6 Sequencing, Assembly, and Annotation of the Alfalfa Genome

Joann Mudge and Andrew D. Farmer

Abstract

While the alfalfa community originally relied on *Medicago truncatula* (especially the reference assembly, A17) for genomic resources, recent changes in sequencing and scaffolding technologies and algorithms have enabled the sequencing and assembly of five different alfalfa accessions, to date. These assemblies include two diploid assemblies, CADL and PI464715, as well as three tetraploid assemblies, NECS-141, Zhongmu No. 1, and XinJiangDaYe. Technological changes within the approximately half a decade over which these assemblies were produced, have allowed for increasingly contiguous assemblies and improved scaffolding resulting in chromosome level assemblies that allow for the detection of large-scale structural rearrangements. They have also made possible the assembly of all four subgenomes of the tetraploid in the XinJiangDaYe assembly. While subgenome haplotypes were very similar and sometimes indistinguishable, nevertheless, structural differences between haplotypes were uncovered. These included local differential gene content between subgenome haplotypes as well as larger structural variants such as inversions. Compared to the *M. truncatula* assembly and annotation, the approximately 75% increase in genome size in alfalfa is mainly due to the expansion of repeats. The availability of five different annotated alfalfa genome assemblies, including those of both diploid and tetraploid accessions, will be a significant asset to the alfalfa community.

6.1 Introduction

6.1.1 The Alfalfa Genome

Plants have very dynamic genomes making plant genome assembly especially challenging. Flexibility and instability in plant genomes is reflected in genome size expansion and contraction and higher rates of polyploidy, heterozygosity, repeats, and pseudogenes compared to eukaryotic organisms from other kingdoms (Schatz et al. 2012; Jiao and Schneeberger 2017).

The alfalfa genome is no exception. Obvious sources of alfalfa genome complexity include autopolyploidy and high rates of heterozygosity. Both heterozygosity and polyploidy can lead to diverging haplotypes that complicate assembly.

J. Mudge (✉) · A. D. Farmer
Department of Bioinformatics, National Center
for Genome Resources, Santa Fe, NM, USA
e-mail: jm@ncgr.org

A. D. Farmer
e-mail: adf@ncgr.org

L.-X. Yu and C. Kole (eds.), *The Alfalfa Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-74466-3_6

The obligate outcrossing reproductive mechanism of alfalfa ensures that heterozygosity rates remain high and polypoidy provides further opportunities for haplotype diversity. While the *Medicago sativa* complex includes both diploid and tetraploid forms, cultivated alfalfa is tetraploid. Most cultivars belong to the *sativa* subspecies or the *varia* subspecies, which represents introgressions of the *falcata* subspecies into the *sativa* subspecies. However, a few cultivars, especially those harboring cold tolerance, are from the subspecies *falcata* (Veronesi et al. 2010).

The autotetraploid genome of cultivated alfalfa allows for up to four different subgenome haplotypes, with the number of distinguishable haplotypes at a locus varying across the genome. In contrast to allopolyploids, whose subgenomes originate from different progenitor species' genomes that are typically relatively divergent, autopolyploids have chromosomes doubled from genomes within the same species and may allow recombination among the homoeologues. Tetraploidy also results in a large genome size requiring an increased sequencing volume to achieve the same genome coverage. While alfalfa has a base (haploid) chromosome number of 8 and a base genome size of ~800 Mb, cultivated and some wild alfalfa species have 32 chromosomes and ~3.2 Gb in its tetraploid genome (Blondon et al. 1994).

Original genomic analyses used the congeneric *Medicago truncatula* as a model (Yang et al. 2008; Young et al. 2011). *M. truncatula* has a smaller genome size (~450 Mb). In addition, this diploid plant has a high rate of selfing (Barker et al. 1990; Cook 1999) resulting in a low heterozygosity rate that makes assembly easier and lowers coverage requirements. Recent advances in sequencing and scaffolding technologies have lowered cost and increased throughput. These advances, along with improved assembly algorithms, and have recently made directly sequencing and assembling plant genomes, including polyploid genomes, more feasible (Mishra et al. 2017; Kyriakidou et al. 2018; Jung et al. 2019; Michael and VanBuren 2020).

6.1.2 Changing Technologies

Long-read sequencing technologies, including Pacific Biosciences Single Molecule Real-time or SMRT sequencing (PacBio) and Oxford Nanopore sequencing (ONT), have vastly improved our ability to generate reference-quality plant genomes, with relative ease and low cost compared to Sanger sequencing. These technologies produce higher quality assemblies compared to short-read or short and long-read hybrid assemblies (Eid et al. 2009; Deamer et al. 2016; Jiao and Schneeberger 2017). Plant assemblies generated based on long-read sequences first began to appear in 2015, with *M. truncatula* having some of the earliest PacBio-based plant assemblies (Berlin et al. 2015; VanBuren et al. 2015; Moll et al. 2017). PacBio-based plant assemblies showed much improved continuity, fewer gaps, and captured more of the genome compared to assemblies based on short reads (VanBuren et al. 2015; Moll et al. 2017; Jiao and Schneeberger 2017). But these assemblies still struggled to span long, closely related repeats or efficiently navigate differing levels of haplotype divergence. This is in part due to the fact that sequence error rates were higher than the divergence levels that needed to be discriminated in order to resolve these types of elements. The use of correction strategies based on consensus among reads taken from different molecules made it difficult to discriminate between closely related repeats or haplotypes, as they would often be lumped together during correction, forming a single chimeric consensus sequence.

But the recent transition from PacBio CLR (continuous long read) to HiFi (high fidelity) reads (Wenger et al. 2019), in which high accuracy consensus (>99%) is achieved by utilizing correction based on multiple sequencing passes on the same template molecule, has improved our ability to discriminate between closely related repeats or slightly diverged haplotypes. PacBio HiFi reads, because of their accuracy, require reduced consensus read coverage for assembly. Reduced coverage combined with increased throughput on the Sequel II, mean that higher quality plant assemblies can be obtained for a lower cost with reduced computational requirements and time compared to the original PacBio-based assemblies. In addition, the high read accuracy makes it possible to distinguish alternate haplotypes and similar but not identical repeat sequences, increasing continuity and improving our ability to phase haplotypes.

Oxford Nanopore Technology (ONT) is another long read technology. Since first conceptualized in the late 1980s, ONT has made recent advances in both length and accuracy (Deamer et al. 2016). In just a few short years, ONT-based assemblies have gone from bacteria (Deschamps et al. 2016) to higher organisms, including plants (Michael et al. 2018; Belser et al. 2018; Deschamps et al. 2018). A recent study on the comparison of PacBio HiFi and ONT in rice (Lang et al. 2020) indicated that while PacBio's high read accuracy enabled higher accuracy at the nucleotide level, including fewer artificial SNPs and small indels in the assembly, the longer ONT read length (up to 2 Mb) enabled higher assembly continuity and better spanning of repetitive genomic sections and resolution of gene family copy number. While both technologies were able to assemble some rice chromosomes in a single contig, ONT technology captured more chromosome length contigs (10 compared to 3 with PacBio HiFi) and 7 of these appeared to be gapless assemblies extending into telomeres on either end. The two technologies appear to complement each other with PacBio delivering high continuity and accuracy and ONT delivering even higher continuity tempered by a small reduction in accuracy.

Whichever long-read technology is used for plant genome assembly, additional scaffolding technologies are often applied to improve contiguity of the assembly. Recently, new technologies have replaced more expensive and cumbersome methods of scaffolding such as physical and genetic maps. Long-range, whole genome scaffolding technologies, including optical mapping and chromatin conformation technologies provide high-throughput and relatively inexpensive methods to scaffold contigs together, improving contiguity, often to the pseudo-chromosome level (Burton et al. 2013; O'Bleness et al. 2014; Steinberg et al. 2014; Mostovoy et al. 2016; Staňková et al. 2016).

In less than a decade, technology improvements have allowed the alfalfa community to move from reliance on *M. truncatula* genome assemblies to sequencing the alfalfa genome directly. To date, the community has generated five publicly available alfalfa genome assemblies. The sequencing and assembling of these genomes have been pursued at different times across a rapidly changing technological background, providing an interesting view into not only how changes in technology and strategies affect genome assemblies, but also elucidating structural challenges inherent in the alfalfa genome.

6.2 Genome Assemblies

6.2.1 Diploid Assemblies

6.2.1.1 Cultivated Alfalfa at the Diploid Level

The first genome assembly was generated from a plant from the Cultivated Alfalfa at the Diploid Level (CADL) population. This population is a stable diploid alfalfa population that is able to reproduce by seed (Bingham and McCoy 1979). It took advantage of diploid cultivated alfalfa germplasm generated by the *4x* − *2x* cross method (Bingham 1969). Though fertility of the diploid lines was low, crossing them as females to diploid *M. sativa* subspecies *falcata* lines improved fertility of the F_1, allowing backcrossing to the *2x* diploids. This resulted in a stable diploid, fertile population whose germplasm is estimated to be derived from at least 98% cultivated germplasm. A single, clonally propagated plant was chosen from the CADL population for genome sequencing and assembly to avoid any interplant variability.

Using a diploid plant for sequencing and assembly reduced the amount of sequence data needed and the complexity of the assembly, which should lead to a high-quality assembly while requiring fewer computational resources.

The eventual goal was to use this assembly as a scaffold for assembling a tetraploid genome. Whole-genome PacBio continuous long read (CLR) sequencing was begun in early 2015 with a preliminary version of the assembly publicly released in mid-2016. Just over 100X subread coverage (based on an 800 Mb genome size) or just over 50X coverage per haplotype was generated. Subreads, sequencing reads resulting from each of the multiple sequencing passes of a DNA fragment, had a mean length of 8.0 kb and an N50 length of 13.1 kb (Table 6.1).

Several assembly iterations were tried, though computational constraints made it difficult to test assembly strategies extensively. The current assembly version (version 1.0) was generated as follows. DAligner (Myers 2014) was used to align reads. Using these alignments, Falcon (Chin et al. 2016) was used to assemble the reads. The resulting assembly was polished using Quiver (https://github.com/PacificBiosciences/GenomicConsensus). The polished assembly was scaffolded with long-distance maps generated from chromatin conformation Dovetail libraries using the HiRise algorithm (Putnam et al. 2016). A final polish with Quiver completed the assembly.

The resulting assembly was fragmented (5,753 pieces) but, nevertheless, contained most of the genespace (96.7%) based on capture of single-copy eukaryotic orthologous genes with Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al. 2015) (Tables 6.1 and 6.2), which is a surrogate for overall gene capture. Even with the fragmentation, much of the assembly was in megabase-sized pieces with a contig N50 of 1.27 Mb (half of the assembly is in pieces of 1.27 Mb or larger). While still far short of expected chromosome sizes, this is, nevertheless, an important improvement over the short-read plant genome assemblies that had previously dominated.

The total assembly size of 1,200 Mb is approximately 50% larger than the expected 800 Mb base genome size. This is due to the assembly of multiple haplotypes in some, but not all regions of the genome. In comparing haplotypes that were divergent enough to assemble separately, it became clear that different haplotypes of this diploid genome were often missing genes from the syntenic haplotype (Fig. 6.1a). Therefore, the full gene complement was not present in a single haplotype. This might be an artifact of creating a diploid from cultivated autotetraploid germplasm. Upon plant whole-genome duplication that results in an autopolyploid, differential gene loss between haplotypes can occur (Doyle et al. 2008; Hufton and Panopoulou 2009). The assembly of multiple haplotypes in about half of the genome is confirmed by the BUSCO results, with 57.4% of the typical single-copy orthologs duplicated (Table 6.2 and Fig. 6.2). The CADL assembly shows good coverage of the *M. truncatula* genome (Fig. 6.3). Regions of the assembly showing one or two haplotypes assembled are visible as double versus single diagonals (Fig. 6.4).

Table 6.1 Sequence read and assembly statistics for the five alfalfa genome assemblies

Accession	Ploidy	Sequencing technology	Read N50 (kb)	Scaffolding technology	Scaffold length (Mb)	Contig N50 (Mb)	Scaffold N50 (Mb)
CADL	*2x*	PacBio	13.1	Dovetail	1,251	1.27	1.27
PI464715	*2x*	Oxford Nanopore	27.9	Hi-C	793	3.86	102.49
NECS-141	*4x*	PacBio	17.4	BioNano	2,698	0.22	2.21
Zhongmu No. 1	*4x*	PacBio	12.2	BioNano and Hi-C	817	3.92	102.29
XinJiangDaYe	*4x*	PacBio HiFi	12.6	Hi-C	3158	0.46	84.27

Table 6.2 Gene statistics for the five alfalfa genome assemblies

Accession	Ploidy of source	Assembly ploidy[a]	Protein coding genes (thousands)	BUSCO database	Complete BUSCO Genes (%)	Complete and duplicated BUSCO Genes (%)
CADL	*2x*	*2x*	111	eudicotyledons_odb10	96.7	57.4
PI464715	*2x*	*1x*	47	embryophyta_odb10	97.7	8.7
NECS-141	*4x*	*4x*	103	eudicotyledons_odb10	95.7	75.8
Zhongmu No. 1	*4x*	*1x*	50	embryophyta_odb9	93.3	5.5
XinJiangDaYe	*4x*	*4x*	165	unavailable	97.2	90.1

[a]Upper estimate as some haplotypes were collapsed in assemblies at more than 1X

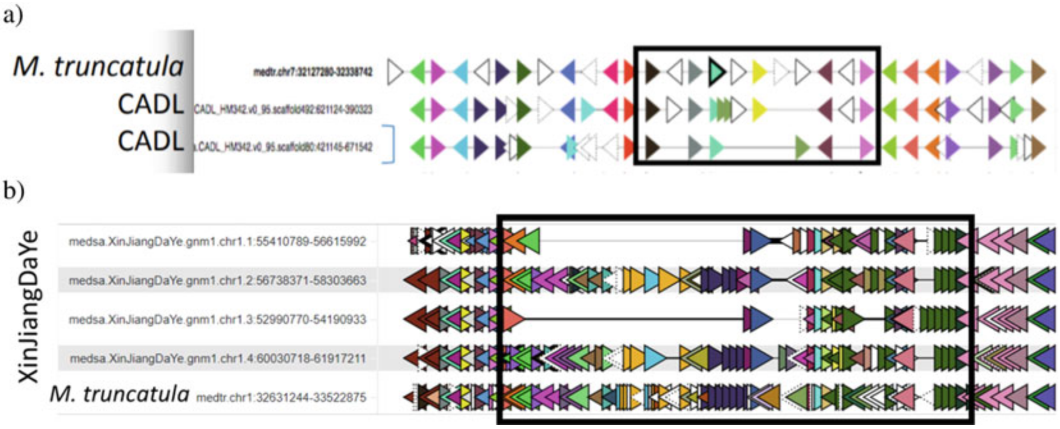
a)
M. truncatula
CADL
CADL
medtr.chr7:32127280-32338742
CADL_HM342.v0_95.scaffold492:621124-390323
CADL_HM342.v0_95.scaffold80:421145-671542
b)
XinJiangDaYe
medsa.XinJiangDaYe.gnm1.chr1.1:55410789-56615992
medsa.XinJiangDaYe.gnm1.chr1.2:56738371-58303663
medsa.XinJiangDaYe.gnm1.chr1.3:52990770-54190933
medsa.XinJiangDaYe.gnm1.chr1.4:60030718-61917211
M. truncatula
medtr.chr1:32631244-33522875

Fig. 6.1 Comparison of alfalfa genomic regions to syntenic *Medicago truncatula* regions. Triangles represent genes with orientation indicated by the direction of the pointed side. Genes are colored by gene family. The boxed regions show synteny breaking down through the differential loss of genes between haplotypes in alfalfa assembly. **a** *M. truncatula* (top) chromosome 7 compared to CADL's two syntenic haplotypes. **b** *M. truncatula* (bottom) chromosome 1 compared to XinJiangDaYe's four syntenic haplotypes

6.2.1.2 Pi464715

In 2020, another diploid alfalfa genome assembly was published (Li et al. 2020). This germplasm with plant introduction (PI) 464715, belongs to *Medicago sativa* subsp. *caerulea* and is thought to be the diploid progenitor of the autotetraploid alfalfa (Small and Jomphe 1989). This wild diploid provides an important contrast to CADL, a diploid derived from a tetraploid.

PI464715 was sequenced and assembled using ONT reads, currently the only alfalfa genome assembly based on ONT technology. ONT reads were corrected with Illumina sequences. With ~145X read coverage of the 800 Mb haploid genome size, the sequencing coverage is higher than that in CADL (just over 100X) and read lengths are longer. The mean read length of 19.7 kb and N50 read length of 27.9 kb are both more than twice those seen in the CADL data (Table 6.1). PI464715 sequence was corrected, assembled, and polished with NextDenovo with additional rounds of correction with Illumina and ONT reads. This resulted in an assembly of 1.35 Gb in length, with part of the genome likely assembled into two haplotypes, as in CADL. Indeed BUSCO duplication rates and alignments to *M. truncatula* for this intermediate assembly support this conclusion (Figs. 6.2 and

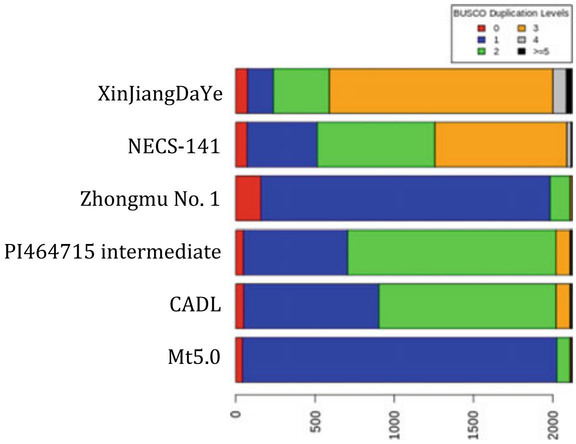

Fig. 6.2 A modified BUSCO analysis was run on each of the five alfalfa genomes and *M. truncatula* that enabled counting of duplication number for each captured BUSCO. To facilitate haplotype analyses, the version of the PI464715 assembly before haplotypes were collapsed and before scaffolding was used, which the authors kindly made available, rather than the final version of the assembly. The analysis was run with BUSCO 3.1.0 in genome mode using the eudicotyledons_odb10 database

6.5). Finally, duplicate haplotypes were removed using purge_haplotigs, which collapsed the assembly to 793.2 Mb in length, consistent with the haploid genome size (Table 6.1). The resulting assembly had a contig N50 of 3.86 Mb, approximately 3-fold that of CADL (Table 6.1). Long-range scaffolding of the assembly was accomplished with Hi-C data using LACHESIS. The final assembly consisted of 355 contigs scaffolded into 8 pseudo-chromosomes that cover 98.5% of the assembly and captured 97.7% of BUSCO gene orthologs (Table 6.2).

The final assembly covered the *M. truncatula* genome well (Fig. 6.6). Assembly contiguity is high enough that it is easy to see the chromosome 4/8 translocation, known to have occurred in the A17 accession of *M. truncatula,* as well as several small inversions (Figs. 6.6 and 6.7). The presence of only one haplotype in the final assembly is supported by the lack of double diagonals when compared to *M. truncatula* (Figs. 6.6 and 6.7) as well as by the low number of duplicate BUSCO genes (8.7%) (Table 6.2). It is interesting to note that the initial assembly length (1.35 Gb) before haplotypes were collapsed and scaffolding was run was very close in size to that of CADL (1.25 Gb), indicating that a similar proportion of the genome had diverged enough to assemble haplotypes independently despite a difference in sequencing technology and assembly strategy.

6.2.2 Tetraploid Assemblies

6.2.2.1 NECS-141

The first tetraploid alfalfa genome to be sequenced was that of NECS-141, a semi-dormant breeding line developed in Iowa (Khu et al. 2010). While originally meant to be a hybrid PacBio and Illumina assembly, because of the complications of the genome such as the repeat and ploidy structure, additional PacBio data was obtained as costs of the technology came down and accuracy, read length, and assembly algorithms improved.

The PacBio CLR sequencing reads were obtained in 2014 and 2015, around the time that fully PacBio plant genome assemblies were first being contemplated. It was sequenced around the same time as CADL with slightly higher coverage overall (~115X per haplotype vs ~100X in CADL) but lower coverage per haplotype (~29X vs ~50X for CADL). It is not surprising that it, too, is fragmented, with lack of continuity exacerbated by the increase in ploidy compared to CADL. However, scaffolding with BioNano

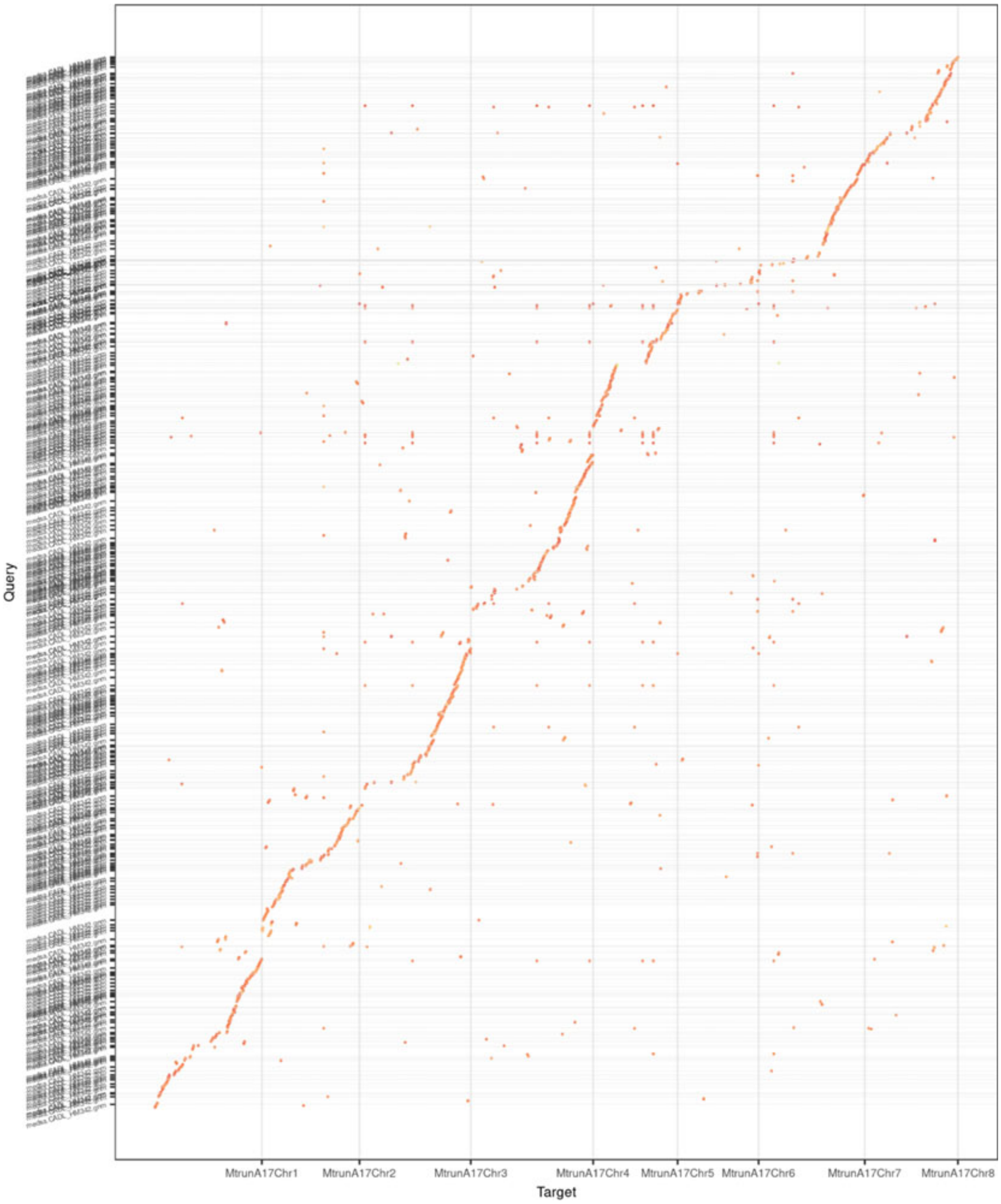

Fig. 6.3 Dotplot comparing the CADL assembly (y-axis) to the eight *Medicago truncatula* v. 5.0 chromosomes (x axis). Nucleotide level alignments were generated with minimap2 (Li 2018) using the asm20 preset, which allows up to 5% sequence divergence. Dotplots were generated using dotplotly with a minimum query length of 50 kb and a minimum alignment length of 10 kb (https://github.com/tpoorten/dotPlotly)

optical maps merged it into fewer pieces, though it still contained approximately twice the number of pieces as CADL, but with a scaffold N50 that exceeded that of CADL. More specifically, BioNano scaffolding was able to collapse the approximately 67 k contigs (N50 = 221 kb) into

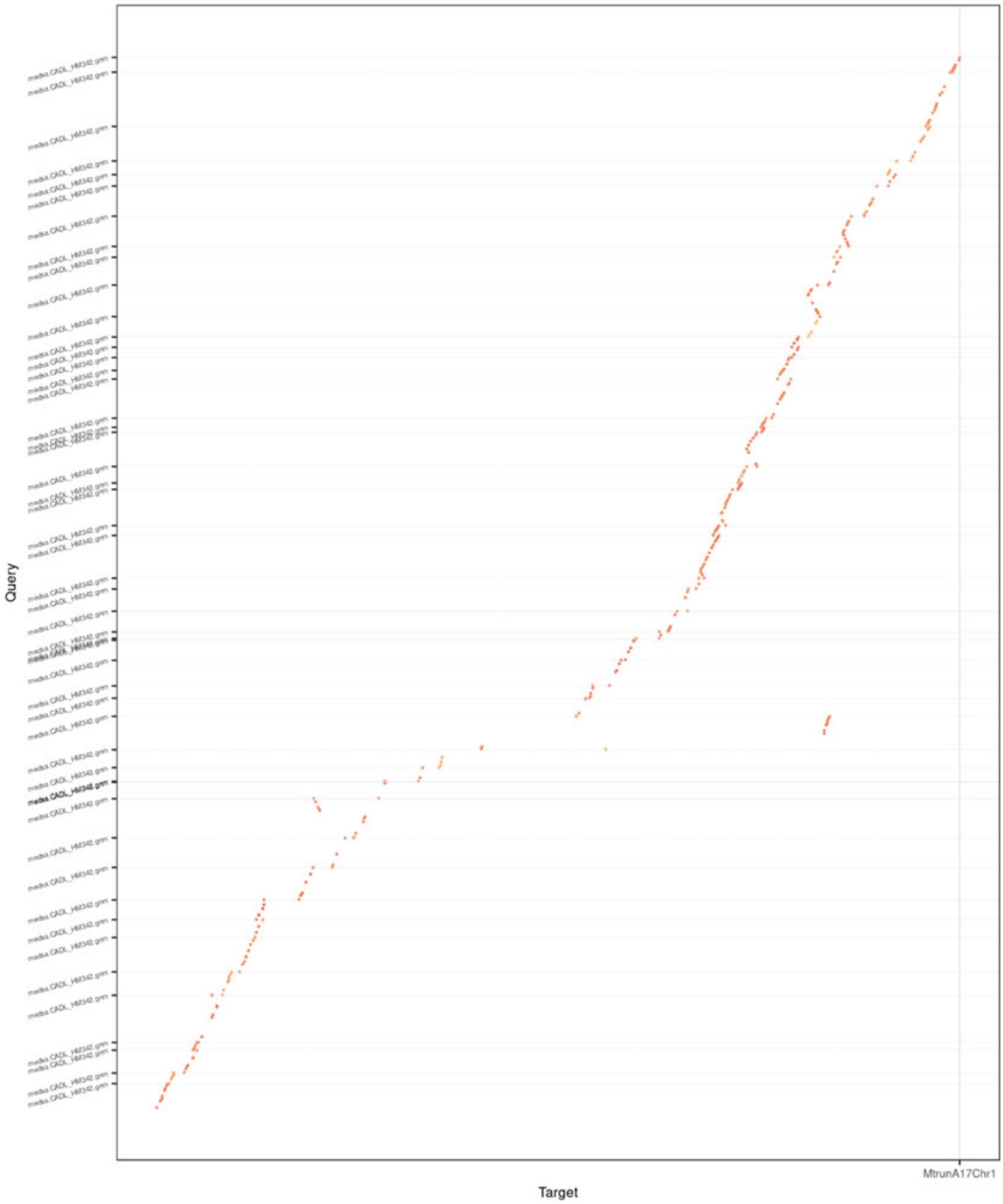

Fig. 6.4 Dotplot comparing CADL (y-axis) to *Medicago truncatula* v. 5.0 chromosome 1 (x-axis) showing capture of differing numbers of haplotypes across the genome, likely due to differing levels of haplotype divergence. Dotplots were generated as described in Fig. 6.3

just under 10 k scaffolds (N50 = 2.2 Mb), increasing the N50 by 10-fold (Table 6.1). Hi-C was also obtained but pieces were small enough that the Hi-C assembly was not able to separate out the haplotypes nor resolve local ordering so it was left out of the final assembly (Unpublished). The assembly covers the *M. truncatula* genome well, indicating that it is largely complete (Fig. 6.8).

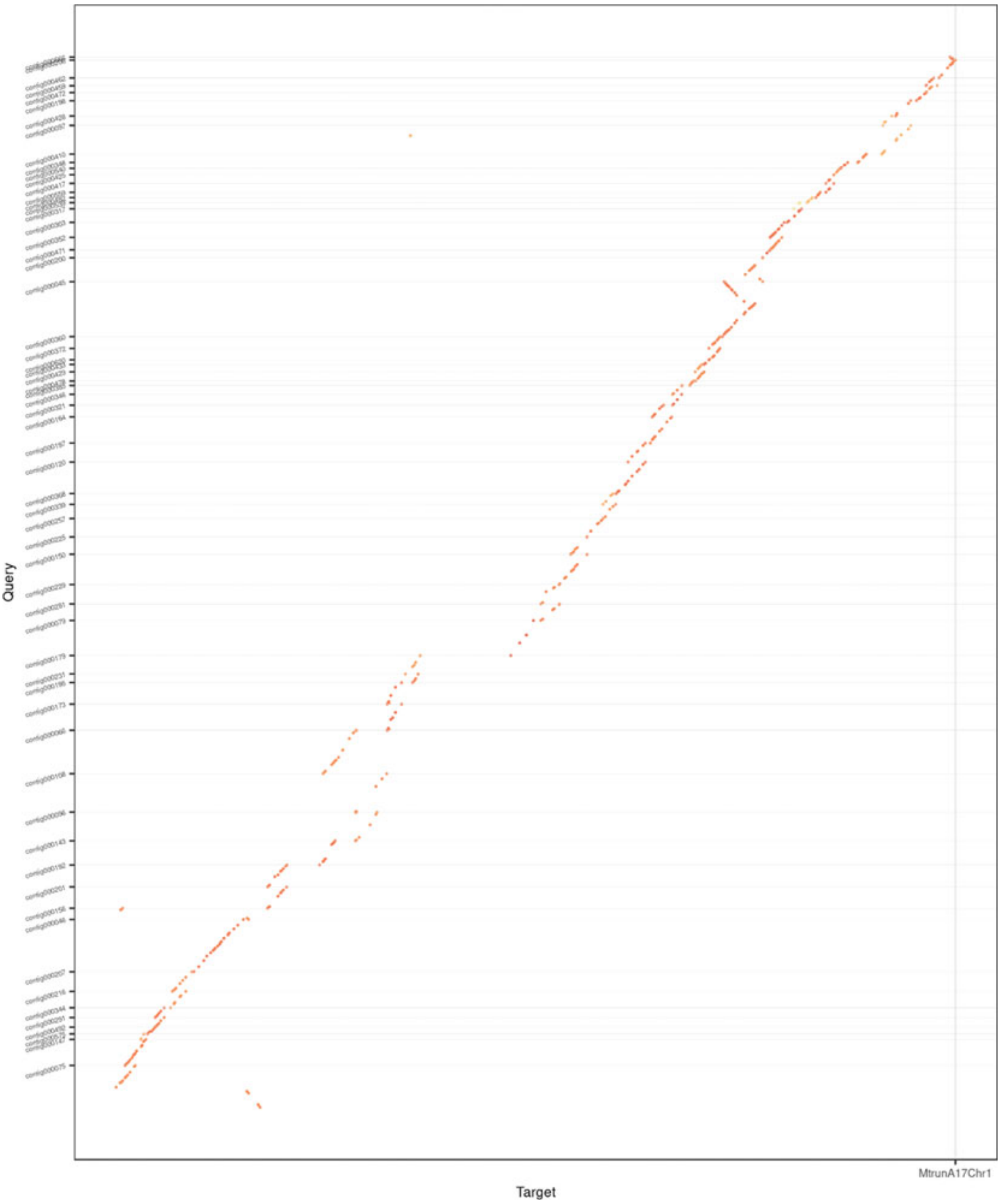

Fig. 6.5 Dotplot comparing a preliminary version of the PI464715 assembly before haplotypes were collapsed and before scaffolding (y-axis) to *Medicago truncatula* v. 5.0 chromosome 1 (x-axis). This shows capture of differing numbers of haplotypes across the genome, likely due to differing levels of haplotype divergence. Dotplots were generated as described in Fig. 6.3

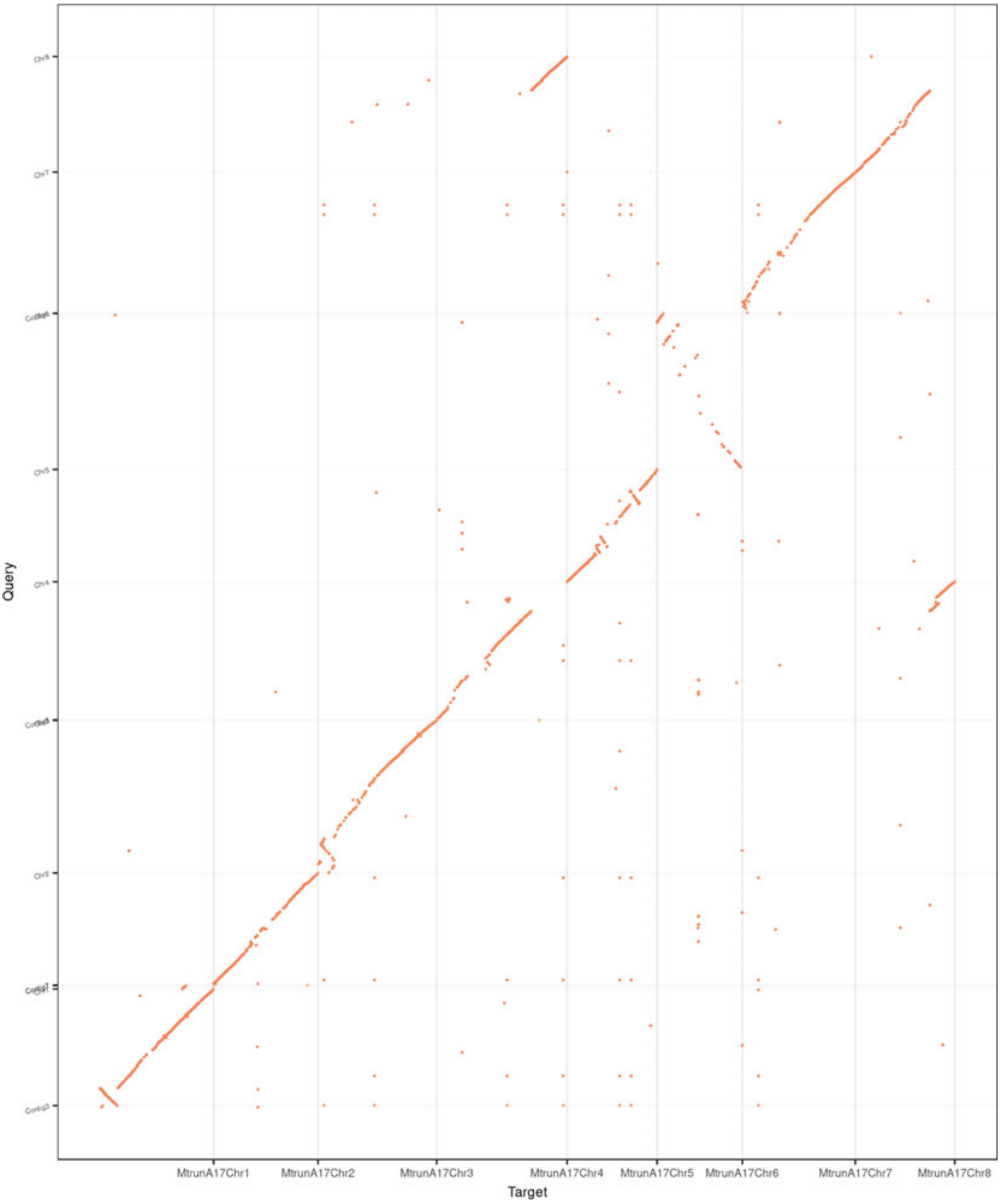

Fig. 6.6 Dotplot comparing the final PI464715 assembly (y-axis) to the eight *Medicago truncatula* v. 5.0 chromosomes (x-axis). Dotplots were generated as described in Fig. 6.3

BUSCO results on the percentage of core eukaryotic orthologs captured reveal some interesting insights (Table 6.2 and Fig. 6.2). As with the diploid genome assemblies, NECS-141 captured the vast majority of BUSCO genes (95.7%) despite its fragmentation. Approximately 76% of these genes were duplicated in the assembly, likely indicating that about three-quarters of the genome had multiple haplotypes assembled. Unlike the diploid genomes, these

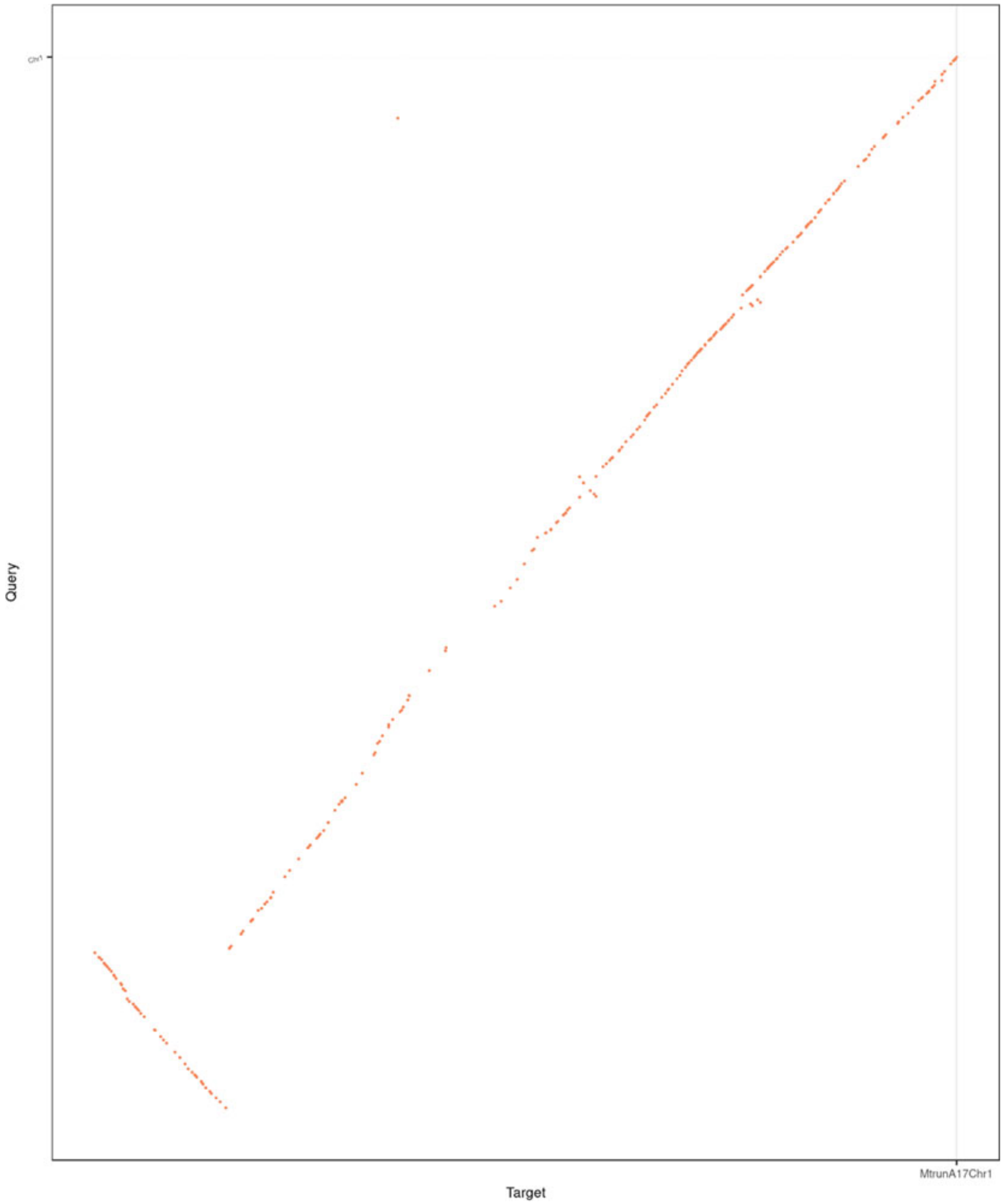

Fig. 6.7 Inversion shown in PI464715 final assembly chromosome 1 (y-axis) compared to *Medicago truncatula* chromosome 1 (x-axis). Dotplots were generated as described in Fig. 6.3

duplicated genes not only included those with two assembled haplotypes but also included slightly more genes with 3 assembled haplotypes, though 4 or more assembled haplotypes were rare (Fig. 6.2). In addition, duplicate or even triplicate haplotypes are visible when aligning to *M. truncatula* (Fig. 6.9). The assembly of multiple haplotypes in some but not genomic regions is further supported by the assembly length (2.70 Gb including 2.35 Gb of

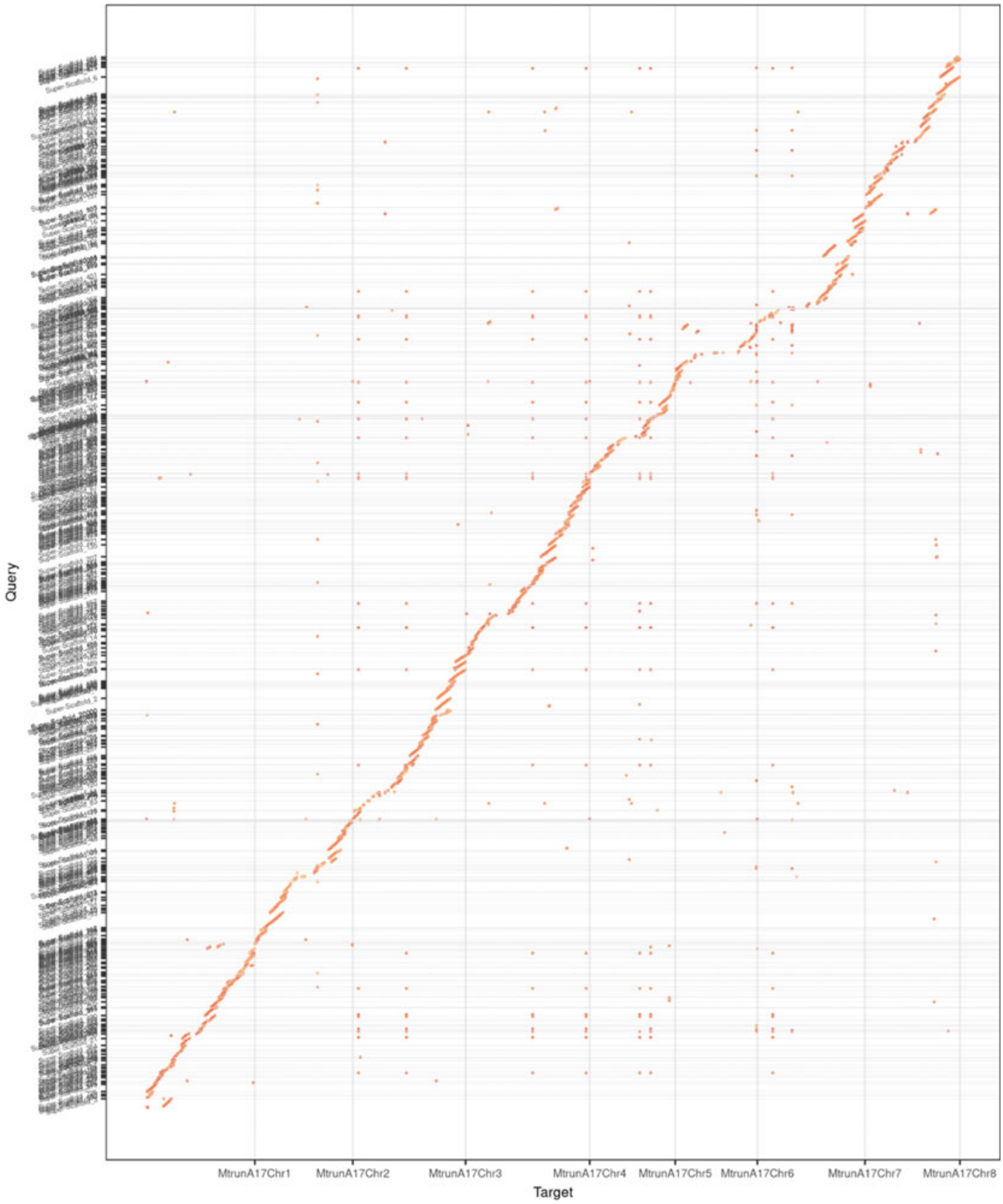

Fig. 6.8 Dotplot comparing the NECS-141 assembly (y-axis) to the eight *Medicago truncatula* v. 5.0 chromosomes (x-axis). Dotplots were generated as described in Fig. 6.3

non-gap sequence), representing approximately 84% and 73% of the expected 3.2 Gb genome covered and captured, respectively, by the assembly (Table 6.1). It appears, therefore, that haplotypes were collapsed to a single version in roughly one-quarter of the genome.

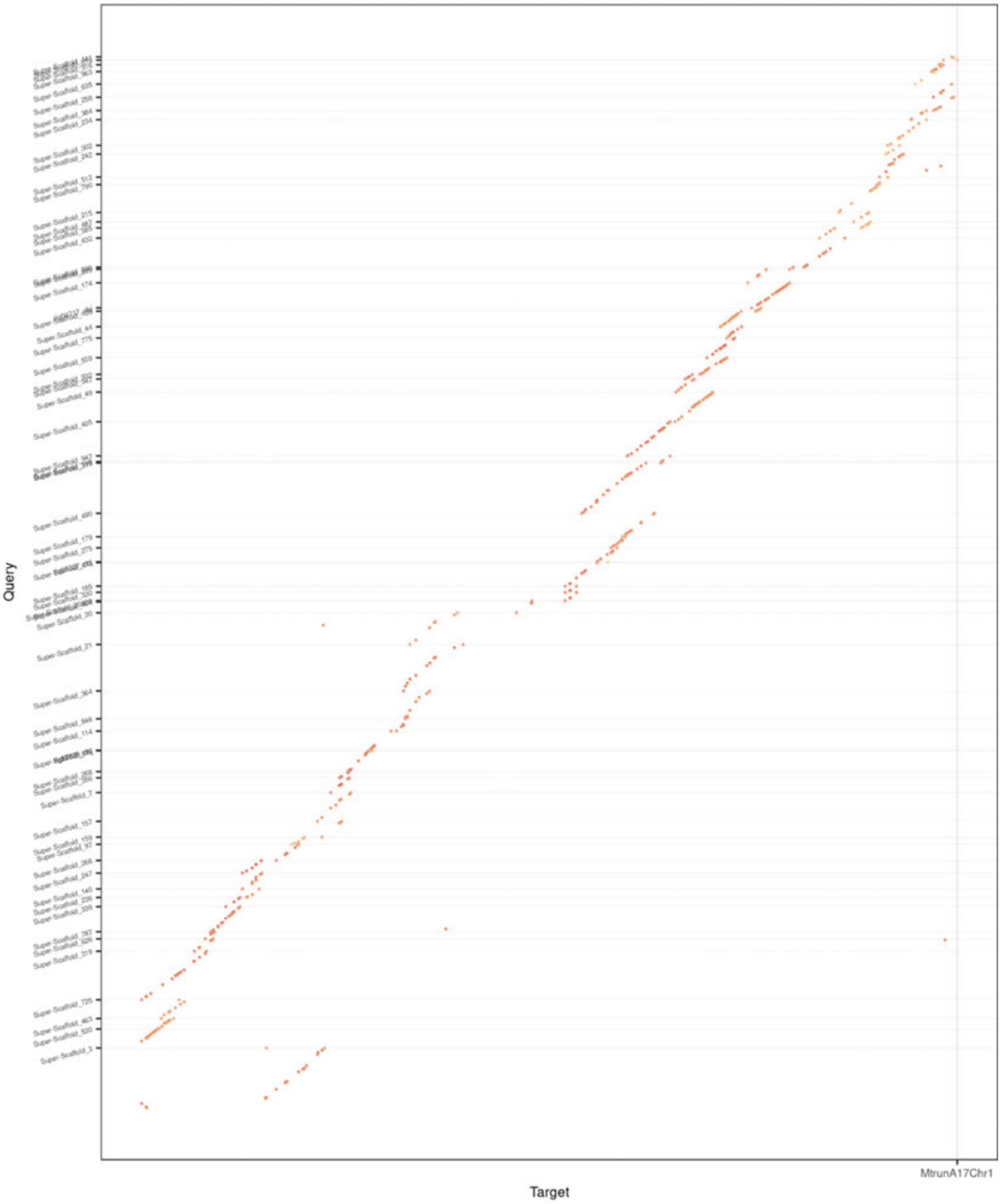

Fig. 6.9 Dotplot comparing NECS-141 (y-axis) to *Medicago truncatula* v. 5.0 chromosome 1 (x-axis) showing capture of differing numbers of haplotypes across the genome, likely due to differing levels of haplotype divergence. Dotplots were generated as described in Fig. 6.3

6.2.2.2 Zhongmu No. 1

A pseudo-chromosome level assembly of the tetraploid cultivar Zhongmu No. 1 was recently published (Shen et al. 2020). Zhongmu No. 1 is a subspecies *sativa* cultivar from Northern China that is salt tolerant (Shi et al. 2017). The continuity at the pseudo-chromosome level is made possible by a combination of PacBio sequencing and BioNano and Hi-C scaffolding.

This assembly is based on PacBio long reads with subread lengths comparable to that of CADL (N50 = 12.1 kb vs 13.1 kb for CADL) (Table 6.1). With approximately 300X coverage (based on the haploid genomic content of 800 Mb) or 75X coverage per haplotype (based on a 3.2 Gb genome size), there was about 3X higher coverage than in CADL. In addition, Illumina data was used to improve assembly accuracy. Finally, BioNano and Hi-C were used to scaffold the assembly.

PacBio data was corrected using Canu (Koren et al. 2017). Corrected PacBio reads were assembled using MECAT (Xiao et al. 2017) and scaffolded with BioNano data. Repeat resolution of the resulting contigs was done with HERA (Du and Liang 2019). Haplotypes were collapsed using Redundans (Pryszcz and Gabaldón 2016) and Purge Haplotigs (Roach et al. 2018). Then Hi-C scaffolding was applied, resulting in a final assembly containing 8 pseudo-chromosomes 816 Mb in length with a contig N50 of 3.9 Mb (Table 6.1). Further refinements to remove low-quality (<Q30) regions and three rounds of genome polishing were done with samtools (Li et al. 2009) and pilon (Walker et al. 2014), respectively. The authors note that this assembly does not match particular subgenomes, but, rather, is a mixture of the subgenomes. Note that this is also likely the case for all the assemblies described here. Though slightly lower than the other assemblies, the Zhongmu No. 1 assembly captured the majority of genes as estimated by BUSCO (93.3%), with most genes captured as single copy (Table 6.2 and Fig. 6.2), reflecting the deredundification step. Dotplots show good coverage of *M. truncatula* and structural variation including the chromosome 4/8 translocation and inversions compared to *M. truncatula* (Figs. 6.10 and 6.11).

6.2.2.3 XinJiangDaYe

Finally, an “allele-aware” tetraploid genome assembly has been published (Chen et al. 2020), using PacBio Circular Consensus Sequencing (CCS) technology and Hi-C scaffolding, that was able to assemble all 32 chromosomes representing the four different haplotypes of each of the 8 base chromosomes. The sequenced accession is XinJiangDaYe, a large-leaved alfalfa cultivar from Xinjiang Provence of northwest China that has good regeneration properties (Zhang et al. 2010; Shi et al. 2017).

Approximately 88X coverage of the haploid complement (800 Mb genome size) or 22X coverage of each haplotype (3.2 Gb genome size) was generated. The PacBio CCS reads were assembled using Canu, yielding an assembly of 3.15 Gb in length with a 459 kb contig N50 (Table 6.1). While this is a relatively fragmented assembly, it is expected that keeping all the haplotypes will yield a lower N50 than that if haplotypes are collapsed, keeping the longest version of each. Furthermore, the high similarity between related haplotypes likely makes it difficult to extend through regions of identity that are longer than the reads. Continuity was improved through scaffolding with Hi-C data using HiC-Pro (Servant et al. 2015) for alignment, removal of cross-allelic connections through manual scripting, and the use of ALLHiC (Zhang et al. 2019) for the Hi-C scaffolding. JuiceBox was used for manual fine-tuning (Durand et al. 2016). A second round of ALLHiC and JuiceBox yielded an assembly length of 2.7 Gb (Table 6.1). Ordering was confirmed with a genetic map. In addition, 200 longest Oxford Nanopore reads (95–263 kb) were mapped to the assembly for confirmation (89% mapped to a single region with at least 80% query coverage).

The assembly length (2.738 Gb) is slightly short of the expected ~3.2 Gb full genome size and only 40 Mb larger than that of NECS-141 (Table 6.1). XinJiangDaYe’s assembly length is approximately 86% of the expected genome size and matches up well with estimates of captured core eukaryotic conserved orthologs (BUSCO) in each haplotype (88.50, 88.30, 87.50, and 87.20%), with 97.2% captured in at least one haplotype

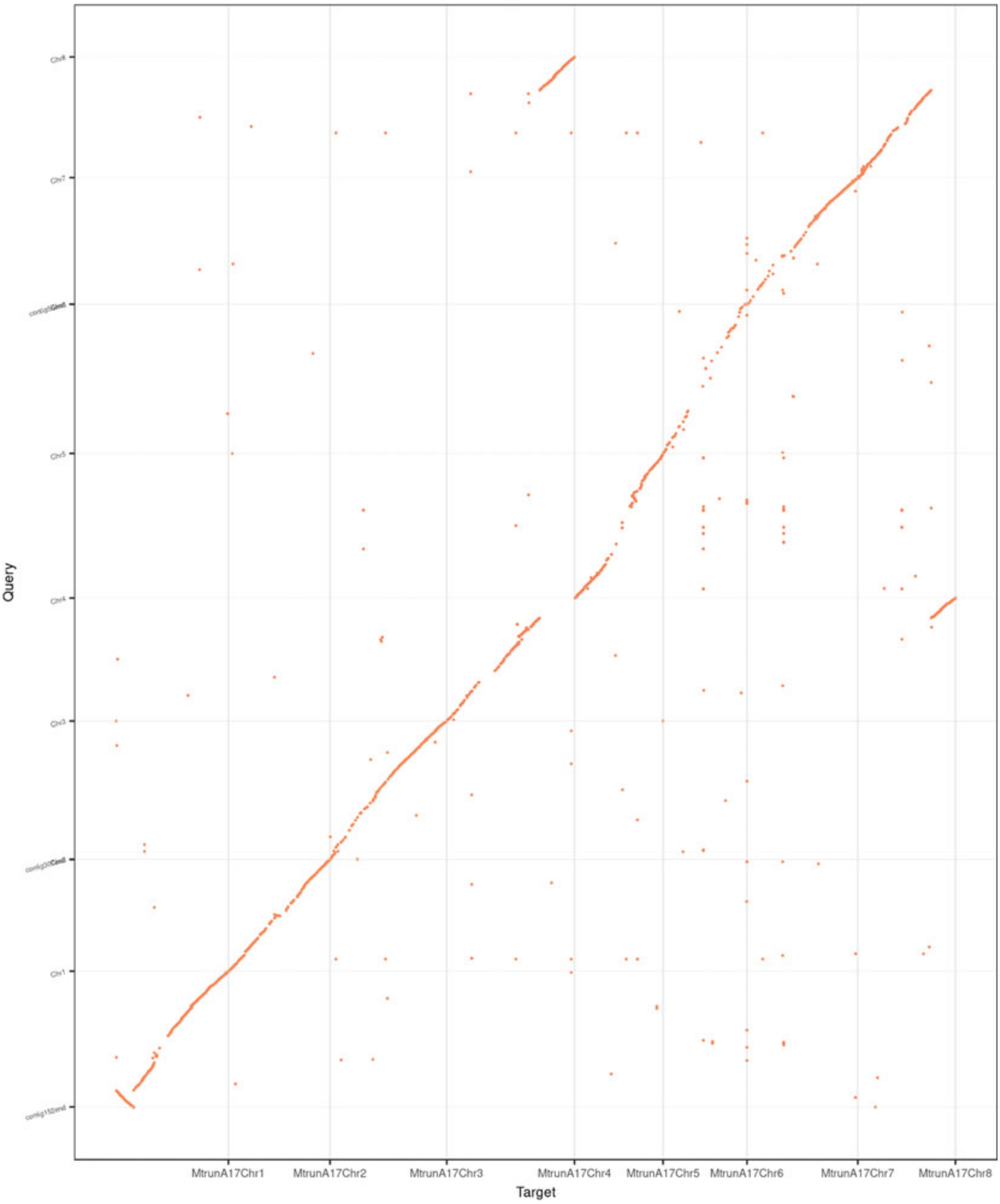

Fig. 6.10 Dotplot comparing the Zhongmu No. 1 assembly (y-axis) to the eight *Medicago truncatula* v. 5.0 chromosomes (x-axis). Dotplots were generated as described in Fig. 6.3

(Table 6.2). The majority of duplicated BUSCO genes were captured three times, though some were captured once or twice, and a small number were captured 4 or more times (Fig. 6.2). While the differential gene capture seen between haplotypes may reflect actual differential gene content as seen in CADL, the authors found that the number of genes is similar between haplotypes, retained synteny is high, and evolutionary pressures on genes are similar between haplotypes.

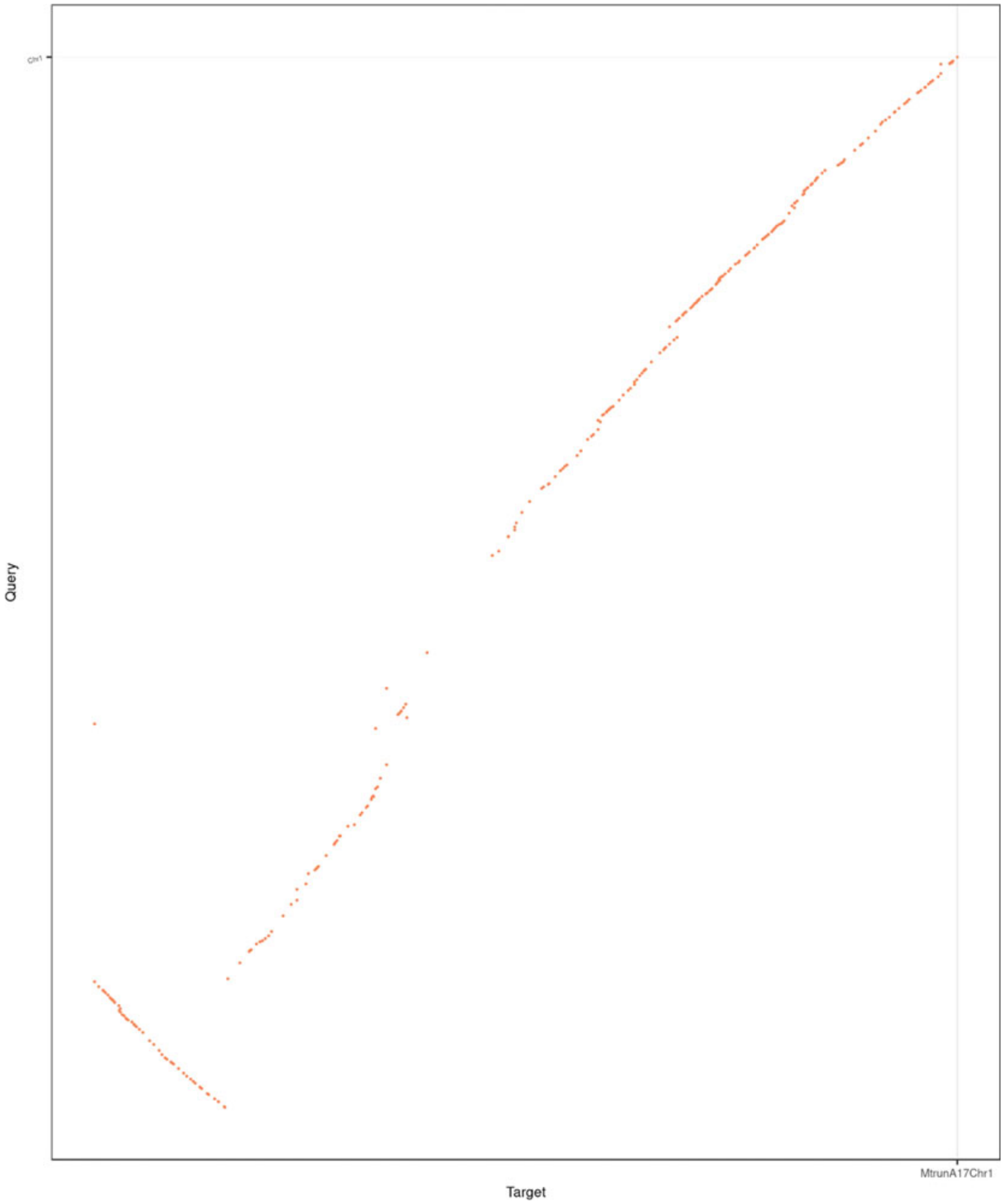

Fig. 6.11 Inversion shown in Zhongmu No. 1 chromosome 1 (y-axis) compared to *Medicago truncatula* chromosome 1 (x-axis). Dotplots were generated as described in Fig. 6.3

Indeed, the percentage of duplicated BUSCO genes in the genome is just over 90%, indicating that most genes are in multiple haplotypes. Furthermore, the percentage of the genome that remains uncaptured is sufficient to explain missing genes in each haplotype. Nevertheless, it is clear that some differential gene loss occurs between the XinJiangDaYe haplotypes (Fig. 6.1b).

At least some of the uncaptured portion of the genome is likely reflected in haplotypes that were collapsed due to strong similarity across chromosome distances that exceed PacBio subread lengths. There was approximately double coverage on 3.2% of the genome, reflecting possible collapse of haplotypes, though collapse of repeats or tandem duplications within a haplotype could also account for some of the double coverage regions. Nevertheless, there is good evidence that, overall, haplotypes were highly similar. Sequence divergence wasn't always sufficient to distinguish haplotypes and the consensus genetic map that was used couldn't distinguish the subgenomes, leading to some possible phasing errors. Nevertheless, it is interesting that, in addition to structural variation identified between XinJiangDaYe and *M. truncatula* (Fig. 6.12), some structural differences between subgenomes were uncovered. For instance, Hi–C data supports the two inversions that occur in only one of the chromosome 1 haplotypes (Fig. 6.13).

6.3 Annotation

6.3.1 Genes

Gene counts vary between the assemblies but there are some interesting patterns (Table 6.2). PI464715, Zhongmu No. 1, and XinJiangDaYe all have between 47 and 50 k genes per 800 Mb haploid genome complement. CADL, on the other hand, has considerably higher at ~71 k genes per haploid genome complement. About ¼ of these are redundant at >98% identity, indicating they might be from alternate haplotypes. Surprisingly, CADL has more protein-coding gene annotations than NECS-141 despite being less than half as long. This not only reflects the high gene count in CADL but also a low gene count in NECS-141 (~30 k genes per haploid genome complement). Differences in annotation pipelines likely account for the differing gene counts. For example, the CADL annotation was the only one that used the SPADA pipeline (Zhou et al. 2013), adding about 8,000 small peptides to the annotation.

BUSCO was originally generated for assessing completeness of genome assembly and annotation (Simão et al. 2015). While it only assays "near-universal single-copy" genes, BUSCO analyses are a reasonable surrogate for overall gene capture. All five genome assemblies had complete gene capture of more than 93% of genes, ranging from 93.3% in Zhongmu No. 1 to 97.7% in PI464715 (Table 6.2). This indicates that, at least in the gene space, all of these assemblies are nearly complete. The capture of duplicate copies of the BUSCO genes mirrors well estimates of duplication based on extra genome length beyond the 800 Mb base genome size and through alignments to *M. truncatula* (Table 6.2, Figs. 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13). The diploid PI464715 and the tetraploid Zhongmu No. 1 have both been compressed into a haploid genome complement. CADL and NECS-141, along with the intermediate PI464715 assembly version, have retained diverged duplicate haplotypes while collapsing highly similar or identical ones, and XinJiangDaYe has separately assembled all four subgenomes.

A customized BUSCO analysis that identifies copy number of captured genes reflects differing levels of haplotype capture in different assemblies (Fig. 6.2). The *Medicago truncatula* (A17) and Zhongmu No. 1 assemblies capture mostly single copy BUSCO genes reflecting the haploid assembly strategy. Based on the small percentage of duplicated genes in the final PI464715 genome (Table 6.2), the final PI464715 genome would have looked similar had it been included in this figure. The diploid assemblies, CADL and the intermediate assembly of PI464715, which both captured some haplotype variation, show similar profiles with the largest fraction of BUSCO genes captured with two copies but with a significant fraction collapsed into a single haplotype. The two uncollapsed tetraploid assemblies, NECS-141 and XinJiangDaYe, show BUSCO gene counts that vary. The biggest fraction has a count of three, with that fraction being larger in XinJiangDaYe where a concerted effort to capture all four haplotypes was

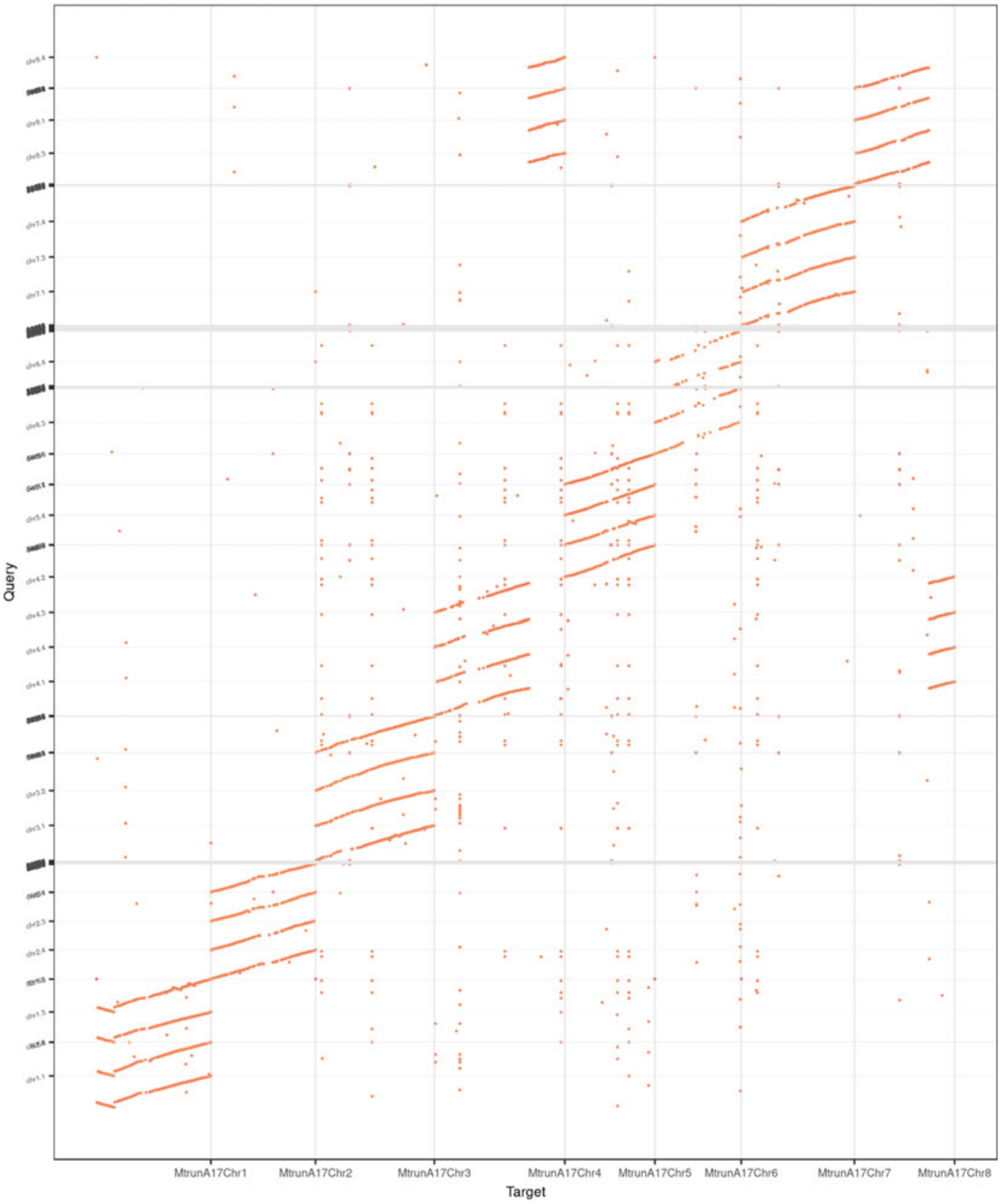

Fig. 6.12 Dotplot comparing the XinJiangDaYe assembly (y-axis) to the eight *Medicago truncatula* v. 5.0 chromosomes (x-axis). Dotplots were generated as described in Fig. 6.3

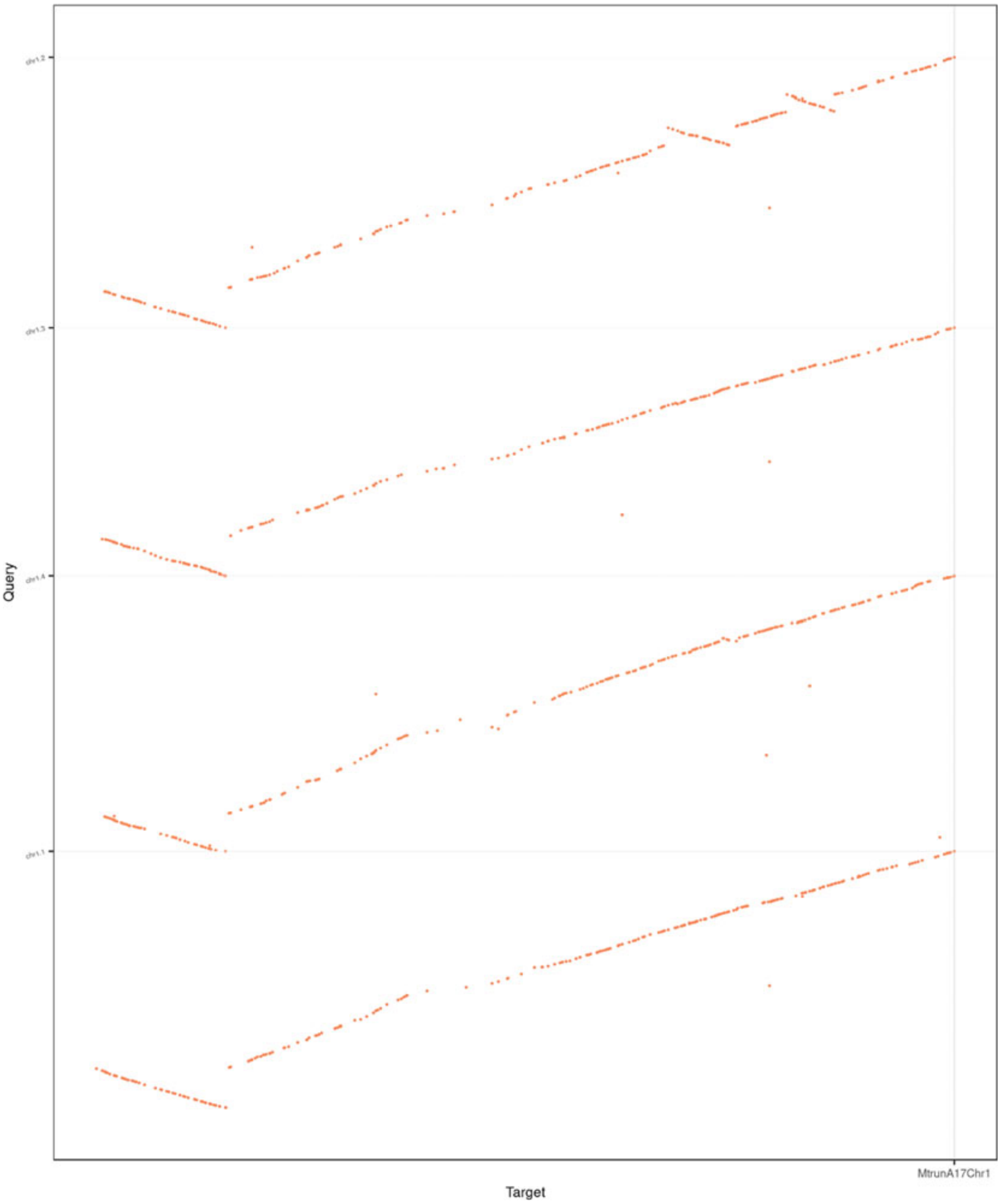

Fig. 6.13 Inversion shown in four XinJiangDaYe chromosome 1 subgenomes (y-axis) compared to *Medicago truncatula* chromosome 1 (x-axis). Dotplots were generated as described in Fig. 6.3

employed. Surprisingly, only a small fraction of BUSCO genes with a count of 4 were captured in the assembly, though more were captured in XinJiangDaYe than in NECS-141.

6.3.2 Repeats

The alfalfa haploid genome size (~800 Mb) is much larger than that of *Medicago truncatula* (~450 Mb). The difference between the two genomes appears to be due mainly to repeat expansion rather than genome duplication. Approximately 55% of the assembled genome consists of transposable elements (TEs), which more than doubles the number of Mb of TEs in the *M. truncatula* genome and provides significant challenges to assembly (Chen et al. 2020; Li et al. 2020).

The long terminal repeat (LTR) class of TEs is the most expanded, nearly quintupling in total length from approximately 65 Mb in *M. truncatula* to 315 Mb in the Zhongmu No. 1 alfalfa genome (Shen et al. 2020) and more than doubling the percentage in the genome from 13.37% in *M. truncatula* to 27.36% in XinJiangDaYe (Chen et al. 2020). This expansion was fueled by LTR bursts that occurred much more heavily in alfalfa than in *M. truncatula* after the two species split (Shen et al. 2020; Chen et al. 2020). Within the LTRs, the Ty3/Gypsy element superfamily is the biggest contributor to the increased alfalfa genome size compared to *M. truncatula*, accounting for nearly a third of the increase (Chen et al. 2020). While repetitive sequence is clearly the major contributor to genome expansion in alfalfa compared to *M. truncatula*, non-repetitive sequence contributes to about one-quarter of the expansion over *M. truncatula* (Chen et al. 2020). Further evidence that large-scale duplications do not appear to have contributed significantly to genome expansion in alfalfa is confirmed by comparisons of the alfalfa genomes to *M. truncatula* (Figs. 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13).

6.3.3 Variation

Alfalfa is an outcrossing species, and so heterozygosity is expected to be high. Genome sequencing data confirms this. In the diploid PI464715, the average heterozygosity rate estimate is ~1.9%, or nearly 2 heterozygous nt per 100 nt (Li et al. 2020). The heterozygosity rate estimate is nearly double (3.7%) in the tetraploid XinjiangDaye, reflecting the increased variation present in the tetraploid genome with four haplotypes rather than two (Chen et al. 2020).

Tetraploid alfalfa is an autotetraploid with tetrasomic inheritance (Stanford 1951), allowing for recombination between haplotypes that keeps them highly similar. Nevertheless, structural differences between haplotypes can be clearly seen in these genome assemblies. These structural differences include differential gene content between haplotypes as shown in the diploid CADL, which was derived from a tetraploid, as well as in the tetraploid XinJiangDaYe (Fig. 6.1). In addition, the XinJiangDaYe assembly, because it has assembled all four subgenomes with high continuity, shows the presence of larger differential structural variation, including inversions that might affect local recombination (Figs. 6.12 and 6.13).

6.4 Conclusion

Within the last five years, five alfalfa assemblies have been generated, allowing alfalfa researchers to work directly within the alfalfa genome rather than relying on *M. truncatula* genomic resources. The five genome assemblies discussed in this chapter utilize a changing spectrum of sequencing and scaffolding technologies that lead to improved assembly continuity and an increased ability to distinguish between repeats and subgenome haplotypes. This ability to distinguish nearly identical sequences is critical in alfalfa genome assembly because transposable repeats alone comprise more than half of the alfalfa genome. Furthermore, haplotypes present in the different subgenomes are highly similar, as evidenced by insufficient sequence divergence to distinguish some haplotypes in the CADL, NECS-141, and XinJiangDaYe assemblies, even with genetic map support, as well as in preliminary versions of the PI464715 assembly. Nevertheless, these assemblies uncover local haplotype differences in gene content as well as larger structural rearrangements that distinguish some of the subgenomes. The increase in repeat content compared to *M. truncatula*, as well as heterozygosity and ploidy challenges, are now more easily navigable with improved, highly accurate PacBio HiFi long reads or even longer, moderately accurate ONT reads, as well as rapid, inexpensive whole-genome scaffolding technologies such as BioNano and chromatin conformation technologies. Given these technological breakthroughs, we fully expect to see additional alfalfa genome assemblies released in upcoming years as alfalfa researchers sequence additional alfalfa germplasm that has important scientific or breeding applications.

6.5 Assembly Availability

CADL is available under a Ft. Lauderdale usage agreement at https://legumeinfo.org/data/index/public/Medicago_sativa/CADL_HM342.gnm1.rVNY/. NECS-141 is available under MTA from the Noble Research Institute in Ardmore, Oklahoma. All other assemblies are available as described in their manuscripts (Shen et al. 2020; Chen et al. 2020; Li et al. 2020).

Acknowledgements We would like to thank the following people for helping with the CADL and/or NECS-141 assembly or annotation: Connor T. Cameron, Xinbin Dai, Nicholas P. Devitt, Diego A. Fajardo, Joseph Guhlin, Jason R. Miller, Maria J. Monteros, Thiru Ramaraj, Kevin A. T. Silverstein, Haibao Tang, Christopher D. Town, Nevin D. Young, Patrick X. Zhao, Peng Zhou, Zhaohong Zhuang. We would like to thank the *Medicago* HapMap project (NSF Project IOS-1237993) and Noble Research Institute for funding for the CADL and NECS-141 assemblies. Acknowledgements for the other genome assemblies can be found in their respective manuscripts.

References

Barker DG, Bianchi S, Blondon F, Dattée Y, Duc G, Essad S, Flament P, Gallusci P, Génier G, Guy P, Muel X, Tourneur J, Dénarié J, Huguet T (1990) *Medicago truncatula*, a model plant for studying the molecular genetics of the Rhizobium-legume symbiosis. Plant Mol Biol Rep 8:40–49

Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R, Deniot G, Denoeud F, Duffé P, Engelen S, Lemainque A, Manzanares-Dauleux M, Martin G, Morice J, Noel B, Vekemans X, D'Hont A, Rousseau-Gueutin M, Barbe V, Cruaud C, Wincker P, Aury J-M (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants 4:879–887

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

Bingham ET (1969) Haploids from cultivated alfalfa, *Medicago sativa L.* Nature 221:865–866

Bingham ET, McCoy TJ (1979) Cultivated Alfalfa at the diploid level: origin, reproductive stability, and yield of seed and forage 1. Crop Sci 19:97–100

Blondon F, Marie D, Brown S, Kondorosi A (1994) Genome size and base composition in *Medicago sativa* and *M. truncatula* species. Genome 37:264–270

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

Chen H, Zeng Y, Yang Y, Huang L, Tang B, Zhang H, Hao F, Liu W, Li Y, Liu Y, Zhang X, Zhang R, Zhang Y, Li Y, Wang K, He H, Wang Z, Fan G, Yang H, Bao A, Shang Z, Chen J, Wang W, Qiu Q (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun 11:2494

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054

Cook DR (1999) *Medicago truncatula*—a model in the making! Curr Opin Plant Biol 2:301–304

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

Deschamps S, Mudge J, Cameron C, Ramaraj T, Anand A, Fengler K, Hayes K, Llaca V, Jones TJ, May G (2016) Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from Agrobacterium tumefaciens. Sci Rep 6:28625

Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun 9:4844

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42:443–461

Du H, Liang C (2019) Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nat Commun 10:5360

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 3:99–101

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Hufton AL, Panopoulou G (2009) Polyploidy and genome restructuring: a variety of outcomes. Curr Opin Genet Dev 19:600–606

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

Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P (2019) Tools and strategies for long-read sequencing and De Novo assembly of plant genomes. Trends Plant Sci 24:700–724

Khu D-M, Reyno R, Charles Brummer E, Bouton JH, Han Y, Monteros MJ (2010) QTL mapping of aluminum tolerance in tetraploid alfalfa. In: Sustainable use of genetic diversity in forage and turf breeding, pp 437–442

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV (2018) Current strategies of polyploid plant genome sequence assembly. Front Plant Sci 9:1660

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L, Wang D, Hu F, Wang W, Liu S (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. Gigascience 9:giaa123. https://doi.org/10.1093/gigascience/giaa123

Li A, Liu A, Du X, Chen J-Y, Yin M, Hu H-Y, Shrestha N, Wu S-D, Wang H-Q, Dou Q-W, Liu Z-P, Liu J-Q, Yang Y-Z, Ren G-P (2020) A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. Hortic Res 7:194

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun 9:541

Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr Opin Plant Biol 54:26–33

Mishra DC, Lal SB, Sharma A, Kumar S, Budhlakoti N, Rai A (2017) Strategies and tools for sequencing and assembly of plant genomes. Compend Plant Genomes 81–93

Moll KM, Zhou P, Ramaraj T, Fajardo D, Devitt NP, Sadowsky MJ, Stupar RM, Tiffin P, Miller JR, Yound ND, Silverstein KAT, Mudge J (2017) Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model *Medicago truncatula*. BMC Genomics 18:578

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok P-Y (2016) A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods 13:587–590

Myers G (2014) Daligner: fast and sensitive detection of all pairwise local alignments

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak ACY, Lai YYY, Lin C, Chu C, Graves T, Kwok P-Y, Wilson RK, Sikela JM (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genom 15:387

Pryszcz LP, Gabaldón T (2016) Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 44:

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

Roach MJ, Schmidt SA, Borneman AR (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinform 19

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

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16

Shen C, Du H, Chen Z, Lu H, Zhu F, Chen H, Meng X, Liu Q, Liu P, Zheng L, Li X, Dong J, Liang C, Wang T (2020) The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. Mol Plant 13:1250–1261

Shi S, Nan L, Smith KF (2017) The current status, problems, and prospects of alfalfa (*Medicago sativa L.*) breeding in China. Agronomy 7:1

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Small E, Jomphe M (1989) A synopsis of the genus Medicago (Leguminosae). Can J Bot 67:3260–3294

Stanford EH (1951) Tetrasomic inheritance in alfalfa 1. Agron J 43:222–225

Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo M, Batley J, Edwards D, Doležel J, Šimková H (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J 14:1523–1531

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, Church DM, Eichler EE, Wilson RK (2014) Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res 24:2066–2076

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, Freeling M, Bartels D, Ten Hallers B, Hastie A, Michael TP, Mockler TC (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature 527:508–511

Veronesi F, Brummer EC, Huyghe C (2010) Alfalfa. In: Boller B, Posselt UK, Veronesi F (eds) Fodder crops and amenity grasses. Springer, New York, New York, NY, pp 395–437

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162

Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, Luo F, Xie Z (2017) MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 14:1072–1074

Yang S, Gao M, Xu C, Gao J, Deshpande S, Lin S, Roe BA, Zhu H (2008) Alfalfa benefits from *Medicago truncatula*: the RCT1 gene from *M. truncatula* confers broad-spectrum resistance to anthracnose in alfalfa. Proc Natl Acad Sci U S A 105:12164–12169

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KAT, Tang H, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Bergès H, Bidwell S, Bisseling T, Choisne N, Couloux A, Denny R, Deshpande S, Dai X, Doyle JJ, Dudez A-M, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, González AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong D-H, Jing Y, Jöcker A, Kenton SM, Kim D-J, Klee K, Lai H, Lang C, Lin S, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun J-H, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin B, Qu C, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi R, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang B-B, Wang K, Wang M, Wang X, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing Y, Yang L, Yao Z, Ying F, Zhai J, Zhou L, Zuber A, Dénarié J, Dixon RA, May GD, Schwartz DC, Rogers J, Quétier F, Town CD, Roe BA (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524

Zhang H, Huang Q-M, Su J (2010) Development of alfalfa (*Medicago sativa L.*) regeneration system and agrobacterium-mediated genetic transformation. Agric Sci China 9:170–178

Zhang X, Zhang S, Zhao Q, Ming R, Tang H (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants 5:833–845

Zhou P, Silverstein KA, Gao L et al (2013) Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). BMC Bioinform 14:335