position was defined as the midpoint between the first unaligned restriction enzyme site and the last aligned restriction enzyme site at the conflicting regions. In addition, it was shifted accordingly if an InDel with two or more restriction enzyme sites were adjacent to the breakpoint. To determine whether the consensus maps or assembly sequences were misassembled, all conflicting alignments related to the sequences or maps were searched among alignments under relaxed alignment cutoff P < 5e-8. If at least two maps or sequences indicated the same conflicting alignment region in one sequence or map, the sequence or map was misassembled and would be broken at the misassembled breakpoint. If this conflicting alignment was only indicated by one map or sequence, the alignments between the

maps and the sequences from the other assembly (Falcon or PBcR) were checked to see whether the conflicting alignment region was also indicated by another sequence. If not, the sequence was misassembled and would be also broken but the sequence from another assembly was not at this conflicting region. If there is no alignment information at this conflicting region from another assembly, both sequence and map were split. After all these conflicting alignments were solved, I did hybrid scaffolding again with the tool RefAligner (P < 1e-9). This step was performed for the Falcon and PBcR assembly, respectively. Finally, the scaffolds from Falcon-based hybrid scaffolding were aligned to the hybrid consensus maps from the PBcR-based hybrid scaffolding using the tool RefAligner (P < 1e-9). Again, the misassembled scaffold sequences or maps were split as above and were further used for the final hybrid scaffolding with alignment P < 1e-9.

In both of the two approaches above, the gap length between the contigs or scaffolds was estimated based on the distance of spanning restriction sites. The gaps were filled with the respective number of ambiguous base “N”.

3.1.8 Integration of Dovetail Genomics chromatin conformation capture data

The reads of chromatin conformation capture sequencing were generated by the Dovetail Genomics Company. Reads were mapped to *A. alpina* Falcon and PBcR assembly contigs using BWA with mapping quality cutoff of 30 and then used for scaffolding using HiRise software. In order to improve the HiRise scaffolding performance, *in silico* consensus maps were transformed from the PBcR-based HiRise scaffolds based on the nick site of BspQI, then aligned to Falcon-based HiRise scaffolds using RefAligner with P < 1e-9 as cutoff. Again, the scaffolds and maps were broken before scaffolding to remove conflicting alignments as described above. Due to the variable insert length of Chicago read pairs, the gaps of between adjacent contigs were hard to be determined and were simply filled 100 base “N”.

3.1.9 Integration of optical mapping and chromatin conformation capture data

To integrate optical mapping and chromatin conformation capture data, firstly the second optical mapping scaffolding workflow was performed as described above only except that the chromatin capture read pairs were also used to help determining whether the contigs or the consensus maps were misassembled in conflicting alignment regions. Secondly, the Chicago read pairs were aligned to the resulting hybrid scaffold and used for scaffolding by HiRise. Then, the hybrid consensus maps from the first step of optical mapping data integration were aligned to scaffold sequences from the second step of HiRise scaffolding (alignment cutoff $P < 1e\text{-}9$). Again, the potentially misassembled maps and scaffolds were broken and used for final scaffolding.

3.1.10 Estimation of centromeric regions

Firstly, the centromere positions in *A. lyrata* genome assemblies were defined based on the *A. thaliana* centromere positions (The Arabidopsis Genome Initiative 2000)., the cytogenetic maps (Schranz et al. 2006) and a whole genome alignment between *A. lyrata* and *A. thaliana* using the nucmer tool from MUMmer version 3.23 (Kurtz et al. 2004) with parameters “--mum -l 40 -g 90 -c 90 -b 200”. Then, I did whole genome alignment between *A. lyrata* genome and each of the three assemblies using nucmer with parameters “--mum -l 20 -g 90 -c 65 -b 200”. In each alignment, the adjacent alignment regions were merged into homology blocks. A scaffold potentially spanned the centromere when both the flanking sides of an *A. lyrata* centromere were homology with this scaffold. A scaffold may only partially cover the centromere when only side of an *A. lyrata* centromere was homology with this scaffold. Additionally, assembly sequences were also screened to identify centromeric tandem repeats (Melters et al. 2013). The tandem repeat arrays were predicated by the tool TRF with parameters “1 1 2 80 5 200 2000 -h”. The candidate centromeric repeat units were defined from the clustering of the longest tandem arrays.

3.1.11 Annotation and finalization of the assemblies

For general usage, the A. alpina assembly further improved to remove residual misjoined sequences and sequences were anchored into chromosomes based on the genetic map and the previous cytogenetic map (Willing et al. 2015).

Protein-coding genes of *E. syriacum* and *C. planisiliqua* were annotated by integrating evidences from *ab initio* predictions and homology protein sequence alignments. Three ab initio prediction tools were used including GlimmerHMM (v3.0) (Majoros et al. 2004), SNAP (v2013) (Korf 2004) and Augustus (v3.0) (Stanke and Waack 2003). Protein sequences from eight Brassicaceae species (*A. thaliana, A. lyrata, Capsella rubella, Brassica rapa, Eutrema salsugineum, Schrenkiella parvula, A. alpina, Arabis montbretiana*) were aligned to the assembly using the tool Scipio (v1.4) (Keller et al. 2008). These evidences were further integrated into consensus gene models by EVidenceModeler (EVM) software (v2012) (Haas et al. 2008). For A. alpina, the earlier gene models were updated based on the new assembly (Willing et al. 2015). Repeats were annotated using the tool RepeatMasker (v4.0) based on a custom Brassicaceae repeat library.

3.1.12 Scripts and data access

All the scripts related the workflows in this project can be found in the on-line Supplement Material of the publication (Jiao et al. 2017) and the GitHub website (https://github.com/wen-biao/OM-HiC-scaffolding). All raw genome sequencing data, optical mapping data and the assembly sequences have been uploaded into the European Nucleotide Archive with the BioProject ID PRJEB16743. Gene annotations can be found in the on-line Supplemental Material of the publication (http://genome.cshlp.org/content/early/2017/04/05/gr.213652.116/suppl/DC1).

3.2 Comparative genomic and genotypic characterization of annual-perennial interspecific introgression lines constructed from two divergent *Arabis* species

3.2.1 Plant sample preparation, genome sequencing and RNA-seq of *A. montbretiana*

All respective samples were prepared by my colleagues in MPIPZ. Two short DNA fragment libraries and two long DNA fragment libraries were constructed for Illumina whole genome sequencing. Seven RNA-seq samples were collected at different growth periods. These sequencing data is summarized in Table 15 and Table 16.

Table 15 The Illumina reads for *A. montbretiana* genome *de novo* assembly

Library	Read Length(bp)	Sequencing Depth
Paired-End 180bp	2*100	179.9
Paired-End 500bp	2*121	81.4
Mate-Pair 3Kb	2*30~50	1.0
Paired-End_20Kb	2*30~100	17.8

Table 16 The Illumina RNA-seq reads of *A. montbretiana*

Tissue	Read Length	Read Number
leaves	2*101	58,422,899
cotiled	2*101	54,709,620
vegetapic	2*101	57,748,666
seedlings	2*101	62,577,078
cauline	2*101	56,173,527
florabuds	2*101	54,867,518
siliq	2*101	56,781,274
total		**401,280,582**

3.2.2 Cytomolecular comparative maps of *A. montbretiana*

This work was done by our collaborators Martin Lysák and his colleagues at CEITEC – Central European Institute of Technology, Masaryk University, Brno, Czech Republic. In Brassicaceae, most species share the common chromosome karyotype blocks with the Ancestral Crucifer Karyotype (ACK) where 24 blocks are defined according to the model plant *A. thaliana*. Using comparative chromosome painting based on *A. thaliana* BACs, the conservation of (or possible rearrangements in) these blocks can be identified in any Brassicaceae genome. The arrangement of these blocks displays the cytomolecular comparative maps.

3.2.3 Genome assembly of *A. montbretiana*

Genome sequences were assembled by the tool ALLPATHS-LG with default parameters using the Illumina sequencing reads (Gnerre et al. 2011). The assembled scaffold sequences were blasted against sequences of *A. thaliana* mitochondrial (The Arabidopsis Genome Initiative 2000), *A. alpina* chloroplast (Melodelima and Lobréaux 2013) and NCBI bacterial nucleotide sequences, to detect organellar and remove contaminant sequences.

To anchor the scaffolds to chromosomes, I first improved the *A. alpina* genome reference sequences version 5 (Jiao et al. 2017) by adjusting the order and orientation of several scaffolds based on the genotyping result of *Arabis* introgression lines. This updated *A. alpina* genome sequences and cytomolecular comparative maps of *A. montbretiana* were used to guide the anchoring of scaffolds to chromosomes. All scaffolds were aligned against *A. lyrata* genome using the tool nucmer in MUMmer package with parameters setting of "--mum –l 20 –g 90 –c 65 –b 200" (Kurtz et al. 2004), to find the Ancestral Crucifer Karyotype (ACK) blocks. Here, the ACK block positions in *A. lyrata* were identified based on the whole genome alignments between *A. thaliana* and *A. lyrata*, ACK of *A. lyrata* (Schranz et al. 2006) and the ACK block positions of *A. thaliana* (Lysak et al. 2016). Then all scaffolds that unambiguously

aligned against ACK blocks were anchored and oriented into eight chromosomes based on the cytomolecular comparative maps (Fig. 10 A).

3.2.4 Genome annotation of *A. montbretiana*

Protein-coding genes were annotated by a pipeline integrating evidences from *ab initio* prediction, RNA-seq paired-end reads mapping and homologous protein alignments. Firstly, *ab initio* gene predictions were performed using the tool AUGUSTUS (version 3.2.3) (Stanke and Waack 2003) and GlimmerHMM (version 3.0.3) (Majoros et al. 2004). Besides, nearly 200 million paired-end RNA-seq reads from seven libraries sampled from tissues under different growth periods were generated, and mapped to the *A. montbretiana* assembly using tophat2 (version 2.1.0) with default parameters (Kim et al. 2013a), followed by assembling the transcripts using the tool Cufflinks (version 2.2.1) (Trapnell et al. 2010). Moreover, I aligned protein sequence from *A. thaliana* (The Arabidopsis Genome Initiative 2000), *Populus trichocarpa* (Tuskan et al. 2006), *Vitis vinifera* (French et al. 2007), *Oryza sativa* (Goff 2002) to *A. montbretiana* genome using tblastx (Altschul et al. 1997). All proteins with alignment e-value less than 1e-5 were aligned to *A. montbretiana* genome again, this time using exonerate with alignment percent cutoff of 50% (Slater and Birney 2005). Then, all evidences from above were combined to generate consensus gene models using the tool EVidenceModeler (version 1.1.0) (Haas et al. 2008). Finally, I used the tool PASA (version 2.0.0) to update the consensus gene models by adding UTR annotations and models for alternatively spliced isoforms (Haas et al. 2003).

To identify transposable element related genes, I integrated the result of transposable element annotation, scanning result of TE-related protein domain HMM profiles using the tool HMMER 3 (version 3.1; http://hmmer.org/) and blastp result with *A. thaliana* proteins. When genes coding regions were overlapped by a region with similarity to a TE with at least 20% or had TE-related protein domains, they were annotated as TE-related genes unless they additionally featured a blastp hit to an *A. thaliana* protein.

TEs were annotated using RepeatModeler (version 1.0.8) and RepeatMasker (version 4.06; http://www.repeatmasker.org) with default parameters. Intact LTR were identified using the tool LTRharvest in GenomeTools package version 1.3.7 (Ellinghaus et al. 2008). To predict the LTR insertion time, all pairwise intact LTRs were firstly aligned using the tool MUSCLE (version 3.8.31) (Edgar 2004). Then the distance *k* of each pair LTR was calculated using the tool dismat from the EMBOSS package version 6.6.0.0 (Rice et al. 2000). The insertion time T of each LTR pair was calculated based on the formula T=k/(2r). Here the nucleotide substitution rate r was set 7e-9 per site per year and base (Ossowski et al. 2009).

3.2.5 Identification of genome synteny and rearrangements

To find syntenic and rearranged genomic regions between *A. alpina* and *A. montbretiana*, I run the whole genome sequence alignment tool nucmer from the software package MUMmer version 3.2.3 (Kurtz et al. 2004) with parameter setting of "---mum –l 40 –g 90 –c 90 –b 200". After filtering redundant short alignment blocks, the uniquely aligned blocks were used to find the longest syntenic regions using the Heaviest Increasing Subsequence algorithm, which is based on the specific sorting of the genome coordinates of *A. alpina* and *A. montbretiana*. Based on the syntenic block (SYN) backbone, other rearranged blocks including inversion (INV), intra-chromosome translocation (ITX), and inter-chromosome translocation (CTX) were assigned accordingly. For duplicated alignments, I clustered them into alignment groups where each group contained all regions which could be aligned to one or more of others. Assuming there are N regions of *A. alpina* and M regions of *A. montbretiana* in one group, a matrix with N rows and M columns was built. In this matrix, each element represents one possible alignment pair. This type of alignment could be assigned as SYN, INV, ITX, or CTX according to the syntenic backbone. One region from *A. alpina* can have multiple aligned regions in *A. montbretiana* with different alignment types. For these cases, the final alignment type of this region was assigned based on the priority order: SYN > INV > ITX > CTX. Once an alignment pair was assigned a type, the two regions would be deleted from the matrix and their pair combination would be

fixed. When regions from one genome or both genomes were removed from the matrix, this process of assignment of pair combination and alignment type was finished. Finally, the remaining regions from one genome were assigned as gained copy number variation (CNV-G).

3.2.6 Comparative gene family analysis of *A. montbretiana* and *A. alpina*

The ortholog gene family analysis was done by iteratively running the tool OrthoMCL (Li et al. 2003). The remaining un-clustered genes from one genome were assigned their best blastp hit in another genome. These paired ortholog genes were assigned as syntenic (SYN) gene or rearranged (INV, ITX or CTX) genes based on their genome positions and the syntenic genome regions defined by whole genome sequence alignment. For those none one-to-one ortholog families, I checked all the possible ortholog pair combinations to assign them as syntenic or rearranged pairs. This would generate a matrix with row and column number equal to gene number of *A. alpina* and *A. montbretiana*. In the matrix, each element presented the ortholog type (SYN, INV, ITX or CTX) for each gene pair combination. For each gene of one genome, its final paired ortholog would be determined following the ortholog types priority (priority order: SYN > INV > ITX > CTX). Once the gene pair was assigned an ortholog type, it would be deleted from the matrix and their combination was fixed. Finally, those unpaired genes in the gene family would be assigned as CNV-Pre (for *A. alpina*) or CNV-Abs (for *A. montbretiana*) using the *A. alpina* genome as the reference.

3.2.7 Genotyping of the introgression population

The *Arabis* interspecific introgression population were generated by our collaborators, Christiane Kiefer and other colleagues in MPIPZ. Briefly, the population were constructed by firstly crossing between *A. montbretiana* and *A. alpina*, then backcrossing to the recurrent parent *A. alpina*, followed by one, two or three arounds of self-crossing (BC1S1, BC1S2, and BCSF3) as shown in Fig. 13 . 460 progenies were selected for Genotyping-by-Sequencing (GBS).

Each individual's sequencing reads were extracted from the GBS reads files according to their specific barcode. Only "good" reads were kept for analysis. Here, I defined "good" reads, when they satisfied the requirements introduced in a previous study (Elshire et al. 2011): (1) Containing barcode and the four-base remnant of the ApeKI cut site (CWGC); (2) No ambiguous base "N" in the first 64 bases after the barcode end; (3) Not containing the eight bases of common adaptor in their first 50 bases after the barcode end. Reads were trimmed if they contained the common adaptor "AGATCGGA", then mapped to *A. alpina* reference genome using the tool BWA (Li and Durbin 2009). Genotypes were called with the tool SAMTools (Li et al. 2009). I used a GBS read-based analysis of the two parental genomes of *A. alpina* and *A. montbretiana* to select SNP markers for genotyping. I selected SNPs from *A. montbretiana* reads alignment with mapping coverage>=5, mapping quality >=25 and *A. montbretiana* allele's reads mapping ratio>=0.9, while I discarded those SNPs which were also called with the *A. alpina* reads or with less than five reads supporting *A. alpina* allele. Moreover, I filtered those close to the assembly gaps with distance less than 500bp and only kept one of adjacent SNPs whose distances were less than 50bp. To reduce the disturbance from repeats during genotyping, I firstly extracted 100bp genome sequence from each flanking side of above filtered SNP markers, and aligned these fragments to the *A. alpina* and *A. montbretiana* genomes. If an alignment showed more than one hit (mismatch<=4, gap<=1 or mismatch=5, gap=0), I discarded the respective SNP marker.

The remaining SNPs were used as genotyping markers. The genotype at each marker position of each recombinant individual was assigned according to the respective number of supporting reads from *A. montbretiana* and *A. alpina*. When the total mapped reads were less than four, the genotype was defined as undetermined "U". When the ratio of mapped reads from allele *A. montbretiana* was (1) larger than 0.9, the genotype was determined as homozygous *A. montbretiana* "MM"; (2) less than 0.1, the genotype was determined as homozygous *A. alpina* "AA"; (3) between 0.1 and 0.9, the genotype was determined as heterozygous "AM".

To identify the recombination breakpoints as accurately as possible, only initial

genotypes at syntenic markers were remained. To correct residual genotyping errors, a sliding window with seven marker was used for imputation. The recombination breakpoints were defined as the middle point between adjacent markers showing different genotypes.

After removing those progenies with low number of sequencing reads, low number of informative genotypes, or weird genotyping result due to possible sample contamination, 416 progenies across nine BC1S1 families and seven BC1S2 families remained for downstream analysis.

3.2.8 Genotyping the copy number of translocated genes

The copy number variation of translocated genes were determined according to the genotypes of their flanking syntenic regions as shown in Fig. 22 and Fig. 23.

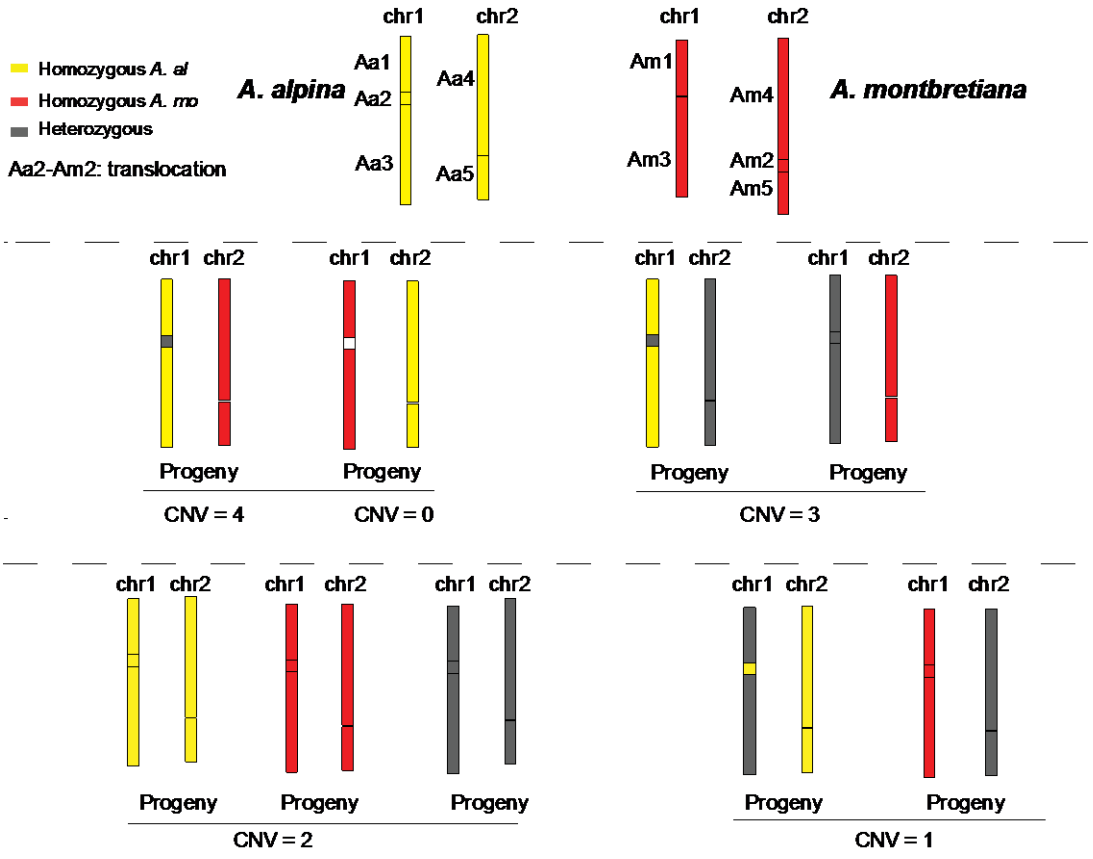


Figure 22 Recombination mediated copy number variations of translocated genes. Different colors indicates different genotypes.

The top row shows the chromosome of two parental genomes. It indicates an inter-

chromosome translocated region Aa2-Am2. The flanking regions Aa1, Aa3, Aa4, and Aa5 are accordingly syntenic with Am1, Am3, Am4, and Am5. The middle and bottom row show the genotype based on the reference genome of *A. alpina*. All genotypes at regions of Aa1, Aa3, Aa4 and Aa5 together indicate the genotype of translocated region Aa2 and the copy number of genes in Aa2 in different scenarios. For example, when a diploid progeny inherits both copies of chromosome 1 both from *A. alpina* and both copies of chromosome 2 from *A. montbretiana*, it will have four copies of the translocated gene. As the GBS analysis is performed with *A. alpina* as reference sequence the reads of all four genes will align to the chromosome 1 (as this is where the gene resides in *A. alpina*) and lead to a heterozygous genotype in this regions, while the flanking (non-translated) regions while display a homozygous genotype. Likewise, other progenies can present different copies of the translocated genes leading to different genotypic patterns if the chromosomes are inherited in a difference way.

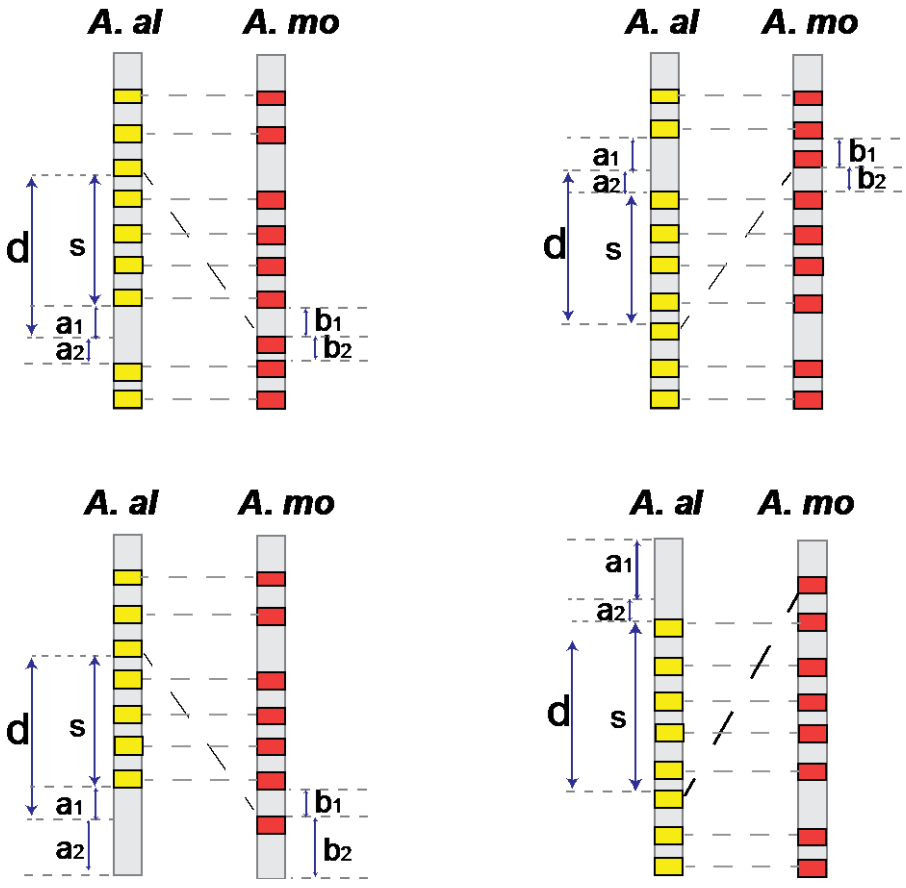


Figure 23 Large-scale arrangements of intra-chromosome translocations.

A.al: *A. alpina*, *A.mo*: *A. montbretiana.* The gray rectangles represent the homologous chromosomes. The small yellow and red rectangles shows the aligned regions. The dashed grey lines connect the homologous regions between *A. alpina* and *A. montbretiana*. The dashed black lines indicate the intra-chromosome translocations. Distances are marked with s, b1, b2, a1, a2 and d, which are described in detail below.

For the translocated genes with the homologous chromosome, I focused on genes with a translocated distance *d* larger than a particular cutoff. The translocated distance *d* was defined and calculated according to the four cases shown in Fig. 22. For example in the case of Fig. 23 A, *d* is calculated with the formula: $d = s + (a1+a2)*b1/b2$. Here, *b1* and *b2* are the distances between the translocation region in *A. montbretiana* and its upstream first (Am.TU1) and downstream first (Am.TD1) syntenic region, respectively. The *a1+ a2* represents the distance of these two syntenic region in *A. alpina*. While *s* indicates the distance between the translocated region in *A. alpina* and the syntenic region referring to that region Am.TU1. Here, the threshold for *d* was set as 5 Mb, any translocated gene pair with *d* larger than this were used for CNVs analysis among all progenies.

4 Discussion

4. 1 High-quality plant genome assembly using long-read sequencing and long-range scaffolding technologies

This section (4.1) was the basis a manuscript which was published into a research article in Genome Research (Jiao et al. 2017), which lists me as first author. All the analysis results not generated by me were not shown here. Data from my colleagues or my collaborators was clearly pointed out as described in the corresponding subsections of 3.1.

Authors list (Jiao et al. 2017):

Wen-Biao Jiao[1], Gonzalo Garcia Accinelli[2], Benjamin Hartwig[1], Christiane Kiefer[1], David Baker[2], Edouard Severing[1], Eva-Maria Willing[1], Mathieu Piednoel[1], Stefan Woetzel[1], Eva Madrid-Herrero[1], Bruno Huettel[3], Ulrike Hümann[1], Richard Reinhard[3], Marcus A. Koch[4], Daniel Swan[2], Bernardo Clavijo[2], George Coupland[1], Korbinian Schneeberger[1]

Author affiliations:

[1] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2] Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK. [3] Max Planck-Genome-center Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. [4] Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, 69120 Heidelberg, Germany.

Authors contributions (Jiao et al. 2017):

Conceived and designed the project: KS, WBJ. Sample preparation: BHartwig, CK, SW, UH, MAK. PacBio sequencing: BHuettel, RR. Optical mapping: DB, DS. Genetic map: ES, EMH, GC. Data analysis: WBJ, GGA, EMW, MP, BC. Wrote the paper: KS,

WBJ with help of all other authors.

In this project, three Arabidopsis relatives' genomes were firstly assembled only using PacBio long reads. I developed different scaffolding workflows to integrate optical mapping and chromatin conformation capture data. For evaluating assembly contiguity independent of chromosome number, a new statistic chromosome-N50 (CN50) was introduced.

Although the high error rates of PacBio reads, the PacBio assemblies had nucleotide error rates lower than one in 10kb, which were as accurate as the Sanger sequencing-based assemblies. However, these primary PacBio assembly contigs suffered from large-scale and inter-chromosome misassemblies mostly due to the highly repetitive regions. My improved workflows could identify and resolve such misassemblies during the integration of optical mapping and/or chromatin conformation capture data. The integration of optical mapping data revealed most of conflicts between assembly sequences and consensus maps. More relaxed parameters for *de novo* assembly of consensus maps might increase the contiguity of maps and more misassemblies. However, as these misassemblies could be resolved during scaffolding, the final assembly might still have higher assembly contiguity and accuracy, implying an optimal parameter setting will allow to better the final assembly.

Integration of chromatin conformation capture data could generate similar improvements on assembly contiguity and also resolve large-scale or inter-chromosome misassemblies. Furthermore, this data did not show highly redundant scaffolding information with optical mapping data. Short contigs are frequently not scaffolded by optical mapping data since they may not be accurately aligned to optical maps due to insufficient restriction sites. However, contig size does not greatly affect the scaffolding using chromatin conformation capture data as this scaffolding can also perform well when the primary assembly only has N50 of 25kb (Putnam et al. 2016). In fact, more short contigs were scaffolded during integration of chromatin contact data.

Additionally, assembly contiguity could be also improved by integrating contiguity information from different assembly tools. Actually, I found that no inter-chromosome

misassemblies was shared between Falcon and PBcR contigs, implying that both Falcon and PBcR could perform better in some regions while the other does not, also the algorisms of assemblers should be improved accordingly.

In the future, longer sequencing reads might make scaffolding technologies obsolete when reads increase longer enough to span any complex or repetitive regions. Nevertheless, before this surprising advancements appear, the art of genome assembly may still require more sophisticated methods to integrate both the sequence and scaffolding data together to untangle complex graph in the initial construction of contigs. This integration will do the initial contig assembly and scaffolding in a one-step procedure, promising significant improvements on assembly efficiency and quality.

4. 2 Comparative genomic and genotypic characterization of annual-perennial interspecific introgression lines constructed from two divergent *Arabis* species

Previous researches revealed considerable variation of genome size and chromosome karyotype among closely related species diverged over short time scales, such as *Arabidopsis*. Here, the annual species *A. montbretiana* has a genome, which is 100 Mb smaller than its perennial sister *A. alpina*. Our analyses suggested that their genome size difference was mostly caused by differences in their recent transposon element activity. This mechanism of genome sequence expansion or contraction was similar with that of Arabidopsis where the smaller genome of *A. thaliana* was generated by reduced transposon element activity or elimination and shortening of intergenic and intronic sequences (Hu et al. 2011). Most of currently available genome size estimation in Arabideae indicated that perennials tend to have larger genomes (Bennett et al. 2011), such as perennial *Arabis procurrens* (352Mb), *Arabis tibetica* (323 Mb), *Arabis nordmanniana* (C Kiefer et al., 2017) (assembly size 342Mb). However, detailed evolutionary processes of genome sequences remain still undetermined without high quality genome assemblies of ancestors or outgroups. Apart from genome size change, chromosome karyotype also showed rearrangements in the two *Arabis* species.

Comparison with the ancestral crucifer karyotype (Schranz et al. 2006; Lysak et al. 2016), both *A. alpina* and *A. montbretiana* have experienced several large-scale karyotype rearrangements, but both still remained the ancestral chromosome number. Even though they only diverged five million years ago, one chromosome arm translocation and one centromere reposition occurred. More cytomolecular comparative maps of other *Arabis* species will highlight the underlying chromosome karyotype rearrangements concurrent with evolutionary life-cycle transitions.

However, higher resolution of genome sequences relies on accurate chromosome-level genome assembly. The high-quality *A. montbretiana* and *A. alpina* genome assemblies provided us an opportunity to deeply investigate the genome evolution. Detailed comparative genomic analysis between *A. montbretiana* and *A. alpina* showed numerous genome rearrangements and ortholog gene family differences including a large number of translocated regions and more than one thousand inter-chromosome translocated orthologous genes. In conclusions, this study indicated that substantial genome divergences at different scales were possibly concurrent with the life-cycle transition.

These two highly divergent parental genomes resulted in some special genotypes in their inter-species progenies. Firstly, crossover mostly emerged in syntenic regions. This pattern is similar with intraspecific hybridization. As our genetic markers were not dense enough due to the common problem of GBS, I were not able to accurately determine the recombination positions at very high resolution. Whole genome sequencing of meiotic tetrads should find more accurate recombination positions, which was often tailored for genome-wide study of recombination and gene conversion events (Lu et al. 2012; Wijnker et al. 2013). Besides, recombination is often repressed in pericentromeric regions. However, it may occur more frequently in inter-specific crossing when the donor genome have different organization of pericentromeric regions. Numerous crossovers in pericentromeric regions were observed among progenies. However, as the study was not designed for investigating inter-specific recombination, I cannot conclude whether the *Arabis* with smaller pericentromeric regions increase the recombination frequency significantly. Moreover, as substantial

translocations exist between the parental genomes, progenies presented copy number variations of translocated genes. This phenomena of recombination-mediated CNVs of translocations was previously observed in *A. thaliana* crossing between two different accessions (Wijnker et al. 2013). Here, I identified 1,256 inter-chromosome translocated genes at whole genome-wide level and determined their copy number in each introgression line. Interestingly, some gene presented all possible copy number among progenies while some genes not, even had no zero copy. Gene copy number variation often result in diverse phenotypic traits (Sutton et al. 2007; Hanikenne et al. 2008). Besides, progenies with different genetic background or copy number of allelic genes may present allelic biased or specific gene expression. Thus, the CNV of translocated genes may influence gene expression and even specific phenotypic traits. Interestingly, seven translocated flower-control related genes might affect the flowering when they present different copies in progenies. Further experiments will elucidate these in the future.

Finally, my workflow on identifying CNVs of translocated genes could be also applied in other inter-specific or intra-specific genotyping studies. Especially, the inter-specific crossing lines between cultivates and wild types are widely utilized in crop breeding, where it can be expected that substantial translocated genes exist between cultivars and wild accessions. While genome-wide study of recombination-mediated CNVs relies on high-quality assembly of parental genomes. As the current genome sequencing cost decreases greatly and multiple long-read or long range genomic technologies advance, more genome assemblies within the same species or closely-related species will become available, which in turn contributes to this study.

4. 3 Concluding remarks

In my thesis, I assembled plant genomes using both second and third generational sequencing technologies. Compared to the assemblies based on the second generation sequencing data, genome assemblies by utilizing third generational sequencing method have much higher contiguity, similar accuracy, and substantial

completeness. Especially, those genomes with highly repetitive or heterozygous regions cannot be assembled well by using short reads, which now can be available by using long reads. Besides, the long-range scaffolding genomic methods can also improve the assembly contiguity and accuracy. As these genomic technologies are complementary in assembly improvements, they can be efficiently combined together to detangle other more complex plant genomes, such as extremely large, heterozygous genomes or autopolyploid genomes.

Simultaneously, other researches can benefit from high-quality genome assemblies, like comparative genomic studies. Before the widely application of third generation genomic technologies, only a few of comparative genomic studies are performed among phylogeny closed species. These phylogeny closed species or sister species are often used for constructing introgression populations. However, they may have greatly genomic divergence which frequently results in failures in single parent genome based genomic analysis. Conversely, parents' genomes aware based analyses can shed light on the understanding of divergent evolution of parent genomes and help to identify causal genes controlling important traits.

Supplement

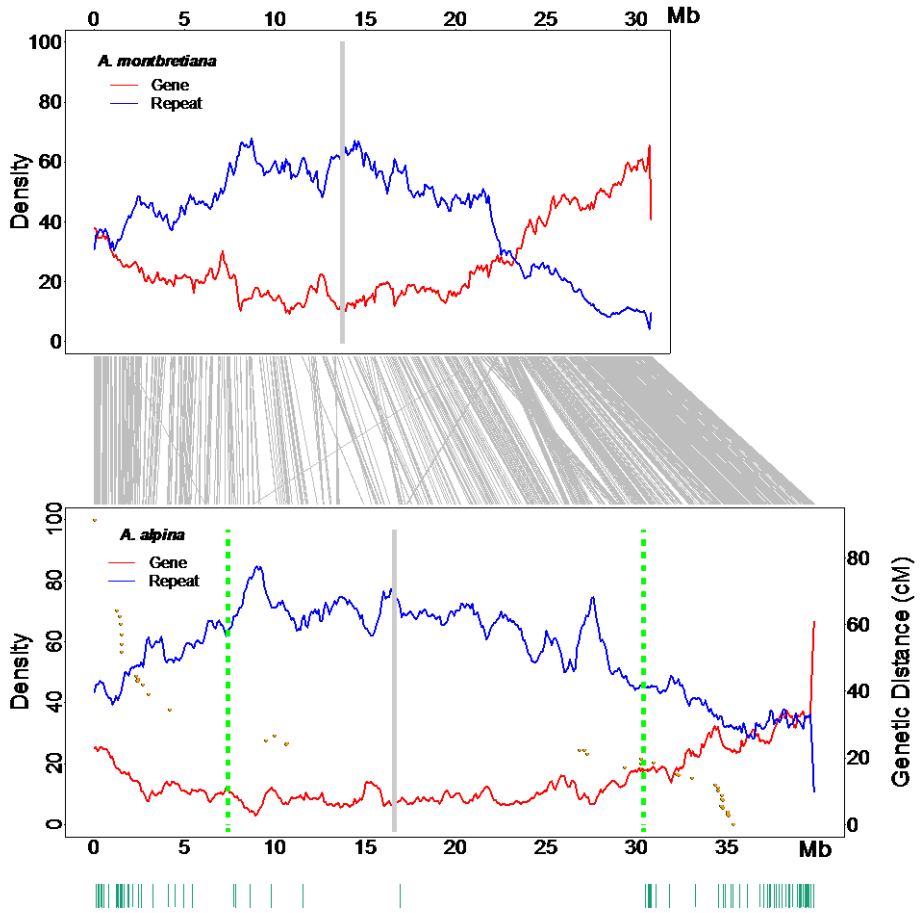


Figure S1 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 2.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

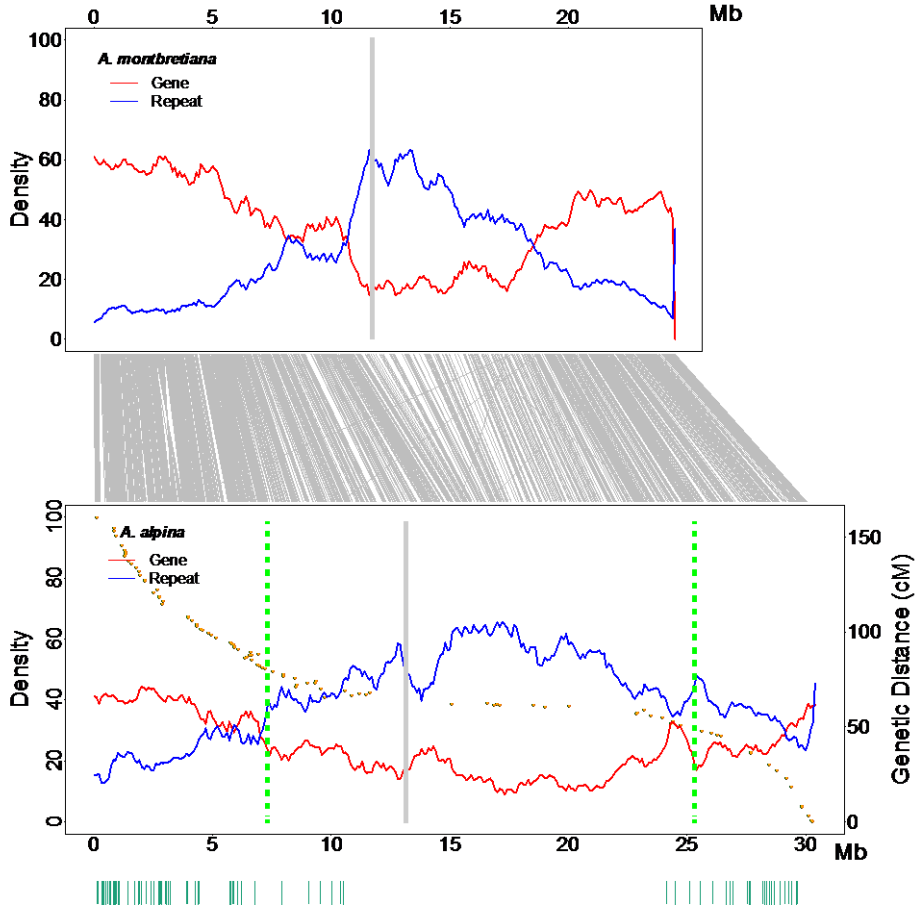


Figure S2 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 3.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

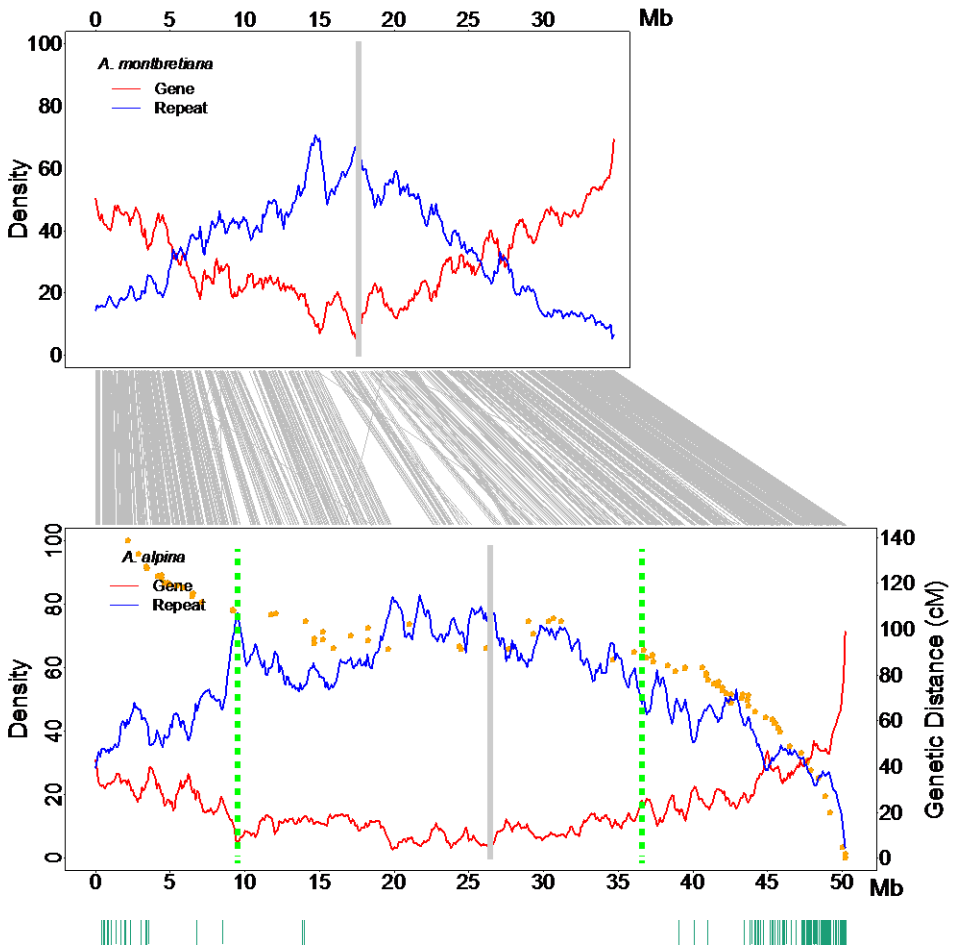


Figure S3 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 4.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

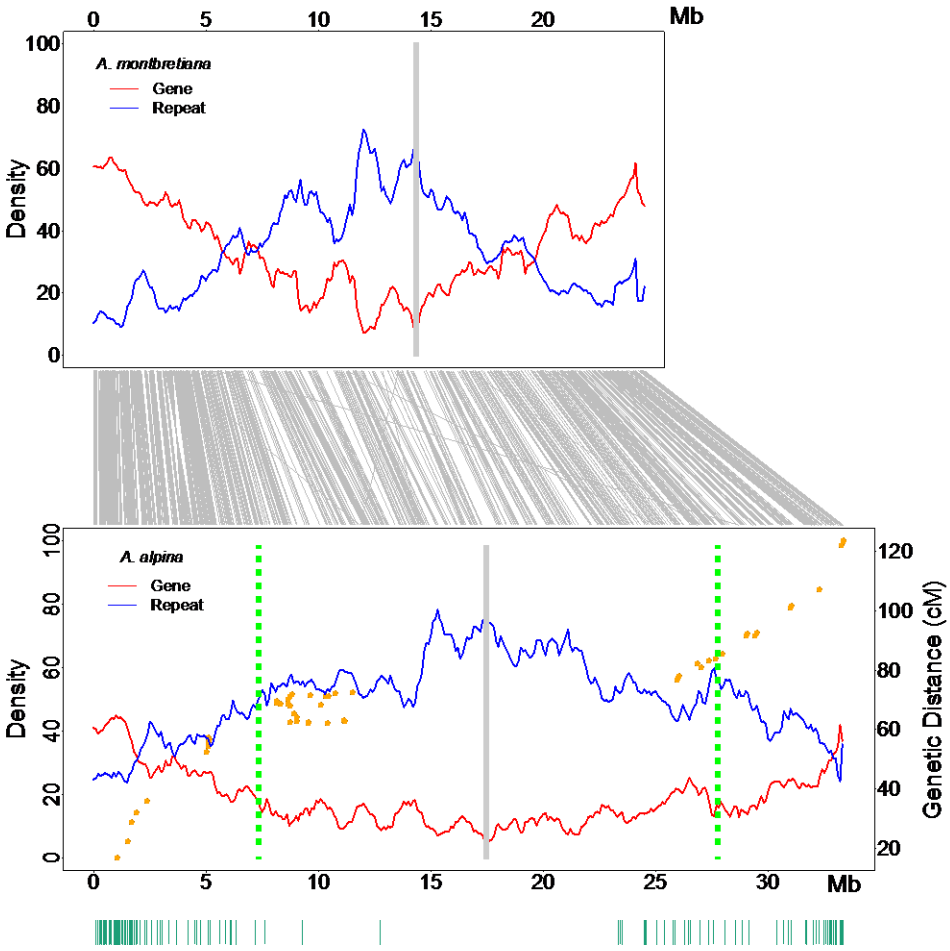


Figure S4 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 6.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

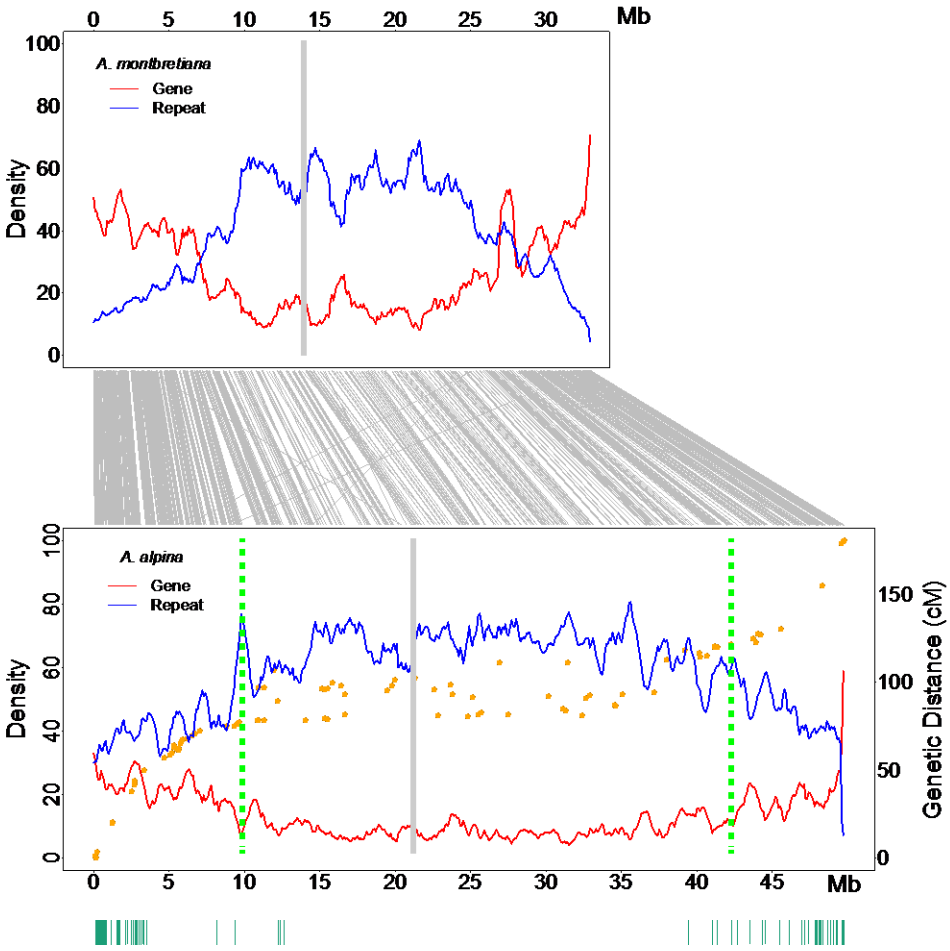


Figure S5 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 7.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

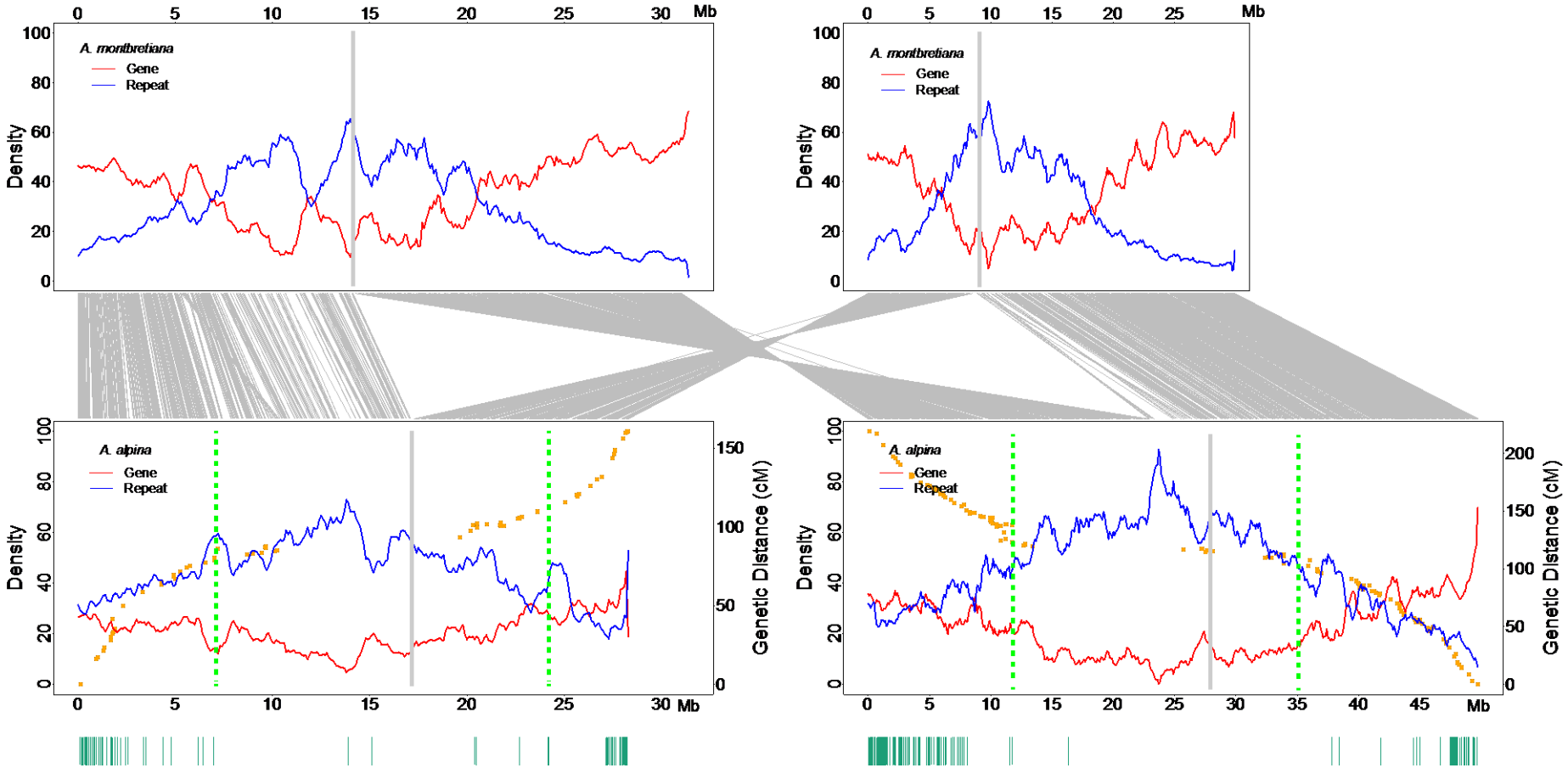

A. montbretiana
Gene
Repeat
Density
Mb
A. alpina
Gene
Repeat
Density
Genetic Distance (cM)
Mb

Figure S6 Pericentromeric recombinations in inter-species introgression lines. Comparison between *A. montbretiana* and *A. alpina* on chromosome 5 and 8.

The red and blue curves indicate the gene and repeat density along the chromosome, respectively. The density was calculated by using a 1 Mb sliding window with step size of 100kb. The grey lines between the upper and lower plots represent all one-to-one orthologous gene pairs. The bottom green bars indicate the recombination breakpoint positions identified in the *Arabis* introgression lines. The narrow gray bars in the plots correspond to the centromere positions. The dashed green bars indicate the range of pericentromeric region in *A. alpina* based on the genetic distance varying along the physical positions. X axis corresponds to the genome position, unit: Mb. Left Y axis indicates the gene or repeat percent. The orange points represent the genetic marker positions on the physical position as the x-axis shows, and genetic positions as the right Y axis shows.

Bibliography

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, Turk C, Pignatelli N, Adey A, Kitzman JO, Vijayan K, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* **46**: 1343–1349.

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. *Curr Opin Genet Dev* **17**: 513–518.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**: 1–7.

Bankevich A, Pevzner PA. 2016. TruSPAdes: barcode assembly of TruSeq synthetic long reads. *Nat Methods* **13**: 248–250.

Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, et al. 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol* **30**: 701–707.

Bena G, Lejeune B, Prosperi JM, Olivieri I. 1998. Molecular phylogenetic approach for studying life-history evolution: the ambiguous example of the genus Medicago L. *Proc Biol Sci* **265**: 1141–1151.

Bennett MD, Leitch IJ, Gardens RB, Tw S. 2011. Nuclear DNA amounts in angiosperms : targets , trends and tomorrow *Ann Bot* **107**: 467–590.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623–630.

Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al. 2016. On the origin and evolutionary consequences of

gene body DNA methylation. *Proc Natl Acad Sci* **113**: 9111–9116.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry C, Bliek M, Boersma M, Borghi L, Bruggmann R, et al. 2016. Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nat Plants* **2**: 16074

Brenchley R, Spannagl M, Pfeifer M, Barker GL a, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**: 1119–1125.

C Kiefer1, E Severing1, R Karl3, S Bergonzi2, M Koch3, A Tresch1, 4, G Coupland1 5. 2016. Divergence of annual and perennial species in the Brassicaceae and the contribution of cis-acting variation at FLC orthologues. *Mol Ecol* **38**: 42–49.

Chalhoub B, Denoeud F, Liu S, Parkin I a. P, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**: 950–953.

Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert V a, dePamphilis CW, Leebens-Mack J, Rounsley S, Schuster SC, et al. 2013. Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science* **342**: 1516–1517.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563–569.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**:1050-1054.

Damas J, O'Connor R, Farré M, Lenis VPE, Martell HJ, Mandawala A, Fowler K,

Joseph S, Swain MT, Griffin DK, et al. 2016. Upgrading short read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res* **27**: 875-884.

Dassanayake M, Oh D, Haas JS, Hernandez A, Hong H, Ali S, Yun D, Bressan RA, Zhu J, Bohnert HJ, et al. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. *Nature Genetics* **43**: 913–918.

Datson PM, Murray BG, Steiner KE. 2008. Climate and the evolution of annual/perennial life-histories in *Nemesia* (Scrophulariaceae). *Plant Syst Evol* **270**: 39–57.

Deamer D, Akeson M, Branton D. 2016. Three decades of nanopore sequencing. *Nat Biotechnol* **34**: 518–524.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

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

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: 1–10.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the gap: upgrading genomes with pacific biosciences rs long-read sequencing technology. *PLoS One* **7**.

Eshed Y, Zamir D. 1995. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* **141**: 1147–1162.

Fierst JL. 2015. Using linkage maps to correct and scaffold *de novo* genome assemblies: Methods, challenges, and computational tools. *Front Genet* **6**: 1–8.

French T, Consortium IP, Characterization GG. 2007. The grapevine genome

sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–468.

Friedman J, Rubin MJ. 2015. All in good time: Understanding annual and perennial strategies in plants. *Am J Bot* **102**: 497–499.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* **108**: 1513–1518.

Goff SA. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* **296**: 92–100.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res* **25**: 1750–1756.

Haas BJ, Delcher AL, Mount S.M. SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7.

Hall MC, Willis JH. 2006. Divergent selection on flowering time contributes to local adaptation in *Mimulus guttatus* populations. *Evolution* **60**: 2466–2477.

Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Krämer U. 2008. Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature* **453**: 391–395.

Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J, et al. 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. *PLoS One* **8**.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res* **16**: 1252–1261.

Heavens D, Accinelli GG, Clavijo B, Clark MD. 2015. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *Biotechniques* **59**: 42–45.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**: 1098–1102.

Hirsch C, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov D, Baruch K, Lu F, Hernandez AG, et al. 2016. Draft assembly of elite inbred line ph207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell* **28**: 2700-2714.

Hoffmann MH, Schmuths H, Koch C, Meister A, Fritsch RM. 2010. Comparative analysis of growth, genome size, chromosome numbers and phylogeny of *Arabidopsis thaliana* and three cooccurring species of the Brassicaceae from uzbekistan. *J Bot* **2010**: 1–8.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**: 2770–2784.

Hoshino A, Jayakumar V, Nitasaka E, Toyoda A, Noguchi H, Itoh T, Shin-I T, Minakuchi Y, Koda Y, Nagano AJ, et al. 2016. Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. *Nat Commun* **7**: 13295.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476-481.

Hunkapiller T, Kaiser RJ, Koop BF, Hood L. 1991. Large-scale and automated DNA sequence determination. *Science* **254**: 59–67.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436**: 793–800.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. *De novo* assembly and

genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44**: 226–232.

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, Ohyanagi H, Mineta K, Michell CT, Saber N, et al. 2017. The genome of *Chenopodium quinoa. Nature* **542**: 307–312.

Jiao WB, Garcia Accinelli G, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, et al. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* **27**: 778-786.

Jiao WB, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Curr Opin Plant Biol* **36**: 64–70.

Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, Chuang H-Y, Kruglyak S, Ronaghi M, Eberle M a, et al. 2013. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci* **110**: 5552–5557.

Karl R, Koch MA. 2013. A world-wide perspective on crucifer speciation and evolution: Phylogenetics, biogeography and trait evolution in tribe Arabideae. *Ann Bot* **112**: 983–1001.

Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al. 2016. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**: 492–505.

Keller O, Odronitz F, Stanke M, Kollmar M, Waack S. 2008. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* **9**: 278.

Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al. 2016. Ensembl Genomes 2016: More genomes, more complexity. *Nucleic Acids Res* **44**: D574–D580.

Kiefer M, Schmickl R, German DA, Mandáková T, Lysak MA, Al-Shehbaz IA, Franzke A, Mummenhoff K, Stamatakis A, Koch MA. 2014. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant Cell Physiol* **55**: e3.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013a. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and

gene fusions. *Genome Biol* **14**: R36.

Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA, et al. 2013b. Reference-assisted chromosome assembly. *Proc Natl Acad Sci* **110**: 1785–1790.

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J, et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data* **1**: 140045.

Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, et al. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**: 59–63.

Koch MA, Kiefer M, German DA, Al-Shehbaz IA, Franzke A, Mummenhoff K, Schmickl R. 2012. BrassiBase: Tools and biological resources to study characters and traits in the Brassicaceae—version 1.1. *Taxon* **61**: 1001–1009.

Koren S, Phillippy AM. 2015. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* **23**: 110–120.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**: 693–700.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Kosambi DD. 1943. The estimation of map distances from recombination values. *Ann Eugen* **12**: 172–175.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, et al. 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30**: 771–776.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The *Arabidopsis* Information

Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* **40:** D1202-1210.

Lander ES, Botstein S. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185-199.

Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**: 566–568.

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, et al. 2015a. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat Biotechnol* **33**: 524–530.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Li R, Hsieh C-L, Young A, Zhang Z, Ren X, Zhao Z. 2015b. Illumina synthetic long read sequencing allows recovery of missing sequences even in the “finished” *C. elegans* Genome. *Sci Rep* **5**: 10814.

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, et al. 2014. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**: 1045–1052.

Lieberman-aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC, Pop M. 2012. AGORA: Assembly Guided by Optical Restriction Alignment. *BMC Bioinformatics* **13**: 189.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods* **12**: 733–735.

Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T, Ma H. 2012. Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Res* **22**: 508–518.

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. 2009. The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol Biol Evol* **26**: 85–98.

Lysak MA, Mandáková T, Schranz ME. 2016. Comparative paleogenomics of crucifers: Ancestral genomic blocks revisited. *Curr Opin Plant Biol* **30**: 108–115.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**: 2878–2879.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Martin NH, Willis JH. 2007. Ecological divergence associated with mating system causes nearly complete reproductive isolation between sympatric mimulus species. *Evolution* **61**: 68–82.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**.

Melodelima C, Lobréaux S. 2013. Complete *Arabis alpina* chloroplast genome sequence and insight into its polymorphism. *Meta Gene* **1**: 65–75.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14**: R10.

Moncada P, Martínez CP, Borrero J, Chatel M, Gauch J, Guimaraes E, Tohme J,

McCouch SR. 2001. Quantitative trait loci for yield and yield components in an *Oryza sativa* × *Oryza rufipogon* BC2F2 population evaluated in an upland environment. *Theor Appl Genet* **102**: 41–52.

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, et al. 2016. A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat Methods* **13**: 587–590.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet* **14**: 157–167.

Nagarajan N, Read TD, Pop M. 2008. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24**: 1229–1235.

Neale DB, Wegrzyn JL, Stevens KA, Zimin A V, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* **15**: R59.

Nordström KJ V, Albani MC, James GV, Gutjahr C, Hartwig B, Turck F, Paszkowski U, Coupland G, Schneeberger K. 2013. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat Biotechnol* **31**: 325–330.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2009. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana. Science* **327**: 92-94.

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, et al. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**: 780–786.

Potato T, Sequencing G. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189-195.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res* **26**: 342–350.

Qi J, Chen Y, Copenhaver GP, Ma H. 2014. Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc Natl Acad Sci* **111**: 10007–10012.

Quick J, Quinlan AR, Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *Gigascience* **3**: 22.

Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K. 2015. Improving the annotation of *Arabidopsis lyrata* using RNA-Seq data. *PLoS One* **10**.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Roach JC, Boysen C, Wang K, Hood L. 1995. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**: 345–353.

Salman-Minkov A, Sabath N, Mayrose I. 2016. Whole-genome duplication as a key factor in crop domestication. *Nat Plants* **2**: 16115.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson a R, Fiddes C a, Hutchison C a, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687–695.

Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, et al. 2014. Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol* **15**: 506.

Schatz MC, Witkowski J, McCombie WR. 2012. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol* **13**: 243.

Scherthan H, Cremer T, Arnason U, Weier H-U, Lima-de-Faria A, Fronicke L. 1994. Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat Genet* **6**: 342–347.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. 2009. Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**: R98.

Schranz ME, Lysak MA, Mitchell-olds T. 2006. The ABC ' s of comparative genomics in the Brassicaceae : building blocks of crucifer genomes. *Trends in Plant Sci* **11**: 535-542.

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. 1993. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**: 110–114.

Selvaraj S, R Dixon J, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**: 1111–1118.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45**: 831-835

Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* **16**: 344–358.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**: ii215-ii225.

Stebbins LG. 1957. Self Fertilization and Population Variability in the Higher Plants. *Am Nat* **91**: 337.

Sutton T, Baumann U, Hayes J, Collins NC, Shi B, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, et al. 2007. Boron-toxicity tolerance in transporter

amplification. *Science* **318**: 1446–1449.

Swerdlow H, Wu S, Harke H, Dovichi NJ. 1990. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr A* **516**: 61–67.

Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli K-P, O'Brien SJ. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *Gigascience* **5**: 38.

Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, Gentzbittel L, Childs KL, Yandell M, Gundlach H, et al. 2014. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**: 312.

Tang H, Lyons E, Town CD. 2015. Optical mapping in plant comparative genomics. *Gigascience* **4**: 3.

Tank DC, Olmstead RG. 2008. From annuals to perennials: Phylogeny of subtribe Castillejinae (Orobanchaceae). *Am J Bot* **95**: 608–625.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28**: 511–515.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604.

Valouev A, Schwartz DC, Zhou S, Waterman MS. 2006. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci* **103**: 15770–15775.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**: 508–511.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B,

Fan HC, Mantalas GL, Palmeri KJ, et al. 2013. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2013**: 1–24.

Wang D, Graef GL, Procopiuk AM, Diers BW. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theor Appl Genet* **108**: 458–467.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* **43**: 1035–1039.

Watson JD, Crick FHC. 1953. The Structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**: 123–131.

Weisenfeld NI, Kumar V, Shah P, Church D, Jaffe DB. 2016. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Wijnker E, James GV, Ding J, Becker F, Klasen JR, Rawat V, Rowan BA, de Jong DF, de Snoo CB, Zapata L, et al. 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* **2013**: 1–22.

Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJ V, Becker C, Warthmann N, Chica C, Szarzynska B, et al. 2015. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat Plants* **1**: 14023.

Wright SI, Kalisz S, Slotte T, B PRS, Wright SI, Kalisz S, Slotte T. 2013. Evolutionary consequences of self-fertilization in plants. *Proc R Soc B Biol Sci* **280**: 1-10

Yang J, Liu D, Wang X, Ji C, Cheng F. 2016. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat Genet* **48**: 1225-1232

Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C, et al. 2013. The reference genome of the halophytic plant *Eutrema salsugineum*. *Plant Genet Genomics* **4**: 46.

Zamir D. 2001. Improving plant breeding with exotic genetic libraries. *Nat Rev Genet* **2**: 983–989.

Zapata L, Ding J, Willing E, Hartwig B, Bezdan D, Jiao W, Patel V, Velikkakam James

G, Koornneef M, Ossowski S, et al. 2016. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci* **113**: E4052–E4060.

Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song J-M, Xie W, et al. 2016. Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. *Proc Natl Acad Sci* **113**: E5163-E5171

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol* **33**: 531–537.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**: 303–311.

Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, Runnheim R, Forrest DK, Goldstein S, Livny M, et al. 2007. Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**: 278.

Publication

Jiao W-B, Garcia Accinelli G, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, et al. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* **27**: 778-786.

Jiao W-B, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Curr Opin Plant Biol* **36**: 64–70.

Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, Patel V, Velikkakam James G, Koornneef M, Ossowski S, et al. 2016. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci* **113**: E4052–E4060.

Gutjahr C, Gobbato E, Choi J, Riemann M, Johnston MG, Summers W, Carbonnel S, Mansfield C, Yang S-Y, Nadal M, et al. 2015. Rice perception of symbiotic arbuscular mycorrhizal fungi requires the karrikin receptor complex. *Science* **350**: 1521–1524.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegt worden ist, sowie das ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.
Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. George Coupland und Dr. Korbinian Schneeberger betreut worden.

.

Ich versichere, dass ich alle Angaben wahrheitsgemäß nach bestem Wissen und Gewissen gemacht habe und verpflichte mich, jedmögliche, die obigen Angaben betreffenden Veränderungen, dem Dekanat unverzüglich mitzuteilen.

…….. ……

Datum Unterschrift

Wenbiao Jiao

Lebenslauf

Angaben zur Person

Familienname / Vornamen: JIAO / WENBIAO

Geburtsdatum: 08. Sep, 1987

Geburtsort: Xianning, Hubei, VR China

Staatsangehörigkeit: Chinesisch

E-mail: jiao@mpipz.mpg.de

Ausbildung

- Jan. 2014- Doktorand, Biologie, Universität zu Köln und Max Planck Institut für Pflanzenzüchtungsforschung, Köln. Thema der Dissertation: High-quality genome assemblies of plant species using third generation genomic technologies and comprehensive genotypic characterization of *Arabis* inter-species introgression lines.
- Sep. 2010 – Juni. 2013 Master-Abschluss in der Wissenschaft Major: Biochemie und Molekularbiologie, College of Life Wissenschaft und Technologie, Huazhong Landwirtschaft Universität, Wuhan, China
- Sep. 2006 – Juni. 2010 Bachelor-Abschluss in der Wissenschaft Major: Biologie Wissenschaft, College of Life Wissenschaft und Technologie, Huazhong Landwirtschaft Universität, Wuhan, China
- Sep. 2003 – Juni. 2006 E'nan Senior Schule, Xianning, China
- Sep. 2000 – Juni. 2003 Tongshan Shiyan Junior Schule, Tongshan, China

…….. ……

Datum Unterschrift