Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*

Chunting Lang

Thesis committee

Thesis supervisor
Prof. dr. Ton H.J. Bisseling
Professor of Molecular Biology (Development Biology of Plants)
Wageningen University

Thesis co-supervisor
Dr. ir. Rene Geurts
Assistant professor, Laboratory of Molecular Biology
Wageningen University

Other members
Dr. ir. Paul F. Fransz, University of Amsterdam
Dr. ir. Christian W.B. Bachem, Wageningen University
Dr. ir. Guusje A.B. Bonnema, Wageningen University
Prof. dr. Hans J.H. de Jong, Wageningen University

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences

Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*

Chunting Lang

Thesis
submitted in fulfillment of the requirement for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus,
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 31 January 2012
at 4 p.m. in the Aula.

Chunting Lang
Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*,
150 pages

Thesis, Wageningen University, Wageningen, NL (2012)
With references, with summaries in Dutch and English

ISBN 978-94-6173-197-5

Contents

CHAPTER 1

General Introduction

The importance of legumes

The *Leguminosae* family, or *Fabaceae*, which comprises over 650 genera and 18,000 species, is the third largest flowering plant family in the world (Polhill et al., 1981). The family has a worldwide distribution and includes plants of every size from tiny herbs to giant trees. Some legumes are important grain or forage crops. These are grown at around 180 million hectare, up to 15% of the earth's arable surface (Graham and Vance, 2003). Legumes are a major source of protein and lipids for both humans and animals. Humans obtain 33% of the dietary protein from edible legumes (Vance et al., 2000), for instance, soybean (*Glycine max*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), pigeon pea (*Cajanus cajan*) and cowpea (*Vigna unguiculata*). More than 35% of the world's industrial vegetable oil comes from legumes, mainly soybean and peanut (*Arachis hypogeae*). Legumes are also producers of profitable secondary compounds. Isoflavones from soybean, red clover (*Trifolium pratense*) and other legumes have been suggested for medical use to reduce the risks of cancer and lower serum cholesterol (Kennedy, 1995; Molteni et al., 1995). In addition to food and forage uses, legumes are also commonly used as organic fertilizers. This is due to their unique ability to interact with several soil bacteria of the *Rhizobiaceae*, collectively called rhizobia, and results in a nitrogen fixing symbiosis. This symbiosis occurs in specialized organs, the root nodules. In the nodules of most legumes the bacteria are hosted intracellularly and are able to fix atmospheric nitrogen to provide ammonium to the host plant. Because lack of nitrogen is often a rate-limiting condition in growth of non-legume plants, the unique nitrogen-fixing ability of legumes provides one of the most important biological sources of organic nitrogen compounds and thus enriches entire ecosystems. Therefore, the study of symbiosis in legumes is of paramount importance for the ongoing development of sustainable and renewable agriculture and human well-being. Not surprisingly, a great amount of study is dedicated to the elucidation of nitrogen fixation pathways and other aspects of legume biology.

The phylogeny of legumes

The oldest known legume fossils date back to 56 million years ago, based on fruit fossils that are similar to *Diplotropis* and *Bowdichia* species (Herendeen et al., 1992). According to molecular dating methods, the estimated age of the legume crown group is 60 million years, which slightly predates the oldest legume fossils. Legumes are traditionally classified into three subfamilies, *Caesalpinioideae*, *Mimosoideae* and *Papilionoideae* (Polhill et al., 1981), and large clades occur within the three subfamilies (Fig. 1.1). The subfamily *Caesalpinioideae* comprises around 161 genera and 3,000 species (Lewis et al., 2005). The age of the oldest *Caesalpinioid* crown clade is dated to approximately 54 million years ago (Lavin et al., 2005). The *Caesalpinioideae* is most diverse in tropical regions throughout the New World, Africa, and southeast Asia (Lavin et al., 2005), and diverged early after the formation of the subfamilies (Polhill et al., 1981). The subfamily *Mimosoideae*, with an estimated 80 genera and approximately 3,000 species is the youngest subfamily of the legumes. It is monophyletic and can be traced back to a most recent common ancestor with the *Caesalpinioideae* just over 40 million years ago (Polhill et al., 1981; Lavin et al., 2005). The subfamily *Papilionoideae* is the largest and most widely distributed of the three classified subfamilies. It includes approximately 483 genera and 12,000 species (Lewis et al., 2005), and is subdivided into at least four major clades; Genistoid, Dalbergioid, Mirbelioid and Hologalegina. The Papilionoids crown clades diverged at approximately the same time as the Caesalpinioids crown clade. The divergence of the Genistoid clade is estimated to have occurred about 56 million years ago, that of the Dalbergioid about 55 million years ago, Hologalegina about 50 million years ago and the Mirbelioid at about 48 million years ago (Lavin et al., 2005). With more than 4,800 species, Hologalegina is the largest of the four major Papilionoid clades. It contains large numbers of temperate and herbaceous tribes, such as *Phaseoleae*, *Galegeae*, *Carmichaelieae*, *Cicereae*, *Hedysareae*, *Trifolieae*, *Fabeae* (a.k.a. *Vicieae*) plus *Loteae* sens. lat. (*Loteae* and *Coronilleae*) and predominantly New World *Robinieae* (Wojciechowski, 2003). Most domesticated legume crop and forage species belong to the Hologalegina clade, for instance, soybean, pea, common bean (*Phaseolus vulgaris*), faba bean (*Vicia faba*), chickpea, alfalfa (*Medicago sativa*), lentils (*Lens culinaris*), clovers (*Trifolium* spp.). It also includes *Medicago truncatula* and *Lotus japonicus*, both of which are used as legume model species for genetic and genomic studies.

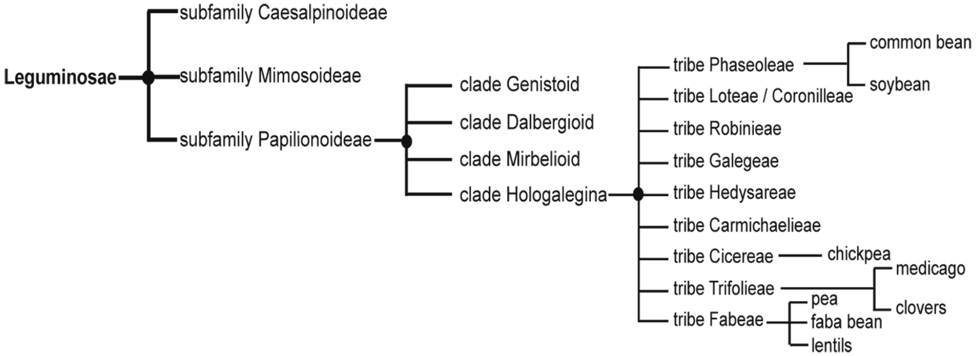


Figure 1.1: Simplified structure of the *Leguminosae*. The family is subdivided in three subfamilies of which Papilionoidae is most prominent.

M. truncatula belongs to the *Trifolieae* tribe that includes the genera *Medicago*, *Trifolium*, *Melilotus*, *Trigonella*, *Parochetus* and *Ononis,* among which *Trifolium* is by far the largest. More than half of the species in the *Trifolieae* tribe belongs to the *Trifolium* genus. The genera *Trifolium* and *Medicago* have diverged 24.7 +/- 2.3 million years, which is relatively recent. In comparison, the divergence with the *Loteae* tribe occurred 50.6 +/- 0.9 million years ago. Because *M. truncatula* is phylogenetically closer to the *Trifolium* genus than *L. japonicus* it is more suitable to use as reference for genome comparison and evolutionary studies within this genus.

Trifolium is one of the largest genera in the legume family with ~255 species (Gillett and Taylor, 2001). All species are herbaceous perennials or annuals, and are widely grown as forage crops as well as organic fertilizers. At least 16 species of the genus are actively cultivated for agricultural use. For instance, white clover (*T. repens*) and red clover (*T. pratense*) are widely grown as highly palatable, nutritious forage for all classes of livestock in the world. Red clover flower extracts are also used in cosmetics. Arrow leaf clover (*T. vesiculosum*) is grown in south-east USA as winter annual for forage because of its cold tolerance. Crimson clover (*T. incarnatum*) is planted as weed control in corn, soybean and wheat fields in Michigan, because of its tall seedlings and shade tolerance. Rose clover (*T. hirtum*) is used for dry-land pasture to increases the protein content of harvested forage in California.

The *Trifolium* genus is subdivided into two subgenera, named *Trifolium* and *Chronosemium* (Ellison et al., 2006). Basic chromosome numbers vary enormously over the two subgenera, from 5 to 24 in the 184 species investigated. Of the examined species, 80% are 2n = 16, inferred to be the ancestral chromosome number in the *Trifolium* genus. Polyploidy occurs frequently, with instances of tetraploidy, hexaploidy, and even dodecaploidy. The subgenus *Chronosemium* is divided into two clades, characterized by different chromosome numbers (2n = 14, 16 or polyploid 2n = 30). Diverging chromosome numbers are also observed in the subgenus *Trifolium*, which is separated into five main sections, *Trichocephalum*, *Trifolium*, *Vesicastrum*, *Trifoliastrum* and *Involucrarium* and several smaller sections (Ellison et al., 2006). Of these five major sections, chromosome number variations are mostly observed in *Trichocephalum* and *Trifolium*. The common ancestor of both sections is 2n = 16, but species with 2n = 10, 12, 14 and 48 also occur (Ellison et al., 2006). In contrast, chromosome number variations are rare in perennial species and absent in the African and American *Trifolium* species (Taylor and Quesenberry, 1996), which occur in sections *Vesicastrum*, *Trifoliastrum* and *Involucrarium* and share a perennial ancestor. Differences in basic chromosome number occur mainly in annual species and species with deviant chromosome numbers rarely give rise to polyploids (Ellison et al., 2006). Within the *Trifolium* section, red clover is the only species with 2n = 14, while all other closely related species have 2n = 16 (Taylor and Quesenberry, 1996; Ellison et al., 2006). Therefore, it is interesting to understand the genome organization and the evolution of chromosome rearrangements that has occurred in red clover.

Organization of legume genomes

Plant nuclear genomes can vary enormously in genome size, chromosome number, numbers of gene clusters and of repetitive sequences. Chromosome numbers are not strongly correlated with genome size as chromosomes can differ enormously in length. For instance, rice (*Oryza sativa*) has a genome size of ~430 Mb and a haploid chromosome number of 12 (The Rice genome, 2005), whereas barley (*Hordeum vulgare*) has a genome size of ~5,700 Mb and a haploid chromosome number of only 7 (Bennett and Smith, 1976). A single chromosome of barley contains more DNA than one complete haploid rice genome (Dubcovsky et al., 2001). Most plants have chromosomes containing a single centromere, necessary for chromosome segregation. One of the major exceptions is *Luzula* members of the *Juncaceae* family which have holokinetic chromosomes with microtubule attachments along the whole chromosome (Nagaki et al., 2005). In the case of regular monocentric chromosomes we find centromeres surrounded by large regions of highly condensed heterochromatin enriched for tandem repeats and transposable elements, with highly methylated DNA and suppression of crossover recombination. In contrast, euchromatin is less condensed than heterochromatin and is rich in genes, and there most of the meiotic recombination occurs. Euchromatin may also contain some smaller heterochromatic regions, called knobs or chromomeres. The telomere ends of plant chromosomes are necessary for stabilizing chromosomes and protecting the chromosome ends against fusion and degradation. The telomere repeat consensus sequence $(TTTAGGG)_n$ was first identified in *A. thaliana* (Richards and Ausubel, 1988) and has been found in most flowering plants with some exceptions (McKnight and Shippen, 2004). In most such cases the consensus sequence is replaced by a different highly repetitive satellite sequence; e.g. the human-type telomere sequence $(TTAGGG)_n$. In some *Allium* species no recognizable minisatellite has been identified (Sykorova et al., 2006), but instead chromosome ends are rich in transposable elements (Pich and Schubert, 1998). The copy number of the telomere repeat unit is species specific, for instance 2 to 5 kb in *A. thaliana* (Richards and Ausubel, 1988) and up to 60 to 160 kb in tobacco (*Nicotiana tabacum*) (Fajkus et al., 1995), and even can be different among the two ends of a single chromosome, as is shown in this thesis for red clover (chapter 4). Although in all plants these chromosome structures are present to some extent, major differences occur in the amount of heterochromatin and in its distribution within genomes. Smaller genomes often have smaller heterochromatic regions than larger genomes. Size, distribution and composition of euchromatin can also be highly variable among plant species and even within a genome. For instance, some euchromatic regions in the wheat genome have a significantly higher gene density than other euchromatic regions (Gill et al., 1996a; Gill et al., 1996b).

The genomes within the legume family vary greatly in size, from 370 Mb in hyacinth bean (*Lablab niger*) to the enormous genome of broad bean (*Vicia faba*) with more than 13,000 Mb (Young et al., 2003); (http://www.rbgkew.org.uk/cval/homepage.html). More than 50% of the known legume genome sizes are smaller than 1,300 Mb, according to the plant DNA C-values database (Young et al., 2003). Most of the cultivated species familiar to researchers are moderate in genome size, for instance, cowpea, common bean, chickpea and red clover all have diploid genomes smaller than 1,000 Mb. The two species that where picked as models for fundamental research, *M. truncatula* and *L. japonicus*, have even smaller genomes; in both cases approximately

470-500 Mb. Variation in genome size is caused by extensive differences in the number of repetitive sequences. Some of the difference in genome size is caused by variations in gene number, especially due to polyploidy or segmental duplications (Tikhonov et al., 1999; Blanc et al., 2000; Grant et al., 2000; Ku et al., 2000; Vision et al., 2000; Wendel, 2000; Bancroft, 2001; Bennetzen, 2002). However, most of the variation in genome size appears to be caused by differences in the number of repetitive sequences, especially RNA-mediated transposable elements. Transposable elements are classified into two classes based on their transposition intermediate. Class I or retrotransposons are intermediated by RNA and consist of two major groups: long terminal repeats retrotransposons (LTR-retrotransposons) and non-long terminal repeat retrotransposons (Non-LTR retrotransposon) (Finnegan, 1989). Class II or DNA transposons contain DNA-mediated transposable elements (Finnegan, 1989). In the case of soybean, 57% of the sequenced genome (genome size of ~1,115 Mb) is repeat-rich and predominantly consists of LTR retrotransposon elements (42% of the sequenced genome portion) (Schmutz et al., 2010). This is still less than Sorghum (~55%) and maize (~79%) (Paterson et al., 2009; Schnable et al., 2009), but significantly higher than found in rice (26%) and *A. thaliana* (10%) (The Arabidopsis Genome, 2000; Yu et al., 2002).

Apart from local duplications generated through the action of transposable elements, studies in *A. thaliana* revealed that whole-genome and large-scale segmental duplications occur in the evolution of higher plants (Vision et al., 2000). Duplicated segments span up to 60% of sequenced *A. thaliana* genome (Blanc et al., 2000; The Arabidopsis Genome, 2000). Large-scale genome duplication has also been demonstrated in maize, where the total duplicated proportion of the genome is up to 70%. Genome duplication in legumes has mainly been studied in soybean. Due to a paleopolyploid evolutionary past, soybean has a complex genome. Based on comparative genetic mapping through molecular markers and sequenced data, the soybean genome appears to consist of many duplicated chromosomal segments (homoeologous segments). Some of those can be clearly observed in two or more chromosomes, up to as many as six times (Schmutz et al., 2010), which suggests that multiple rounds of duplication occurred in soybean. From a functional point of view, genome duplications can accelerate evolution by providing draft copies for mutation and selection to work on without disturbing original functions of genes, creating new functions or expression patterns (Lynch and Conery, 2000). Multiple rounds of duplication may occur to further duplicate genes into gene families (Sankoff, 2001).

Genome sequencing

Genome sequencing provides vital resources for studies of development, function and evolution of plant species. In legumes, information derived from sequencing is essential to understand the genetic networks underlying the nitrogen-fixing symbiosis with rhizobia. For *M. truncatula* and *L. japonicus* large-scale genome sequencing has been conducted. Both are diploid and have relatively small genome sizes of ~500 Mb. Both sequencing projects aim to sequence the euchromatic arms, which contain most of the genes (and is known as the gene-space).

Several strategies can be applied to sequence a genome. The oldest method, employed from the late 1980s onwards to sequence viruses, prokaryotes, and later small eukaryote genomes and eventually even the human genome, is slow but sure. In this method, the genome is cloned into a bacterial artificial chromosome (BAC) library. Each BAC is then sub-cloned into small vectors, which are sequenced completely using Sanger sequencing (or nowadays 454 pyrosequencing). Later, BACs are assembled into contigs based on sequence overlap and on the order in genetic maps of molecular markers within the BACs. The resulting sequence is relatively error-free and contains a limited number of gaps, but the costs and labor involved in this method are prohibitive for any but the largest projects. A newer method introduced in the late 1990s is shotgun sequencing, in which the genome is directly fragmented into millions of tiny pieces, which are sequenced using parallel sequencing methods such as 454 pyrosequencing or Illumina sequencing. These pieces are then assembled into a sequence using sophisticated alignment algorithms. Although this method is much faster than BAC-by-BAC sequencing, the resulting sequence assembly, consisting of scaffolds (contigs pieced together by bridging paired-end sequences), superscaffolds (scaffolds pieced together based on genetic methods) and pseudomolecules (a single sequence per chromosome with gaps filled in with an approximately correct number of unknown basepairs denoted by “N”) contains many more gaps and can often not be assembled into fragments larger than several kb due to the methods inability to span repeat regions. To overcome this problem, nowadays a series of libraries each containing millions or billions of clones are created. Clones in each library contain fragments of a certain insert size, and the set of libraries is chosen such that it covers a range of fragment lengths, e.g. 50 bp, 100 bp, 500 bp, 2 kb, 10 kb, 50 kb. Then, all clones in each of these libraries are end-sequenced using Illumina sequencing, which thus results in pairs of short sequences of approximately 30-90 bp connected by an unknown sequence of which the approximate length is known. Sequences are first assembled into contigs and then into scaffolds, which consist of contigs linked together by the fragments of approximately known length inside each clone. Thus, scaffolds can span across several Mb of sequence.

Each of these methods results in many more fragments than the number of chromosomes known to exist. No sequencing method can bridge kb-long stretches of repeat sequence, and each method will result in many alignment errors due to homologous sequences. Genetic maps can be used to assemble fragments into linkage groups corresponding to chromosomes, but even then errors, gaps and spurious fragments will remain. Ultimately, the only method that can reliably bridge huge stretches of repetitive sequences is *in situ* hybridization on chromosomes, such as we describe in chapter 3 for *M. truncatula*.

Importance of fluorescent *in situ* hybridization (FISH) for mapping and genome sequencing

The *in situ* hybridization (ISH) technology was initially developed to detect and localize specific DNA or RNA sequences using isotope labeling (Pardue and Gall, 1975). This technology could be used to detect the localization of fragments as small as 1 kb on metaphase chromosomes in the early 1980s (Harper et al., 1981; Rabin et al., 1984). However, radioactive labeling requires long exposure times of sometimes several weeks, probe detection can be difficult due to insufficient signal, and the spatial resolution of radio-sensitive materials is low compared to the size of chromosomes. These drawbacks have been overcome by using fluorescent probes, resulting in fluorescent *in situ* hybridization (FISH). Usually, biotin or digoxigenin molecules are incorporated into probes using nick translation or PCR based methods (Langer-Safer et al., 1982; Landegent et al., 1985). While these were first detected using chemical methods, later fluorescent labeled antibodies were bound to the molecules and directly observed by microscopy. In the past 30 years, FISH techniques have been continually improved. The number of probes which can be detected simultaneously have increased, through clever combination of fluorophores with non-overlapping spectra (Liehr et al., 2004; Tang et al., 2008b). Also, the intensity of the labeled probes has been significantly improved.

The FISH technique has long been applied in plant genome mapping (Jiang and Gill, 1994). It has become a useful tool to identify individual chromosomes in karyotypes and chromosome aberrations (Ferguson-Smith, 1997). FISH mapping of single copy or low copy DNA sequences, such as large-insert (~100 kb) genomic DNA clones can be used to facilitate chromosome identification, especially where morphological characteristics such as chromosome lengths, centromere positions, euchromatin-heterochromatin pattern are not sufficiently different between chromosomes to identify them. BAC clones containing sequences with genetically mapped molecular markers, can be used as probes, and thus chromosomes within a karyogram can be directly assigned to linkage groups within the genetic map, such as for instance has been performed in rice (Cheng et al., 2001). For instance, in *A. thaliana* a detailed karyotype has been described based on chromosomes fixed in the pachytene stage of meiosis. In this karyotype DNA sequences in both euchromatic and heterochromatic regions can be mapped using FISH, and chromosomes can be distinguished using ribosomal DNA as FISH probes. The FISH technique has also greatly contributed to understand the organization of the euchromatic and heterochromatic regions in each individual chromosome (Fransz et al., 2000). FISH Probes can be designed to target repetitive sequences which generate recognizable signals on individual chromosomes of the species. Such repetitive sequences are abundant in many plants, for instance, about 77% of the tomato genome consists of heterochromatin that is rich in repetitive sequences (Peterson et al., 1996). The use of FISH has shown that the repetitive sequence THG2 described in tomato is present in the heterochromatin of all pericentromeres (Zhong et al., 1996). Furthermore, using the biochemically isolated repetitive DNA fraction Cot-100 as FISH probes, it is possible to identify all the heterochromatin regions in tomato (Chang et al., 2008). In tomato, BAC clones that contain single copy sequences are used to identify the chromosome 1, 2, 4, 6, 7, 9 and 12 (Chang et al., 2007; Koo et al., 2008; Szinay et al., 2010). A set of 60 BAC clones with known genetic positions are used to construct the pachytene-based cytogenetic map of potato (Tang et al., 2009). This strategy has also been successfully used to identify individual chromosomes in *M. truncatula* and *L. japonicus* (Kulikova et al., 2001; Pedrosa et al., 2002).

FISH can also be used to guide genome sequencing projects. On a pachytene FISH-based physical map, BAC clones can be efficiently anchored to euchromatic or heterochromatic regions of the chromosomes (Peters et al., 2009). Adjacent euchromatic BACs can then be sequenced on a BAC-by-BAC basis, resulting in a high quality sequence. The order of adjacent, but non-overlapping BAC clones can also be determined by anchoring BAC clones to chromosomes using FISH. This provides independent verification of the sometimes error-prone process of aligning BACs using genetic maps. Furthermore, completeness of the genome sequence can be estimated through measuring the gap size between BAC clones using FISH analysis. The information on gap sizes is important for ongoing genome sequencing to eventually close gaps, as is shown for tomato (Szinay et al., 2010) as well as for *Medicago* (chapter 3).

Chromosome painting, in which FISH is performed with pools of chromosome-specific BAC probes, can be used for genome comparisons of related species. Homeologous chromosomes can thus be identified as well as chromosomal rearrangements between related species and similarities and differences in gene content and gene order. If two closely related species contain the same genetic markers, these can be used to link maps together. Similarly, sequence homology can be used to align genome sequences across species. However, both comparisons based on sequence homology and on shared genetic markers are sensitive to small rearrangements and to multiple copies of small fragments, generating many spurious alignments to fragments on other chromosomes. These can complicate the alignment. Probes derived from one species can be hybridized to metaphase or pachytene spreads of other species, thus visualizing where similar sequences occur. The main advantage of using comparative FISH is that probes are large and therefore insensitive to small misalignments. Also, no genetic mapping population is required, which is an important advantage when studying little-known species. Chromosome painting is widely used in mammals (Cremer et al., 1988). For plants it has been successfully applied in *A. thaliana* to paint entire autosomes to study chromosomal rearrangements, homologue association, and interphase chromosome territories (Lysak et al., 2001). It has also been successfully used to reveal chromosomal rearrangements between *A. thaliana* and *Arabidopsis lyrata* (Berr et al., 2006; Lysak et al., 2006) and to reconstruct karyotype evolution in *Brassicaceae* (Mandakova and Lysak, 2008). In this thesis, we describe chromosomal rearrangements in the *Trifolium* genus by using this approach (chapter 5).

Scope of this thesis

In chapter 2, we describe the genome sequence of the legume model species *M. truncatula*. *M. truncatula* is an excellent model for the study of legume-specific biology, especially endosymbiotic interactions with bacteria and fungi. Here we describe the sequence of the euchromatic regions of the *M. truncatula* genome based on a recently completed BAC-by-BAC based assembly sequence that is supplemented by Illumina-shotgun sequencing, together capturing ~94% of all genes. This genome sequence enables valuable insights into legume specific traits. A whole-genome duplication (WGD) approximately 58 million years ago contributed significantly to the genome we see today, including the process of nodulation and symbiotic nitrogen fixation. The *M. truncatula* genome exhibits much higher levels of genome rearrangements in addition to the WGD event than soybean or *L. japonicus*. Our work provides evidence that this recent WGD gave rise to key components in the perception of rhizobial signals and the formation of nitrogen fixing nodules, traits instrumental to the success of legumes and one of the major drivers for their importance in nature and agricultural systems. The major contribution of the work presented in this thesis to the *M. truncatula* genome project is in the mapping of BAC sequences using FISH (Fig. 2.1).

In chapter 3, we describe in more detail the use of FISH guide the *M. truncatula* genome sequencing. The BAC-by-BAC sequencing strategy used for *M. truncatula* results a very high quality genome sequence, however, in practice gaps still remain. We use FISH experiments to determine the completeness of the genome sequence coverage and perform gap-size estimates based on chromosome measurements.

In chapter 4, we construct a molecular cytogenetic map of legume red clover (2n = 2x = 14) based on a DAPI stained pachytene karyogram. We are able to distinguish all 7 seven chromosomes in a high-resolution pachytene karyotype, and identify centromeres, pericentromeric heterochromatin, rDNA loci and telomeric tandem repeat sequences. We observe a consistent divergence in length of the telomeric repeat unit between both chromosome ends in all seven chromosomes. Using FISH, we cytogenetically mapped BAC clones containing red clover sequences with known genetic positions on pachytene chromosomes. Thus we integrate the genetic, cytogenetic and physical map of red clover.

In chapter 5, we use comparative chromosome painting to study the chromosomal rearrangements that occurred in the *Trifolium* genus, focusing on the evolution of red clover and other related *Trifolium* species through a comparison with a model legume species *M. truncatula*. We use pooled BACs as probes to perform comparative chromosome painting in several species of the *Trifolium* genus. We identify chromosome inversions, translocations, and breaks that occurred in some *Trifolium* species using *M. truncatula* sequences as reference. Some of these chromosomal rearrangements occurred in many species of the *Trifolium* and *Trichocephalum* sections, but specific rearrangements for red clover have also been observed. Our results furthermore show that an inter-chromosomal break in red clover is associated with a red clover specific repetitive sequence, a new type of Non-LTR retrotransposable element similar to LINEs.

CHAPTER 2

The Medicago genome provides insight into the evolution of rhizobial symbioses

Published in Nature 480, 520-524. Supplementary data available online at http://www.nature.com/nature/journal/v480/n7378/ full/ nature10625.html#/supplementary-information

Nevin D. Young[1]*, Frédéric Debellé[2,3]*, Giles Oldroyd[4]*, Rene Geurts[5], Steven B. Cannon[6, 7], Michael K. Udvardi[8], Vagner A. Benedito[9], Klaus F. X. Mayer1[0], Jérôme Gouzy[2,3], Heiko Schoof[11], Yves Van de Peer[12], Sebastian Proost[12], Douglas R. Cook[13], Blake C. Meyers[14], Manuel Spannagl[10], Foo Cheung[15], Stephane De Mita[5], Vivek Krishnakumar[15], Heidrun Gundlach[10], Shiguo Zhou[16], Joann Mudge[17], Arvind K. Bharti[17], Jeremy D. Murray[4,8], Marina A. Naoumkina[8], Benjamin Rosen[13], Kevin A. T. Silverstein[18], Haibao Tang[15], Stephane Rombauts[12], Patrick X. Zhao[8], Peng Zhou[1], Valérie Barbe[19], Philippe Bardou[2,3], Michael Bechner[16], Arnaud Bellec[20], Anne Berger[19], Hélène Bergès[20], Shelby Bidwell[15], Ton Bisseling[5, 21], Nathalie Choisne[19], Arnaud Couloux[19], Roxanne Denny[1], Shweta Deshpande[22], Jeff Doyle[23], Anne-Marie Dudez[2,3], Andrew D. Farmer[17], Stéphanie Fouteau[19], Carolien Franken[5], Chrystel Gibelin[2,3], John Gish[13], Steven Goldstein[16], Alvaro J. González[24], Pamela J. Green[14], Asis Hallab[25], Marijke Hartog[5], Axin Hua[22], Sean Humphray[26], Dong-Hoon Jeong[14], Yi Jing[22], Anika Jöcker[25], Steve M. Kenton[22], Dong-Jin Kim[13, 27], Kathrin Klee[25], Hongshing Lai[22], **Chunting Lang[5]**, Shaoping Lin[22], Simone L Macmil[22], Ghislaine Magdelenat[19], Lucy Matthews[26], Jamison McCorrison[15], Erin L. Monaghan[15], Jeong-Hwan Mun[13, 28], Fares Z. Najar[22], Christine Nicholson[26], Céline Noirot[29], Majesta O'Bleness[22], Charles R. Paule[1], Julie Poulain[19], Florent Prion[2,3], Baifang Qin[22], Chunmei Qu[22], Ernest F. Retzel1[7], Claire Riddle[26], Erika Sallet[2,3], Sylvie Samain[19], Nicolas Samson[2,3], Iryna Sanders[22], Olivier Saurat[2,3], Claude Scarpelli[19], Thomas Schiex[29], Béatrice Segurens[19], Andrew J. Severin[7], D. Janine Sherrier[14], Ruihua Shi[22], Sarah Sims[26], Susan R. Singer[30], Senjuti Sinharoy[8], Lieven Sterck[12], Agnès Viollet[19], Bing-Bing Wang[1], Keqin Wang[22], Xiaohong Wang[1], Jens Warfsmann[25], Jean Weissenbach[19], Doug D. White[22], Jim D. White[22], Graham B. Wiley[22], Patrick Wincker[19], Yanbo Xing[22], Limei Yang[22], Ziyun Yao[22], Fu Ying[22], Jixian Zhai[14], Liping Zhou[22], Antoine Zuber[2,3], Jean Dénarié[2,3], Richard A. Dixon[8], Gregory D. May[17], David C. Schwartz[16], Jane Rogers[26], Francis Quétier[19], Christopher D. Town[15], Bruce A. Roe[22]

1. Departments of Plant Pathology and Plant Biology, University of Minnesota, St. Paul, MN 55108, USA
2. INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, BP 52627, F-31326 Castanet-Tolosan CEDEX, France
3. CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, BP 52627, F-31326 Castanet-Tolosan CEDEX, France
4. Department of Disease and Stress Biology, John Innes Centre, Norwich NR4 7UH, UK
5. Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands
6. USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA, 50011, USA

7. Department of Agronomy, Iowa State University, Ames, IA 50011, USA
8. Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA
9. Department of Genetics and Developmental Biology, Plant and Soil Science Division, West Virginia University, Morgantown, WV 26506, USA
10. MIPS/Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstr.1, Neuherberg, Germany
11. University of Bonn, INRES Crop Bioinformatics, Katzenburgweg 2, 53115 Bonn, Germany
12. Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
13. Department of Plant Pathology, University of California, Davis, Davis, CA 95616, USA
14. Department of Plant & Soil Sciences and Delaware Biotechnology Institute, Universityof Delaware, Newark, DE 19711, USA
15. J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
16. Laboratory for Molecular and Computational Genomics, University of Wisconsin-Madison, Wisconsin 53706 USA
17. National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505 USA
18. Masonic Cancer Center, Biostatistics and Bioinformatics Group, University of Minnesota, Minneapolis, MN 55455 USA
19. Genoscope/Centre National de Séquençage, 2, rue Gaston Crémieux, CP 5706, 91057 Evry Cedex, France
20. INRA, Centre National de Ressources Génomiques Végétales (CNRGV), BP 52627, F-31326 Castanet-Tolosan CEDEX, France
21. College of Science, King Saud University, Post Office Box 2455, Riyadh 11451, Saudi Arabia
22. Advanced Center for Genome Technology, Department of Chemistry and Biochemistry, Stephenson Research and Technology Center, University of Oklahoma, Norman, OK 73019, USA
23. Department of Plant Biology, Cornell University, Ithaca, NY, 14853 USA
24. Department of Computer & Information Sciences, and Delaware Biotechnology Institute, University of Delaware, Newark, DE, 19711, USA

25. Max Planck Institute for Plant Breeding Research, Plant Computational Biology, Carl von Linné Weg 10, 50829 Köln, Germany
26. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
27. International Institute for Tropical Agriculture, (c/o P.O. Box 30709 Nairobi, Kenya 00100), Ibadan, Nigeria
28. National Institute of Agricultural Biotechnology, Rural Development Administration, 225 Seodun-dong, Gwonseon-gu, Suwon 441-707, South Korea
29. INRA, Unité de Biométrie et d'Intelligence Artificielle (UBIA), UR875, BP 52627, F-31326 Castanet-Tolosan CEDEX, France
30. Department of Biology, Carleton College, Northfield, MN, 55057 USA

Legumes (*Fabaceae* or *Leguminosae*) are unique among cultivated plants for their ability to carry out endosymbiotic nitrogen fixation with rhizobial bacteria, a process that takes place in a specialized structure known as the nodule. Legumes belong to one of the two main groups of eurosids, the Fabidae, which includes most species capable of endosymbiotic nitrogen fixation (Wang et al., 2009). Legumes comprise several evolutionary lineages derived from a common ancestor 60 million years ago (Mya). Papilionoids are the largest clade, dating nearly to the origin of legumes and containing most cultivated species (Lavin et al., 2005). *Medicago truncatula* is a long-established model for the study of legume biology. Here we describe the draft sequence of the *M. truncatula* euchromatin based on a recently completed BAC assembly supplemented with Illumina-shotgun sequence, together capturing ~94% of all *M. truncatula* genes. A whole genome duplication (WGD) approximately 58 Mya played a major role in shaping the *M. truncatula* genome and thereby contributed to the evolution of endosymbiotic nitrogen fixation. Subsequent to the WGD, the *M. truncatula* genome experienced higher levels of rearrangement than two other sequenced legumes, *Glycine max* and *Lotus japonicus*. *M. truncatula* is a close relative of alfalfa (*Medicago sativa*), a widely cultivated crop with limited genomics tools and complex autotetraploid genetics. As such, the *M. truncatula* genome sequence provides significant opportunities to expand alfalfa's genomic toolbox.

Optical mapping indicates that the eight pseudomolecules of assembly Mt3.5 span a physical distance of 375 million base pairs (Mb), and fluorescence *in situ* hybridization indicates they extend from pericentromeres almost to telomeric ends (Fig. 2.1). Altogether, Mt3.5 consists of 2,536 bacterial artificial chromosomes (BACs; Supplementary Tables 1 and 2) with 273 physical gaps (including centromeres, Supplementary Table 3) and 101 internal sequencing gaps. The pseudomolecules contain 246 Mb of non redundant sequence (Supplementary Table 2) located entirely within the optical map (Supplementary Fig. 3). Another 146 unfinished BACs/BAC pools that cannot be placed on the optical map contribute 17.3 Mb. Regions not represented in pseudomolecules or unanchored BACs were captured through assembly of approximately 40× coverage Illumina sequencing, yielding 104.2 Mb of additional unique sequence. Although not directly tested, the Illumina sequence is expected to lie predominantly within the boundaries of pseudomolecules (see below). On the basis of expressed sequence tag alignments, the combined data sets capture ~94% of expressed genes, providing a highly informative platform for analysing the euchromatin of *M. truncatula*, although still at the draft stage.

Altogether there are 62,388 gene loci in Mt3.5 (Supplementary Table 4 and Supplementary Fig. 4), with 14,322 gene predictions annotated as transposons. Pseudomolecules and unassigned BACs contain a total of 44,124 gene loci, 177,271 retroelement-related regions and 26,487 DNA transposons, and non-redundant Illumina assemblies contribute an additional 18,264 genes, 75,777 retrotransposon regions and 8,476 DNA transposons (Supplementary Tables 5–9) along with 1,418 organellar insertions (Supplementary Data 1). For pseudomolecules and unassigned BACs, this translates to 16.8 genes, 67.6 retrotransposons and 10.1 DNA transposons per 100 kilobases (kb). Within Illumina sequence assemblies, gene density (17.1 per 100 kb) and retrotransposon density (72.2 per 100 kb) are similar to pseudomolecules and unassigned BACs, whereas DNA transposon density is lower (8.2 per 100 kb). Similarities in gene and transposon densities between BAC and Illumina sequences support the assertion that the Illumina sequence is euchromatic, although the possibility that some Illumina assemblies come from low-copy regions within heterochromatin cannot be excluded. Considering only the 47,845 genes with experimental or database support (Supplementary Table 4), the average *M. truncatula* gene is 2,211 bp in length, contains 4.0 exons, and has a coding sequence of 1,001 bp. These values are similar to those observed previously in *Arabidopsis thaliana* (2,174 bp), *Oryza sativa* (3,403 bp) and *Populus trichocarpa* (2,301 bp) (The Arabidopsis Genome, 2000; The Rice genome, 2005; Tuskan et al., 2006).

Recent analyses of plant genomes indicate a shared whole-genome hexaploidy (WGH) preceding the rosid–asterid split at 140–150 Myr ago (Tang et al., 2008a). Duplication patterns and genomic comparisons strongly suggest an additional WGD approximately 58 Myr ago in the

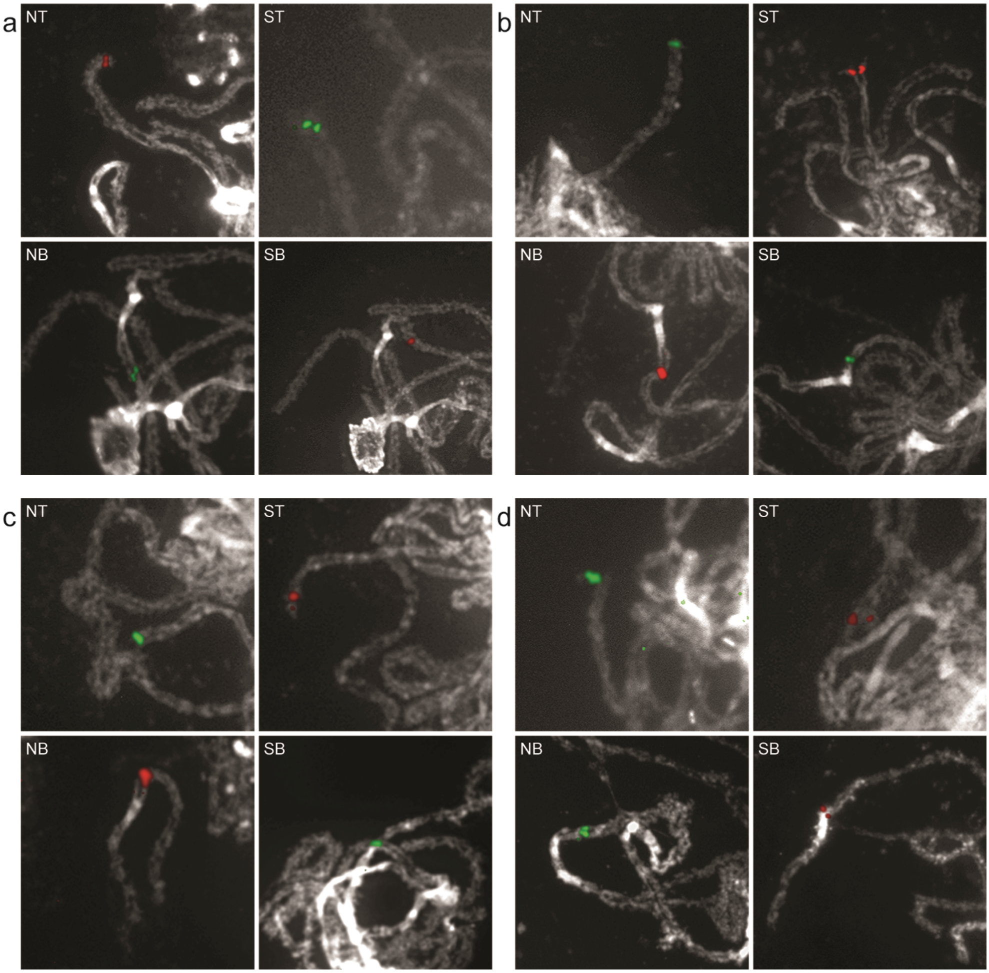


Figure 2.1: FISH pachytene chromosomes of *M. truncatula* using most distal and euchromatin-heterochromatin border BAC clones as probes. (NT) North telomere end. (ST) South telomere end. (NB) North euchromatin-heterochromatin border. (SB) South euchromatin-heterochromatin border. (a) Pachytene chromosome 1: (NT) BAC clone AC134822; (ST) BAC clone AC149471; (NB) BAC clone AC163324; (SB) BAC clone AC150977. (b) Pachytene chromosome 2: (NT) BAC clone AC140551; (ST) BAC clone AC146794; (NB) BAC clone AC145372; (SB) BAC clone AC135798. (c) Pachytene chromosome 3: (NT) BAC clone AC125481; (ST) BAC clone CU151877; (NB) BAC clone CT971479; (SB) BAC clone AC147407. (d) Pachytene chromosome 4: (NT) BAC clone AC146651; (ST) BAC clone AC162440; (NB) BAC clone AC135316; (SB) BAC clone AC174309.

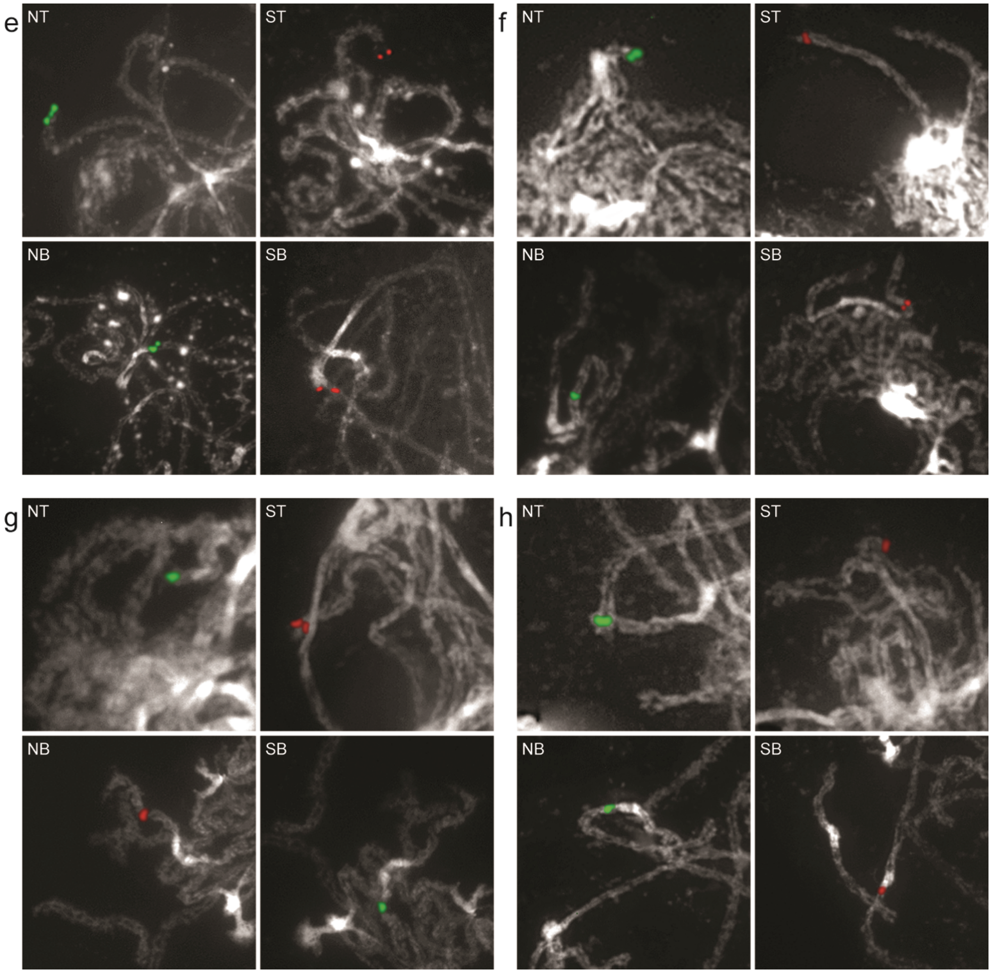


Figure 2.1: FISH pachytene chromosomes of *M. truncatula* using most distal and euchromatin-heterochromatin border BAC clones as probes. (NT) North telomere end. (ST) South telomere end. (NB) North euchromatin-heterochromatin border. (SB) South euchromatin-heterochromatin border. (e) Pachytene chromosome 5: (NT) BAC clone CR931730; (ST) BAC clone CT009656; (NB) BAC clone CR936326; (SB) BAC clone CR962123. (f) Pachytene chromosome 6: (NT) BAC clone AC174342; (ST) BAC clone AC134823; (NB) BAC clone AC135464; (SB) BAC clone AC141436. (g) Pachytene chromosome 7: (NT) BAC Clone AC146972; (ST) BAC clone AC146587; (NB) BAC clone AC146757; (SB) BAC clone AC157777. (h) Pachytene chromosome 8: (NT) BAC clone AC174350; (ST) BAC clone AC161043; (NB) BAC clone AC140024; (SB) BAC clone AC144724

papilionoids (Pfeil et al., 2005; Cannon et al., 2010). Near the time of this WGD, papilionoids radiated into several clades, the largest of which split quickly into two subclades, the Hologalegina (including *M. truncatula* and *L. japonicus*) and the milletioids (including *G. max* and other phaseoloids) at about 54 Myr ago (Lavin et al., 2005). We therefore compared *M. truncatula* pseudomolecules with other sequenced plant genomes to learn more about shared synteny and genome duplication history.

There is significant macrosynteny among *M. truncatula*, *L. japonicus* and *G. max* (Fig. 2.2 and Supplementary Fig. 5a, b). Conserved blocks, sometimes as large as chromosome arms, span most euchromatin in all three genomes. A given *M. truncatula* region is typically syntenic with one other *M. truncatula* region as a result of the approximately 58-Myr-ago WGD, usually in small blocks showing degraded synteny (Fig. 2.3 and Supplementary Fig. 6). A given *M. truncatula* region is most similar to two *G. max* regions via speciation at about 54 Myr ago and the *Glycine* WGD at <13 Myr ago (Schmutz et al., 2010) and less similar to two other *G. max* regions resulting from the ~58-Myr-ago and <13-Myr-ago WGD events. A *M. truncatula* region is likewise most similar to one *L. japonicus* region via speciation at about 50 Myr ago and less similar to a second *L. japonicus* region as a result of the ~58-Myr-ago WGD. Finally, each *M. truncatula* region and its homeologue typically show similarity to three *Vitis vinifera* regions via the pre-rosid WGH. Exceptions to these patterns could be due to gene losses, gains, or rearrangements specific to the *M. truncatula* lineage, resulting in synteny being more evident between *M. truncatula* and other genomes than in self-comparisons. Indeed, self-comparisons within *M. truncatula* reveal few remnants of the legume-specific WGD (Fig. 2.3 and Supplementary Fig. 6). Whereas this seems paradoxical, it is probably explained by extensive gene fractionation between WGD-derived homeologues in *M. truncatula*. In Fig. 2.4, two short regions on Mt1 and Mt3 resulting from the ~58-Myr-ago WGD are displayed beside microsyntenic regions of *G. max* and *V. vinifera*. As expected, many genes are microsyntenic between *M. truncatula* and *G. max* (ranging from 7/19 between Mt3 and Gm14 to 10/20 between Mt1 and Gm17). Between the two *M. truncatula* homeologues, however, only 6 out of 33

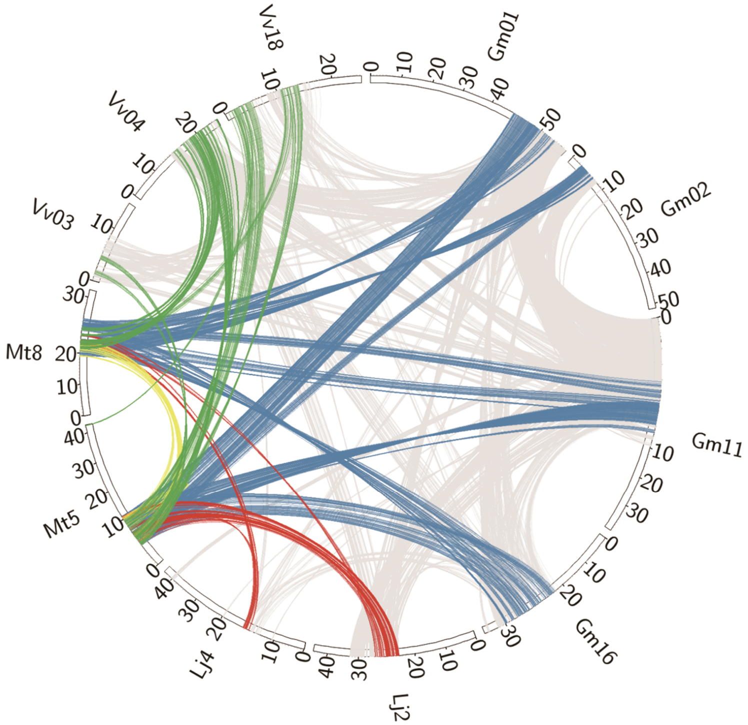


Figure 2.2: Circos diagram illustrating syntenic relationships between *Medicago*, *Glycine*, *Lotus* and *Vitis*. Homologous gene pairs were identified for all pairwise comparisons between *M. truncatula*, *G. max*, *L. japonicus* and *V. vinifera* genomes. Syntenic regions associated with the ancestral WGD events were identified by visually inspection of corresponding dot-plots. The large Mt5–Mt8 synteny block (yellow) was found to have two syntenic regions in *L. japonicus* (red), four syntenic regions in *G. max* (blue) and three in *V. vinifera* (green).

genes (or collapsed gene families) are microsyntenic, with a homeologue missing from one or the other duplicate (Supplementary Table 10). Apparently, there have been many more changes, large and small, in *M. truncatula* than in *G. max* since the legume WGD. This is borne out by the fact that synteny blocks in *M. truncatula* are one-third the length of those remaining from the papilionoid WGD in *G. max* (524 kb against

1,503 kb) with the average number of homologous gene pairs per block correspondingly lower (12.4 against 31.0).

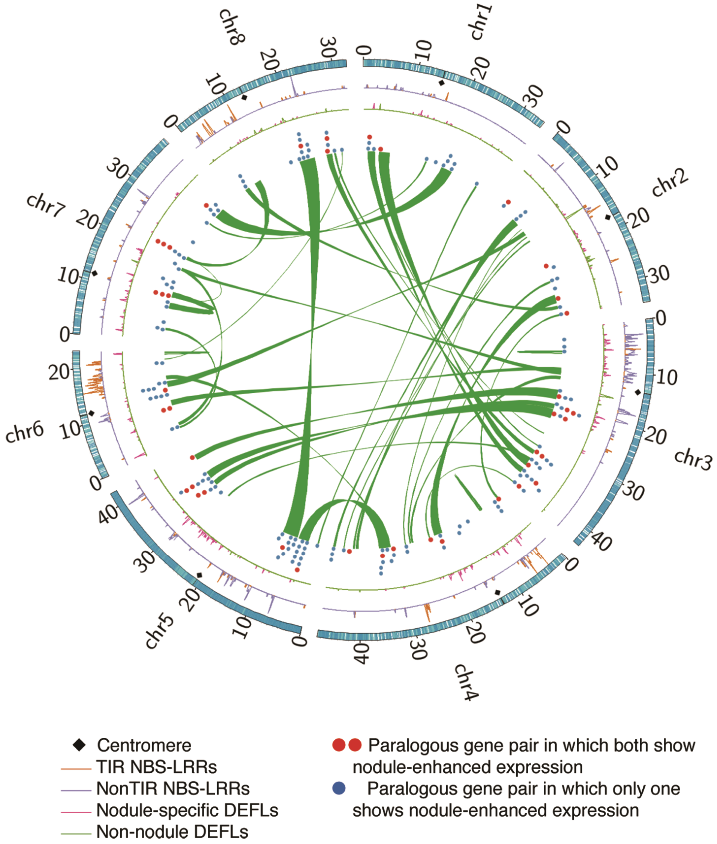


Figure 2.3: Circos diagram illustrating the *Medicago* WGD and selected gene families. The 963 WGD-derived paralogous gene pairs were examined for overlap with the nodule-enhanced gene list (Supplementary Data 2). Resulting gene pairs were joined and plotted as either blue cycles (only one of the duplicates is nodule-enhanced) or red (both nodule enhanced). Gene densities of NBS-LRRs, NCRs and other defensin-like proteins are plotted against chromosome position. Density was calculated using a sliding window (100-kb window with 50-kb steps).

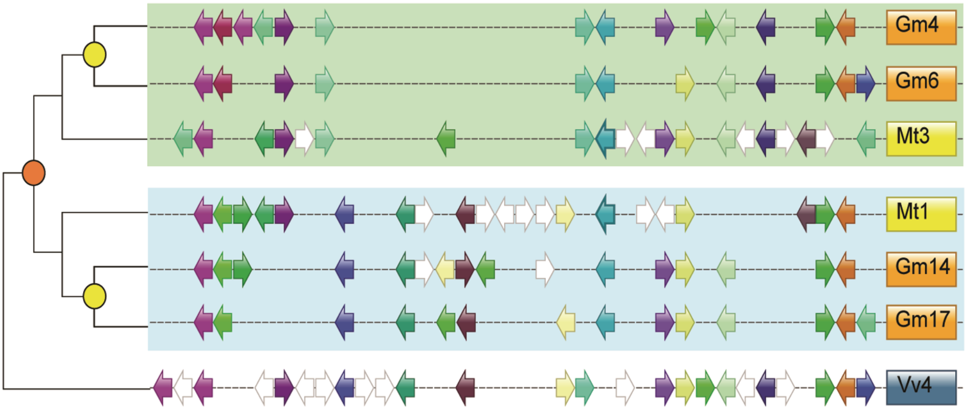


Figure 2.4: Microsynteny comparison between *Medicago* homeologues and corresponding regions of *Glycine* and *Vitis*. Microsyntenic genome segments are centred around Medtr3g104510/Medtr1g015890 (Supplementary Table 10), a duplicated region derived from the ~58-Myr-ago WGD event noted in orange. The <13-Myr-ago G. max-specific WGD is coloured yellow. Orthologous/ paralogous gene pairs are indicated through use of a common colour. White arrows represent genes with no syntenic homologue(s) in this genome region. Some of these genes may actually have a syntenic sequence in soybean but no corresponding model reported in the current annotation (http://www.phytozome.net/soybean).

The *M. truncatula* genome also has undergone high rates of local gene duplication. The ratio of related genes within local clusters compared to all genes in families is 0.339 in *M. truncatula*, 3.1-fold higher than in *G. max* and 1.6-fold higher than in *A. thaliana* or *P. trichocarpa*. ('Local clusters' are defined as genes in a family all within 100 gene models of one another.) The excess of local gene duplications in *M. truncatula* is observed genome-wide and affects many families. There are 2.63 times as many gene families with local duplications in *M. truncatula* compared with *G. max* (2,980 against 1,131), an excess that also is seen in detailed comparisons of syntenic regions in *M. truncatula* and *G. max*. We examined 16.3 Mb of Mt05 showing synteny to two large regions of Gm01 plus homeologous blocks on Gm02, Gm09 and Gm11. In these regions, 25.8% of *M. truncatula* genes are locally duplicated compared with just 8.0% in *G. max*. Local gene duplications and losses have contributed both to synteny disruptions (Fig. 2.4 and Supplementary Fig. 7) and to high gene count (62,388) in *M. truncatula*—a value nearly as high as the 65,781 total gene models in *G. max* despite its additional (<13 Myr ago) WGD. Local gene duplications are evident in certain gene families, such as F-box genes, which have undergone pronounced expansions (Supplementary Fig. 8 and Supplementary Table 11). *M. truncatula* also has experienced higher rates of base substitution compared to other plant genomes (Supplementary Fig. 9). Assuming 58 Myr ago as the date of the legume WGD, then the rate of synonymous substitutions per site per year in *M. truncatula* is 1.08×10^{-8}, 1.8 times faster than estimates in other vascular plants (Lynch and Conery, 2000). Higher rates of mutation and greater levels of rearrangement in *M. truncatula* following the papilionoid duplication may have been driven by factors including short generation times, high selfing rates or small effective population sizes, although these characteristics are not unique to *M. truncatula*.

Legumes and actinorhizal species are capable of forming a specialized organ, the root nodule, a highly differentiated structure hosting nitrogen-fixing symbionts. Phylogenetic studies suggest that nodulation may have evolved multiple times in the Fabidae, but the observation that all nodulating species are contained within this single clade indicates that a predisposition to nodulate evolved in their common ancestor (Soltis et al., 1995). It is unknown whether nodulation with rhizobia preceded the divergence of the three legume subfamilies or evolved on multiple occasions (Doyle and Luckow, 2003). Nevertheless, rhizobial nodulation and the 58-Myr-ago WGD are features common to most papilionoid legumes and both occurred early in the emergence of the group (Lavin et al., 2005). Given that WGDs generate genetic redundancy that potentially facilitates the emergence of novel gene functions without compromising existing ones (Freeling and Thomas, 2006), we examined the *M. truncatula* genome to ask whether the 58-Myr-ago WGD might have had a role in the evolution of rhizobial nodulation in *M. truncatula* and its relatives.

Nod factors are bacterial signalling molecules that initiate nodulation. Previous studies have shown that several of the plant components involved in the response to Nod factors also function in mycorrhizal signaling (Oldroyd and Downie, 2008). However, some Nod factor receptors and transcription factors have distinctly nodulation-specific functions. Among these nodulation-specific components, we found that the Nod factor receptor, *NFP*, and the transcription factor, *ERN1*, each have paralogues, *LYR1* and *ERN2* respectively, that trace back to the papilionoid WGD based on genome location and synonymous substitution rate values (Supplementary Fig. 10 and Supplementary Data 2). Both sets of gene pairs also show contrasting expression patterns and functional specialization. *NFP* and *ERN1* are expressed predominantly in the nodule and are known to function in nodulation (Arrighi et al., 2006; Middleton et al., 2007) , whereas *LYR1* and *ERN2* are highly expressed during mycorrhizal colonization (Supplementary Fig. 11). These observations indicate that two important nodulation-specific signalling components in *M. truncatula* might have evolved from more ancient genes originally functioning in mycorrhizal signalling and then duplicated by the 58-Myr-ago WGD. In the case of *M. truncatula NFP*/*LYR1*, this conclusion is supported by the observation that the apparent orthologue of *NFP* in the nodulating non-legume *Parasponia andersonii* functions in both nodule and mycorrhizal signaling (Op den Camp et al., 2011). Thus, the 58-Myr-ago WGD seems to have led to sub-functionalization of an ancestral gene participating in both interactions, resulting in two homeologous genes that each performs just one of the original functions.

To assess further the contribution of the WGD to *M. truncatula* nodulation, we analysed expression of paralogous gene pairs using RNA-seq data from six different organs (Supplementary Methods 5.1). A total of 963 WGD-derived gene pairs were found (Supplementary Data 2) with 618 pairs (1,046 genes) having RNA-seq data for one or both homeologue. We then determined the number of genes showing organ-enhanced expression (defined as genes with expression level in a single organ at least twice the level in any other) within the pseudomolecule and the WGD-derived gene sets (Supplementary Table 12). In both cases, different organs contained markedly different numbers of genes with enhanced expression (χ^2 with 5 degrees of freedom, $P = 10^{-272}$); however, the rank order among the organs was identical. Roots had the largest number of genes with enhanced expression followed by flower, nodule, leaf, seed/pod and bud. Among gene pairs with nodule-enhanced expression, both paralogues were nodule-enhanced in eight pairs, whereas just a single paralogue was nodule-enhanced in the other 43 pairs. This is consistent with nodulation pre-dating the WGD and further sub- and neo-functionalization emerging afterwards. We went on to examine transcription factors because they can act as regulators of plant growth and development. A total of 3,692 putative TF genes were discovered (Supplementary Data 3), representing 5.9% of all *M. truncatula* gene models (Supplementary Table 13). Of the 1,513 TF genes on pseudomolecules with RNA-seq data, 142 genes (9.4%) derived from the 58-Myr-ago WGD (Supplementary Fig. 12 and Supplementary Data 4), consistent with previous observations indicating greater retention of transcription factors following polyploidy (Thomas et al., 2006). Nodule-enhanced expression was significantly higher among transcription factors (92 out of 1,513 or 6.1%) than among all pseudomolecule genes (1,111 out of 23,478 or 4.7%) (χ^2 with 1 degree of freedom, $P = 0.024$) (Supplementary Table 12). Nodule-enhanced expression was even higher in WGD-derived transcription factors (11 out of 142 or 7.7%), although this enrichment did not reach statistical significance ($P = 0.113$). As expected, *ERN1* is found within this group of WGD-retained, nodule-enhanced transcription factors.

These results show that many paralogous genes retained from the 58-Myr-ago WGD, especially signalling components and regulators, have undergone sub- or neo-functionalization, including several with specialized roles in nodulation. Nevertheless, separate phylogenetic analyses (Supplementary Methods 5.5) indicate that some nodule-related genes derive from the more ancient pre-rosid WGH, with their nodule-related functions pre-dating the 58-Myr-ago WGD (Supplementary Data 5). Taken together, these results are consistent with a model where the capacity for primitive interaction with new symbionts derived from existing mycorrhizal machinery involving genes recruited from the pre-rosid WGH. This capacity would have arisen early in the Fabidae clade and led to the appearance of nodulation in multiple lineages (Kistner and Parniske, 2002; Doyle and Luckow, 2003). Later, the 58-Myr-ago WGD would have resulted in additional genes, including *NFP*, *ERN1* and the transcription factors described above, that went on to become specialized for nodule-related functions in the Papilionoideae.

Medicago contains additional amplified gene families, many nodulation-related and found in tandem clusters. *M. truncatula* has nine symbiotic leghaemoglobins, more than twice the number in *L. japonicus* or *G. max* (Supplementary Fig. 13). Five of these genes are located in a tight cluster on Mt5. The *M. truncatula* genome contains 593 nodule cysteine-rich peptides (NCRs) (Supplementary Data 6), a gene family restricted to *M. truncatula* and its relatives (Kato et al., 2002). NCRs are noteworthy because they include members essential for terminal differentiation of rhizobia (Van de Velde et al., 2010). NCRs are tightly clustered within the *M. truncatula* genome (Fig. 2.3), with 75% found in clusters of up to 11 members. The *M. truncatula* genome also has 764 nucleotide-binding site and leucine-rich repeat (NBS-LRR) genes (Supplementary Data 7), more than other plant genomes that have been sequenced so far (Meyers et al., 2003; Zhou et al., 2004; Yang et al., 2008), many with nodule-specific expression (Supplementary Fig. 14). Almost 90% of NBS-LRRs occur in clusters and genome regions showing limited macrosynteny to other species, such as Mt3 and Mt6, are locations of large NBS-LRR superclusters (Fig. 2.3 and Supplementary Tables 14 and 15). Finally, *M. truncatula* secretes flavonoid signalling molecules to induce the *nod* genes of *Sinorhizobium meliloti* (Peters et al., 1986). In *M. truncatula*, the corresponding biosynthetic pathway has expanded markedly, with 28 *M. truncatula* chalcone synthase genes in clusters of up to seven members compared to just four chalcone synthases in *A. thaliana* (Winkel-Shirley, 2001) (Supplementary Data 8). *M. truncatula* has ten chalcone reductases compared to none in *A. thaliana* (Hegnauer and Grayer-Barkmeijer, 1993) and *M. truncatula* has 11 chalcone isomerase genes, including one cluster of seven members, compared to just one representative in *A. thaliana* (Shirley et al., 1995) (Supplementary Figs 15 and 16).

Analysis of the *M. truncatula* genome supports earlier studies indicating that the dramatic radiation of the legume family (at least the papilionoid subfamily) is partly attributed to the 58-Myr-ago WGD (Singer et al., 2009). Our results indicate that the WGD early in papilionoid evolution allowed the emergence of critical components in Nod factor signalling and contributed to the complexity of rhizobial nodulation observed in this clade. As such, the WGD seems to have had a crucial role in the success of papilionoid legumes, enhancing their utility to humans.

Methods summary

DNA sequencing

Six A17 BAC and one fosmid library were used to create Mt3.5 (Supplementary Table 1). Most were processed by Sanger paired-end sequencing of 3–6-kb shotgun libraries. Sequences were downloaded in February/March 2009 with scaffolding performed by aligning all BAC and fosmid ends against contigs and then anchored and ordered primarily by optical mapping. Separately, 25 billion base pairs (Gb) of Illumina sequence was generated using short (375 nt) inserts plus 2.1 Gb from a 5 kb mate-pair library, then assembled using CLCbio (http://www.clcbio.com) and Soap (http://soap.genomics.org.cn/).

RNA sequencing

Five tissues were used for RNA-seq analysis with ~10 million Illumina 36-bp reads per library (Supplementary Table 12). Three tissues were used for small RNA analysis with ~3 million reads per Illumina library (Supplementary Figs 17–18, Supplementary Table 16 and supplementary data 9).

CHAPTER 3

Cytogenetic characterization of euchromatic gene-space of the model legume species *Medicago truncatula*

Chunting Lang[1], Carolien Franken[1], Marijk Hartog[1], Ton Bisseling[1], Rene Geurts[1]

1. Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands

Abstract

The high quality sequence of the euchromatic portion of the *Medicago truncatula* genome has been produced largely by a BAC-by-BAC sequencing approach. To support sequencing, we determined the precise physical position of selected BACs on pachytene chromosomes by fluorescent *in situ* hybridization (FISH). We delineated the gene space of the 16 euchromatic chromosome arms by identifying the most distal BAC clones that mark telomeres and euchromatin-heterochromatin border regions of all individual chromosomes. These cytogenetic studies revealed that 96% of the euchromatin is located between these terminal BAC sequence chromosomal targets, through gaps remain within the sequence assembly. We measured gap sizes within the genome sequence using either two-color FISH with individual BACs as probes or pooled-BAC FISH. Thus we showed that the sequence coverage on the euchromatic part of *M. truncatula* chromosome 5 is more than 99% complete, which is consistent with sequencing data. However we identified one very large gap in the euchromatic part of short arm and one smaller gap in the long arm of *M. truncatula* chromosome 6. Furthermore, we also resolved an inconsistency between the genetic map and assembled sequence with respect to chromosomes 4 and 8. Taken together these studies underline the importance of cytogenetics to generate a high quality plant genome sequence.

Introduction

The legume family (*Fabaceae*) is, with more than 18,000 species, the third largest family of flowering plants. Many of its members are important for agriculture and the environment due to their capability to form nitrogen-fixing root nodules. One of the model species for the legume family and for study of the process of nodulation is *Medicago truncatula*, for which a knowledge base of genomic information is being developed. One important aspect of this knowledge base is a whole genome sequence, which is being developed largely by using a BAC-by-BAC sequencing strategy. While such strategy has resulted in very high quality genome sequences for other species in the past (The Arabidopsis Genome, 2000; The Rice genome, 2005; Schnable et al., 2009), it remains important to monitor sequencing progress and perform accurate gap-size measurements to determine the quality and reliability of the resulting sequence assembly using independent means. Here, we provide an overview of the cytogenetic studies that we have applied to support genome sequencing of *M. truncatula* using fluorescent *in situ* hybridization (FISH).

To study any biological process, a genome sequence of a model organism in which the process occurs, provides an excellent resource. For instance, the model species for plant biology in general is *Arabidopsis thaliana*, of which the genome sequence has contributed to understand many biological processes and has been used as a reference to study the genome structure of related plant species. However, not all processes of interest in plant biology occur in *A. thaliana*. Members of the legume family are basically the only plant species which can live in symbiosis with nitrogen fixing rhizobium bacteria. The rhizobium-legumes symbiosis leads to the formation of root nodules in which the bacteria partner is hosted. In these nodules molecular nitrogen is converted to ammonia by the bacterial nitrogenase enzyme complex. This forms a first step in the production of many N-containing compounds; e.g. amino acids. Due to the capacity of biological nitrogen fixation legumes can reduce the dependence of agriculture on artificial fertilization.

The taxonomic family of *Fabaceae* is divided into three subfamilies; *Mimosoideae*, *Caesalpinioideae*, and *Papilionoideae*, which all contain species that can form rhizobium root nodules (Doyle and Luckow, 2003). Nodulation is most prominent in the largest subfamily *Papilionoideae*, which contains four major crown clades (Genistoid, Dalbergioid, Hologalegina and Millettioid) (Lavin et al., 2005). Two large clades, Hologalegina and Millettioid, harbor important agricultural crops, for instance *Medicago sativa* (alfalfa), *Cicer arietinum* (chickpea) and *Pisum sativum* (pea) in Hologalegina and *Glycine max* (soybean), *Vigna unguiculata* (cowpea) and *Phaseolus vulgaris* (common bean) in Millettioid (Lavin et al., 2005). Because of the high number of important agriculture crops there is great interest in understanding legume biology, especially symbiotic nitrogen fixation.

To create resources similar to those available for *A. thaliana* among others, *M. truncatula* was selected as legume model (Young et al., 2005). *M. truncatula* was chosen because it was already used to study symbiotic nitrogen fixation. The species is self-pollinating, has a short generation cycle, and a relatively small diploid (2n = 2x = 16) genome size of ~500 Mb, similar to the genome size of *Oryza sativa* (rice) (~430 Mb) and four times larger than *A. thaliana* (~125 Mb). *M. truncatula* genetic maps, a physical map and sequenced EST libraries have been published prior the initiation of the sequencing project (Choi et al., 2004b). The classic genetic map is built based on amplified fragment length polymorphism (AFLP) and randomly amplified polymorphic DNA (RAPD) markers (Thoquet et al., 2002; Choi et al., 2004b). However, the sequence of the *M. truncatula* genome (chapter 2) provides the opportunity to collect expressed sequence tags (EST) markers, which have been positioned on the genetic maps and also used as reference to study the genome conservation between *M. truncatula* and *A. thaliana* or other legume species (Choi et al., 2004b; Choi et al., 2004a). Recently, the genome sequence of *M. truncatula* has been completed (chapter 2).

The *M. truncatula* genome sequencing project focused on the euchromatic arms that presumably contain most of the genes. The remaining heterochromatic chromosome regions include pericentromeric blocks that are made-up largely of repetitive sequences. To sequence the euchromatin, primarily a BAC-by-BAC strategy has been applied that was combined with a whole genome shotgun approach for completion of the project. Bacterial artificial chromosome (BAC) clones were constructed and arranged in a physical map (Gamas et al., 2006). Subsequently a minimum tiling path of BACs has been sequenced and assembled on a BAC-by-BAC basis (Young et al., 2005). This resulted in gene-dense sequence containing 17.7 genes / 100 kb (1 gene every 5.9 kb) (chapter 2), although repetitive elements also occur in these regions. However no BAC clones of the euchromatic regions have been found that consisted exclusively of repetitive sequences. This is further confirmed by cytogenetic research using fluorescent *in situ* hybridization (FISH) on *M. truncatula* pachytene chromosomes. The euchromatin and pericentromeric heterochromatin are relatively easy to distinguish on pachytene chromosome spreads (Kulikova et al., 2001). FISH mapped *M. truncatula* gene-rich BAC clones are all localized on the euchromatin (Kulikova et al., 2001; Choi et al., 2004b; Kulikova et al., 2004). Therefore, the chosen sequencing strategy indeed covers the *M. truncatula* gene-space.

Ideally, the resulting sequence comprises the complete 16 chromosome arms, however, in practice, gaps still remain. To visualize the extent of genome sequence coverage and measure gap sizes, we applied FISH. Here, we present the estimation of the genome sequence coverage between telomeres and the boundaries of euchromatin and pericentromeric heterochromatin. Also we measured gap sizes between physically mapped BACs. Furthermore we resolved an inconsistency between the genetic map and assembled sequence regards to chromosome 4 and 8 using FISH.

Results

Estimation of genome sequence coverage

The *M. truncatula* genome sequencing project initially aimed to generate a high quality sequence of the euchromatic regions of all 16 chromosome arms by using a BAC-by-BAC based strategy. To achieve this a high-density physical map with 12x genome coverage was created and integrated with the genetic map. Based on optical mapping the euchromatic portion is estimated to span 375 Mb. The BAC-by-BAC approach delivered 263 Mb sequence information of this region, resulting in coverage of ~70% (chapter 2). The remaining sequences have been subsequently obtained by applying next-generation sequencing based on a whole genome shotgun approach delivering an additional 104 Mb of unique sequences. Though these novel sequences originate from heterochromatic as well as euchromatic parts of the genome. Taken together, the 367 Mb of unique DNA sequence of the *M. truncatula* genome that has been generated covers ~94% of the expressed genes.

To determine whether the resulting sequence assembly covers most of the *M. truncatula* euchromatin, we aimed to position the euchromatin-heterochromatin borders by integrating a pachytene based cytogenetic map with the genetic and physical maps. For each individual chromosome arm we aimed to identify BAC-based markers that represent the euchromatin-heterochromatin border regions as well as the distal ends of each chromosome. To obtain candidate markers we made use of the available genetic map, which contains mainly EST-derived markers. Heterochromatin is known to be a gene-poor region (low gene density), which means heterochromatic regions are underrepresented on such genetic map. On all 8 linkage groups, such low marker coverage regions can be found (Fig. 3.1). We selected markers that flank these marker poor regions and used the corresponding BAC clones as probes in FISH experiments. Some of these BAC clones did not provide a single FISH hybridization signal on the pachytene chromosome spreads, even with use of Cot-100 as a blocking reagent, presumably due to presence of large numbers of repetitive sequences. In those cases, flanking markers deeper into the euchromatic regions were selected, until BAC clones were obtained that resulted in single signal locus in FISH experiments. Then, the distance was measured between the heterochromatin flanking marker and the end of each chromosome arm.

In *M. truncatula* pachytene chromosome spreads, euchromatin-heterochromatin borders can be directly observed based on staining intensities. As *M. truncatula* has 8 chromosomes, 16 BAC clones for the euchromatin–heterochromatin borders and 16 at the distal chromosome ends were searched by hybridization to pachytene chromosomes. Eleven BAC-marker clones could be identified that hybridized in the direct vicinity of the euchromatin–heterochromatin borders. In case of the distal chromosome ends 7 such BAC-marker clones could be identified that give a unique hybridization signal at the distal end of the chromosome (Fig. 3.2a-h). In the other cases, some physical distance was observed between the hybridization signal and actual telomere ends or euchromatin–heterochromatin borders (Fig. 3.2a-h). We measured the microscopic distances between hybridization signal from our terminal BACs and euchromatin–heterochromatin borders or telomere ends (Table 3.1).

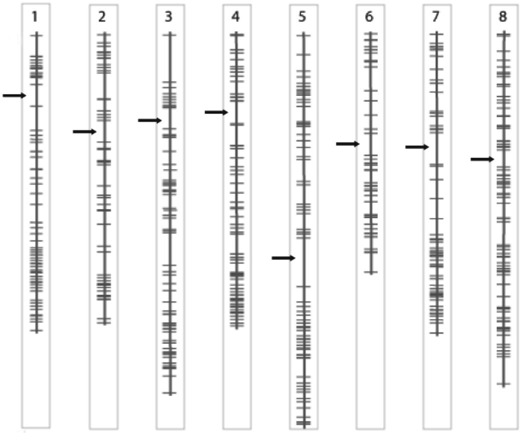


Figure 3.1: Eight genetic linkage groups of *M. truncatula*, constructed using gene-based genetic markers. Arrows point to regions with a low marker coverage (Unpublished, available on http://medicagohapmap.org).

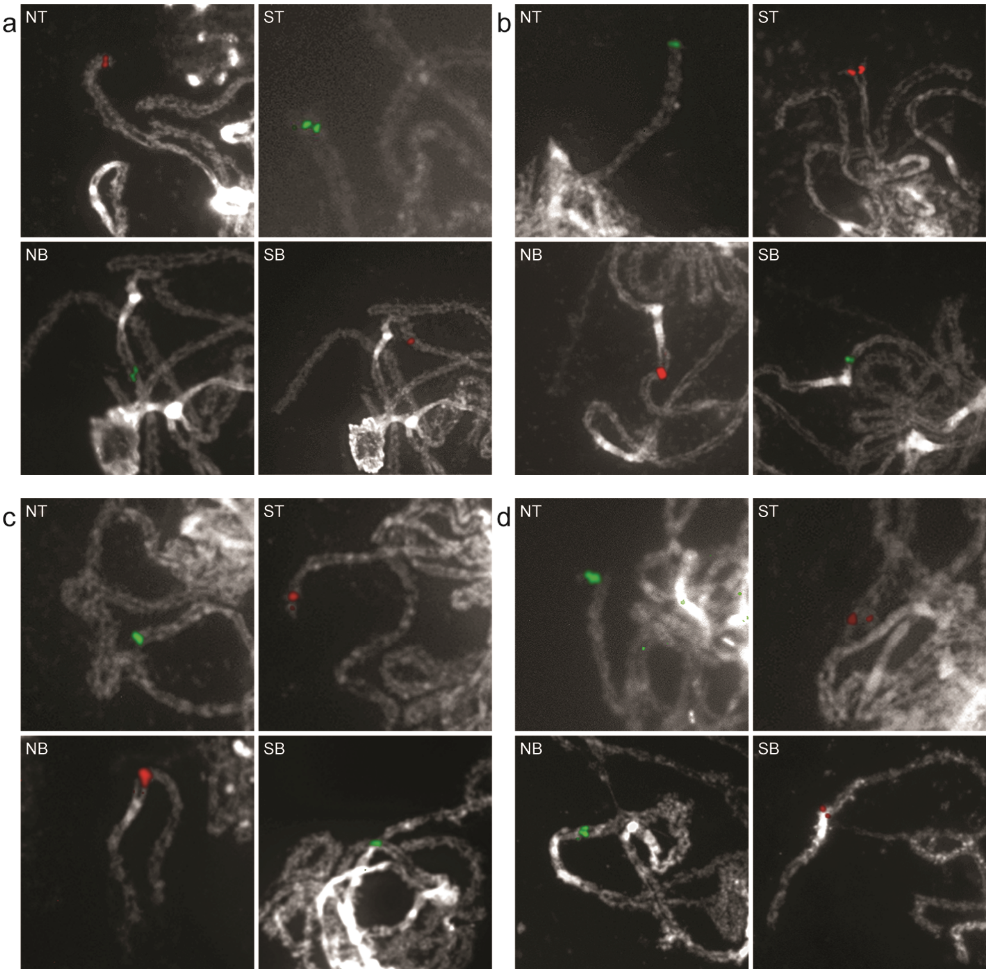


Figure 3.2: FISH pachytene chromosomes of *M. truncatula* using most distal and euchromatin-heterochromatin border (directly observed based on DAPI staining intensities) BAC clones as probes. (NT) North telomere end. (ST) South telomere end. (NB) North euchromatin-heterochromatin border. (SB) South euchromatin-heterochromatin border. (a) Pachytene chromosome 1: (NT) BAC clone AC134822; (ST) BAC clone AC149471; (NB) BAC clone AC163324; (SB) BAC clone AC150977. (b) Pachytene chromosome 2: (NT) BAC clone AC140551; (ST) BAC clone AC146794; (NB) BAC clone AC145372; (SB) BAC clone AC135798. (c) Pachytene chromosome 3: (NT) BAC clone AC125481; (ST) BAC clone CU151877; (NB) BAC clone CT971479; (SB) BAC clone AC147407. (d) Pachytene chromosome 4: (NT) BAC clone AC146651; (ST) BAC clone AC162440; (NB) BAC clone AC135316; (SB) BAC clone AC174309.

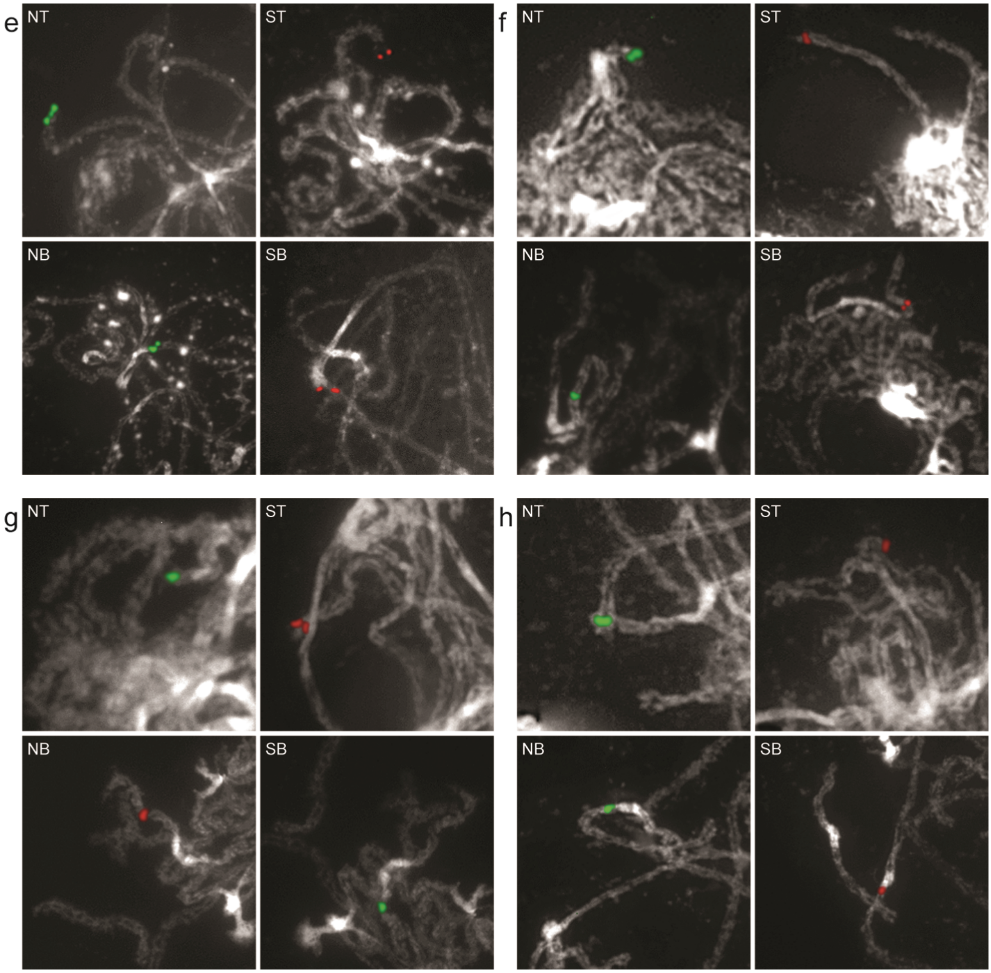


Figure 3.2: FISH pachytene chromosomes of *M. truncatula* using most distal and euchromatin-heterochromatin border BAC clones as probes. (NT) North telomere end. (ST) South telomere end. (NB) North euchromatin-heterochromatin border. (SB) South euchromatin-heterochromatin border. (e) Pachytene chromosome 5: (NT) BAC clone CR931730; (ST) BAC clone CT009656; (NB) BAC clone CR936326; (SB) BAC clone CR962123. (f) Pachytene chromosome 6: (NT) BAC clone AC174342; (ST) BAC clone AC134823; (NB) BAC clone AC135464; (SB) BAC clone AC141436. (g) Pachytene chromosome 7: (NT) BAC Clone AC146972; (ST) BAC clone AC146587; (NB) BAC clone AC146757; (SB) BAC clone AC157777. (h) Pachytene chromosome 8: (NT) BAC clone AC174350; (ST) BAC clone AC161043; (NB) BAC clone AC140024; (SB) BAC clone AC144724

Chr	Physical position	BAC accession	Micro-scopic distance (μm)	BAC to BAC (μm)	Estimated physical distance (Mb)	Actual eu-chromatin (μm)	Sequenced-EU coverage (%)
1-N	NT	AC134822	1.1	15.9 ± 0.3	13.5	19.9 ± 3.8	80
	NB	AC163324	2.8				
1-S	SB	AC150977	2.4	35.9 ± 0.1	30.5	39 ± 8.0	92
	ST	AC149471	0.8				
2-N	NT	AC125481	0.4	20.7 ± 0.9	17.6	22.9 ± 4.7	91
	NB	AC145372	0.8				
2-S	SB	AC135798	0	30.1 ± 0.2	25.6	30.8 ± 6.1	98
	ST	AC146794	0.4				
3-N	NT	AC125481	0	18.6 ± 0.6	15.8	18.8 ± 3.6	100
	NB	CT971479	0				
3-S	SB	AC147407	0	54.3 ± 0.1	46.2	54.8 ± 2.6	99
	ST	CU151877	0.4				
4-N	NT	AC146651	0.6	17.0 ± 0.4	14.5	17.7 ± 2.8	96
	NB	AC135316	0				
4-S	SB	AC174309	0	47.3 ± 0.2	40.2	47.4 ± 5.7	100
	ST	AC161034	0				
5-N	NT	CR931735	0	23.8 ± 0.3	20.2	23.7 ± 3.6	100
	NB	CR936326	0				
5-S	SB	CR962123	0	26.5 ±0.5	22.5	26.9 ± 1.4	100
	ST	CT009656	0				
6-N	NT	AC174372	0	13 ± 0.3	11.0	13.2 ± 5.7	100
	NB	AC135464	0				
6-S	SB	AC141436	1.3	16.3 ± 0.1	13.9	17.6 ± 4.0	92
	ST	AC134823	0.1				
7-N	NT	AC146972	0	16.6 ± 0.2	14.1	19 ± 5.1	87
	NB	AC146757	2.3				
7-S	SB	AC157777	0	35.3 ± 0.2	30.0	35.6 ± 5.0	99
	ST	AC146587	0.5				
8-N	NT	AC174350	0.4	19.8 ± 0.2	16.8	20.2 ± 5.0	96
	NB	AC140024	0				
8-S	SB	AC144724	0	31.2 ± 0.1	26.5	31.2 ± 7.3	100
	ST	AC162440	0				
Total			14.3	422.3 ± 11.1	358.9	438.7 ± 31.7	96

Table 3.1: The size of physical distances between selected BACs and actual telomere, euchromatin and heterochromatin border of *M. truncatula* pachytene chromosomes 1 to 8.

For instance in case of chromosome 1, none of the four mapped BAC-marker clones were at the very end of the euchromatic portion of both chromosome arms. Instead, there was a microscopic distance between terminal BACs of 1.1 µm to the north telomere end, 0.8 µm to the south telomere end, 2.8 µm and 2.4 µm to the north and south borders of pericentromeric heterochromatin, respectively (Fig. 3.2a, Table 3.1). Next, we determined the length of the total euchromatic region. Eight pachytene complements were measured resulting in a total length of 438.7 ± 31.7 µm. As it is estimated that the euchromatin is 375 Mb in length, hence the euchromatin has an average condensation degree of 0.85 Mb / µm. The identified markers at the telomere ends and the euchromatin-heterochromatin borders span a total 422.3 ± 11.1 µm (Table 3.1); ~360 Mb, which is 96% of the total euchromatin. Taken together, we conclude that the identified BAC-markers delineate the euchromatic portion of the *M. truncatula* genome.

FISH-based validation of completeness of genome sequencing

While the existing genome sequence covers largely the euchromatic portion of the genome, only 70% is covered by non-redundant high quality BAC-based sequence. As a consequence 273 physical gaps are present. We aim to investigate the size of some of these gaps on basis of 3 chromosomes; namely 5, 6, and 8. The sequence coverage of these 3 chromosomes differs, where the sequence of chromosome 5 is of high quality, chromosome 8 of average quality and chromosome 6 of relatively poor quality.

For chromosome 5, ~42.5 Mb of non-redundant BAC-based sequence has been obtained. Based on physical length of the euchromatin in the pachytene complements (50.6 ± 4.6 µm) we estimated a total length of ~43 Mb (condensation degree 0.85 Mb / µm). Based on this it can be concluded that the euchromatic portion of chromosome 5 has been largely covered. In total 8 gaps are still present in the assembled sequences. To validate the claim of relative small gaps on this chromosome, we estimated the size of all gaps. FISH was applied using BAC clones on either side of each gap. They were labeled with either digoxigenin-dUTP or biotin-dUTP that can be detected either by green or red fluorescence. For 6 gaps the flanking BAC clones hybridized in a close vicinity to each other, showing yellow fluorescence as result of overlapping signal (Fig. 3.3a-f). No gap between both hybridization signals could be detected on pachytene complements. This suggests that the physical gaps are below the resolution limits of pachytene FISH. The 2 remaining gaps displayed microscopic size of 0.1 µm (0.085 Mb) and 0.6 µm (0.51 Mb), respectively (Fig. 3.3g, h; Table 3.2). However the gap of 0.51 Mb has been closed because previously unmapped *M. truncatula* BAC clones were assembled in this gap, covering 0.7 Mb. Thus, currently 7 gaps remain within the chromosome 5 euchromatin sequence, of which the largest is less than 100 kb. Based on these findings we conclude that the euchromatic arms of chromosome 5 are largely covered by the currently available sequence.

In case of chromosome 8, ~32.6 Mb of non-redundant BAC-based sequence was obtained. Cytogenetic analysis indicated a total length of euchromatic arms is 51 ± 0.4 µm, which could correspond to ~43.4 Mb (average condensation degree 0.85 Mb / µm). This suggests a coverage of ~75% of the total euchromatin. We applied a strategy similar as used for chromosome 5 and studied 8 gaps on chromosome 8. Eight pairs of BAC clone each flanking a gap were used as probes and hybridized to pachytene chromosomes. Seven of these BAC clone pairs visualized a physical gap, whereas for one pair the hybridization signals of both BAC clones partially overlapped (Fig. 3.4a-h). Gap sizes were measured and listed in table 3.3. A cumulative of ~1.85 µm was measured over 8 gaps, representing a missing of ~1.57 Mb of sequence on chromosome 8 (average condensation degree 0.85 Mb / µm).

Next, we studied chromosome 6. Although the shortest chromosome in length based on cytogenetic studies (43.4 ± 3.1 µm), it has more pericentromeric heterochromatin (28.9%) than other chromosomes in *M. truncatula* (Kulikova et al., 2001; Choi et al., 2004b). Also the euchromatin of the chromosome arms is interspersed with small heterochromatic knobs. Approximately 17.9 Mb non-redundant BAC-based sequence has been obtained from chromosome 6.

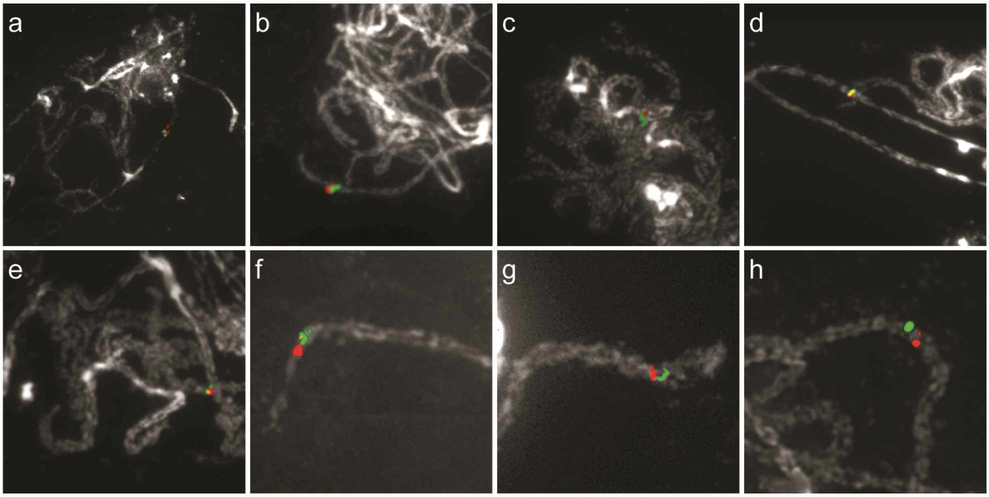


Figure 3.3: Gap sizing of *M. truncatula* (pachytene) chromosome 5 using: (a) BAC clones AC142526 (green) and CR962124 (red). (b) FISH of BAC clones CT009655 (green) and CT009651 (red). (c) FISH of BAC clones CT963108 (green) and CT573076 (red). (d) FISH of BAC clones CG911592 (green) and CR962133 (red). (e) FISH of BAC clones CT027661 (green) and CU459036 (red). (f) FISH of BAC clones CU424494 (green) and CU469551 (red). (g) FISH BAC clones CT963113 (green) and CU326391 (red). (h) FISH of BAC clones AC161405 (green) and CR962188 (red).

Chr	BAC clone	BAC accession	BAC clone	BAC accession	Microscopic distance (μm)	Estimated physical gap size (Mb)
5	mth2-34h20	AC142526	mth2-44c15	CR962124	0	0
5	mth2-121g11	CT009655	mte1-70c15	CT009651	0	0
5	mth4-37c5	CT963108	mth2-181d5	CT573076	0	0
5	Mth2-1p11	CG931592	mth2-17h8	CR962133	0	0
5	Mth2-70c9	CT027661	mte1-64g13	CU459036	0	0
5	Mth2-103o8	CU424494	mth2-79n1	CU469551	0	0
5	mth2-181g12	CT963113	mth2-182c4	CU326391	0.1	0.08
5	Mth2-33p21	AC161405	mth2-2o11	CR962188	0.6	0.48

Table 3.2: Gap sizes in *M. truncatula* chromosome 5. Estimated gap size (Mb) is calculated by using microscopic distance (μm) * average condensation degree (0.85 Mb / μm).

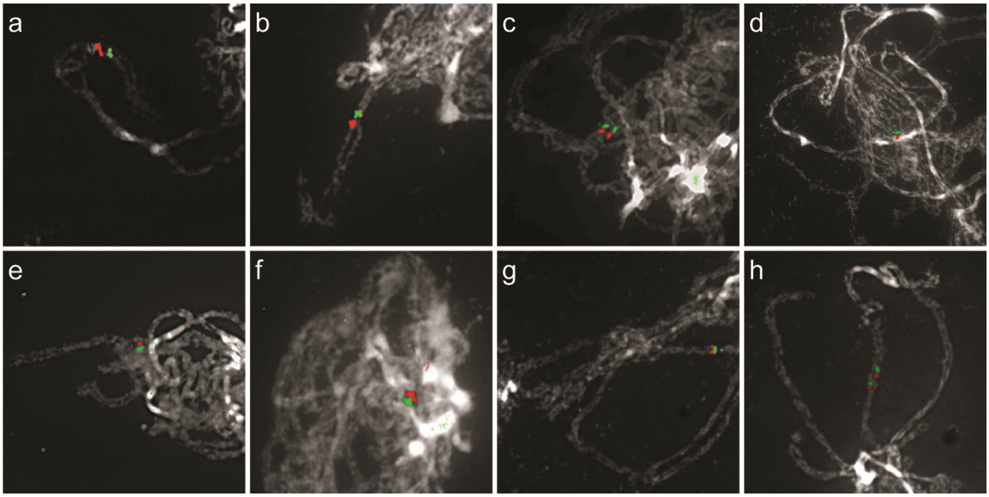


Figure 3.4: Gap sizing of *M. truncatula* (pachytene) chromosome 8. (a) FISH of BAC clones AC148398 (green) and AC146708 (red). (b) FISH of BAC clones AC146720 (green) and AC125480 (red). (c) FISH of BAC clones AC159706 (green) and AC126784 (red). (d) FISH of BAC clones AC137703 (green) and AC147000 (red). (e) FISH of BAC clones AC146791 (green) and AC153643 (red). (f) FISH of BAC clones AC225467 (green) and AC202357 (red). (g) FISH BAC clones AC187278 (green) and AC136839 (red). (h) FISH of BAC clones AC234853 (green) and AC225460 (red).

Chr	BAC clone	BAC accession	BAC clone	BAC accession	Microscopic distance (μm)	Estimated physical gap size (Mb)
8	mth2-24c23	AC148398	mth2-80a22	AC146708	0.62	0.5
8	mth2-17n5	AC146720	mth2-8c24	AC125480	0.15	0.12
8	mth2-187h4	AC159706	mth2-36b12	AC126784	0.22	0.18
8	mth2-11d24	AC137703	mth2-124a16	AC147000	0.14	0.11
8	mth2-123m17	AC146791	mth2-109d3	AC153643	0.22	0.18
8	mth2-174e10	AC225467	mth2-52g1	AC202357	0	0
8	mth2-189b17	AC187278	mth2-13n2	AC136839	0.1	0.08
8	mth2-69f20	AC234953	mth2-166h7	AC225460	0.14	0.11

Table 3.3: Gap sizes in *M. truncatula* chromosome 8. Estimated gap size (Mb) is calculated by using microscopic distance (μm) * average condensation degree (0.85 Mb / μm).

Based on total length of 30.8 ± 4.4 µm of visible euchromatin and a similar condensation degree as used for other chromosomes (Table 3.1), we estimated that the euchromatin of chromosome 6 is ~26.2 Mb. Therefore, the current obtained sequence of chromosome 6 represents approximately 68% of total DNA content of the euchromatic arms. This makes chromosome 6 the least completed chromosome based on the BAC-by-BAC sequencing and implies that relative large gaps may exist. To visualize and measure the gaps on microscopic scale, we applied pool BAC FISH on chromosome 6. BAC clones of *M. truncatula* chromosome 6 were picked along the physical map with a spacing of ~400 kb on average. Twenty *M. truncatula* BACs from the short arm and 18 BACs from the long arm of *M. truncatula* chromosome 6 were selected and labeled with biotin-dUTP (for red fluorescent detection) and digoxigenin-dUTP (for green fluorescent detection), respectively. A large gap was observed on the short arm (Fig. 3.5a), which was ~1.8 µm, corresponding with approximately 1.53 Mb (average condensation 0.85 Mb / µm). A small gap was observed on the long arm (Fig. 3.5b), which was 0.25 µm (0.2 Mb). This indicates that several major gaps are still present in chromosome 6.

Resolving a discrepancy between the genetic map and assembled sequence of *M. truncatula* chromosomes 4 and 8

When the genetic map and assembled sequence of *M. truncatula* were compared a discrepancy was noted. Part of the south bottom arms of chromosome 4 (LG4) and 8 (LG8) are swapped when compared to the genetic map. According to the assembled sequence, the south arm of chromosome 4 belongs to the south arm of chromosome 8 from the telomere to the marker at 58.3 cM that is represented by BAC clone AC146585. The south arm of chromosome 8 is part of south arm of chromosome 4 from the telomere to the marker at 50.2 cM that is represented by BAC clone AC137703. To determine whether the genome sequence assembly reflects the correct genome structure of the sequenced reference line (Jemalong A17), we applied FISH. To do so, BAC clones that contain genetic markers genetically positioned adjacent to the swapped regions were used as probes.

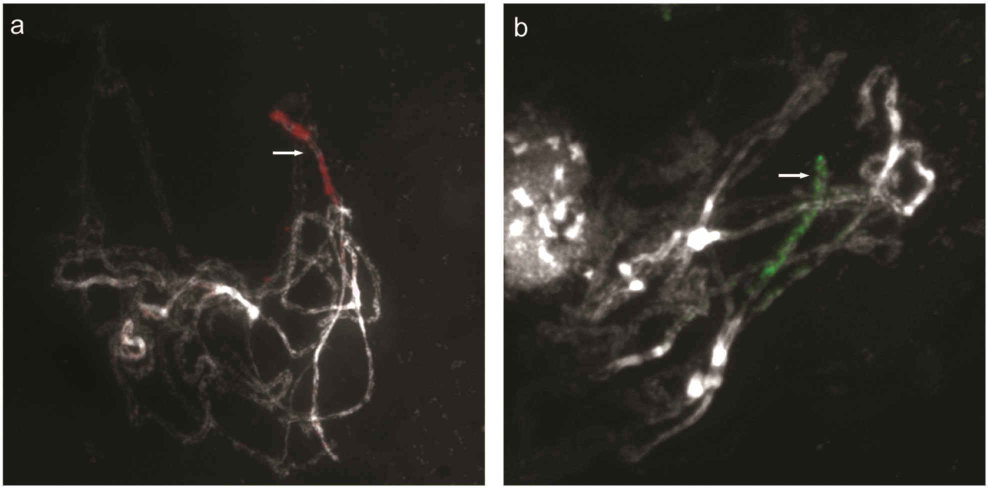


Figure 3.5: Coverage of *M. truncatula* (pachytene) chromosome 6. (a) 20 BACs of chromosome 6 short arm were labeled with biotion-dUTP (red fluorescence). The arrow indicated the region of a large gap. (b) 18 BACs of chromosome 6 long arm were labeled with digoxigenin-dUTP and detected for green fluorescence. The arrow indicated the region of a smaller gap on the long arm.

BAC clone AC144564 and AC174339 are genetically mapped to LG4. These two BACs were used as reference for chromosome 4. BAC clone AC137703 was genetically mapped on LG8 at 50.2 cM, but is part of the sequence assembly of chromosome 4. Together these three BAC clones (AC144564, AC174339 and AC137703) were used as probes in a FISH experiment, in which AC144564 was labeled with Cy3-dUTP (orange signal detection), AC174339 was labeled with FITC-dUTP (green signal detection) and AC137703 was labeled with Cy3.5-dCTP (red signal detection). FISH experiments showed that these three BAC clones were located on the same chromosome arm (Fig. 3.6a). BAC clone AC137703 mapped at a microscopic distance of 5.1 µm apart from BAC clone AC174339 (Fig. 3.6a). To confirm this result, the telomeric BAC AC162440 that was genetically mapped to the south arm of chromosome 8 was used as probe together with BAC clones AC174339 and AC137703 in a independent FISH experiment, where AC174339 was labeled with FITC-dUTP (green signal), AC137703 was labeled with Cy3-dUTP (orange signal) and AC162440 was labeled with Cy3.5-dUTP (red signal). The FISH result indicated that the most south telomere end BAC AC162440 was indeed located on the same chromosome as the other two BACs, which were mapped on the chromosome 4 (Fig. 3.6b). Conversely, BAC clone AC137839 was used as reference for LG8 and BAC clone AC137703 for LG4, together with AC146585 that genetically mapped on LG4 were performed in a FISH experiment. The result showed that BAC clone AC146585 (red) was located on the same chromosome (#8) as AC137839 (green), whereas AC137703 (orange) was located on different chromosome (Fig. 3.6c). The telomeric BAC AC161034 that was genetically mapped to the south arm of chromosome 4 was used as probe together with BAC clone AC146585 and AC171267 in a FISH experiment. This revealed that most south telomere end BAC AC161034 was located on the same chromosome (#8) as the other two BACs (Fig. 3.6d). Taken together, we conclude that part of south bottom arm of chromosome 4 should be swapped with south bottom arm of chromosome 8 in the genetic map (Fig. 3.6e), while the sequence assembly is correct.

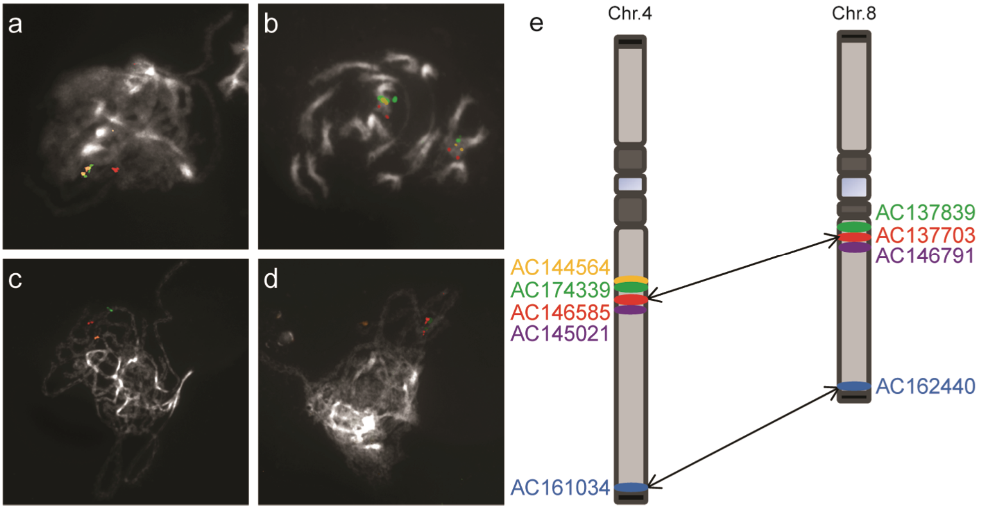


Figure 3.6: FISH mapping of BAC reshuffling between chromosome 4 and 8 of *M. truncatula*. (a) Early pachytene chromosomes with probes AC144564 (orange signal), AC174339 (green signal) and AC137703 (red signal) located on the same chromosome (chr.4). (b) Prophase chromosomes with south telomere BAC AC162440 (red) that was genetically mapped on LG8 hybridized on chromosome 4 together with the other two BACs AC174339 (green) and AC137703 (orange). (c) Pachytene chromosomes with red color labeled BAC clone AC146585 (genetically on LG4) located on the same chromosome (8) as green labeled BAC clone AC137839, on the different chromosome as orange color labeled BAC clone AC137703. (d) Early pachytene chromosomes with south telomere BAC AC161034 (orange) that was genetically mapped on LG4 hybridized on chromosome 8 together with the other two BACs AC146585 (red) and AC171267 (green). (e) Schematic representation of reshuffling between chromosome 4 and 8. The positions of the BACs were drawn based on the genetic map. Chromosome 4, from BAC clone 146585 until the telomere end should be swapped with chromosome 8, from BAC clone 137703 until the telomere end.

Discussion

To sequence the *M. truncatula* genome, largely a BAC-by-BAC approach has been used. Several BAC libraries and one fosmid library were constructed using the enzyme combination of Hind III, BamH I, EcoR I and random shearing, resulting in 452,352 BAC clones. The *M. truncatula* genome project aims to sequence of the euchromatin of all chromosome arms, therefore, euchromatic BACs were identified by EST content, after which contigs were created using BAC end sequences. A tiling path was created from BACs with at least 2 kb overlap and maximal coverage and selected for sequencing. The sequenced BAC clones and all other relevant genomic sequences were aligned and assembled into contigs and then scaffolds. The final step of assembling was anchoring and ordering scaffolds into eight chromosome pseudomolecules by integrating genetic map information obtained using DNA-based genetic markers (Thoquet et al., 2002; Mun et al., 2006). Subsequently, gaps were filled using next-generation whole genome shotgun sequencing. For use as a reference species in biological studies, it is important that the sequence is as complete and accurate as possible. The published *M. truncatula* sequence spans a physical distance of 367 Mb of unique DNA sequence. Although the estimated *M. truncatula* genome size is larger (~500 Mb), the sequencing project aimed to cover just the euchromatic part of the genome, where the vast majority of genes are located. 62,388 genes have been annotated in the *M. truncatula* genome sequence and 94% of expressed genes are covered with currently available sequence (chapter 2). This means *M. truncatula* has a high gene density: 1.68 genes / 10 kb. This gene density is lower than *A. thaliana* (2.2), but higher than other sequenced plant genomes, for instance, rice (0.97), poplar (0.94) cucumber (0.73), sorghum (0.47), and soybean (0.42) (Yu et al., 2002; Tuskan et al., 2006; Huang et al., 2009b; Paterson et al., 2009; Schmutz et al., 2010).

Substantial gaps still remain within the assembly. Also, beyond the ends of the assembled chromosome arms and in areas of low recombination, additional euchromatic sequences could occur which are hard to map using genetic methods. For further improvement of the sequence it is important to know where these gaps are and how big they are approximately. We have shown that the using FISH on pachytene complements is a powerful tool to map the physical positions of BACs. The well-characterized pachytene chromosomes allow us to actually determine the euchromatin-heterochromatin borders, the length of individual chromosome arms, centromere and telomere positions. Thus we can estimate physical distances between BACs mapped by FISH, and estimate the lengths of segments missing in the assembly.

We measured the sizes of all remaining gaps within the sequence assembly of chromosome 5, which is the best sequenced chromosome. Each of the 8 gaps proved to be just a few kb, underlining that BAC-by-BAC-sequencing can produce very high-quality sequence. Other *M. truncatula* chromosomes have been sequenced less extensively, but still gap sizes on chromosomes 6 and 8 proved to be at mostly in the Mb range. This means that the published *M. truncatula* sequence is among the high-quality plant sequences determined so far, and thus provides a great resource for further studies in legumes.

An inconsistency was discovered between the previously published genetic map based on a cross between *M. truncatula* accessions Jemalong A17 and A20, and the assembled sequence of Jemalong A17. The genetic map links parts of the assembly of chromosome 4 with that of chromosome 8, and conversely. We have determined that the sequence assembly corresponds with the physical positions of BACs on pachytene chromosomes of the accession Jemalong A17. That leaves the question why the genetic map is inconsistent with the assembled sequence. Although the inconsistency could arise by chance from skewed recombination rates on chromosomes 4 and 8 in the population used for mapping, it is also possible that an actual reciprocal translocation occurred between accessions Jemalong A17 and A20. In that case, an attempt at genetic mapping would result in inconsistent segregation in the breakpoint region which could easily result in a genetic map that differs from the Jemalong A17 sequence. In chapter 5 of this thesis, we show how FISH experiments could be used to map such translocations.

In this chapter, we have shown that FISH experiments can verify and improve sequence assemblies by providing a direct and independent way to map sequences to chromosomes. While we used FISH to improve an already well-assembled genome sequence with relatively few gaps, the benefits of using FISH experiments to improve less-well assembled genomes, especially those deriving from modern massive-parallel sequencing methods generating short reads and assemblies with thousands of gaps, would be even greater. Especially when the need arises to work with sequences that are only partially present in an assembly or when the quality of assembly at a certain locus is doubtful, FISH can be used to resolve inconsistencies and prevent drawing incorrect conclusions. Thus FISH technique becomes ever more important in the genomic era.

Materials and methods

Plant materials

Plants of *M. truncatula* genotype Jemalong A17 were grown under green house conditions. The young flower buds were collected in the morning and fixed in fixative solution: acetic acid-ethanol (1:3), for three hours. During these three hours, the fixative solution was changed a few times until the solution remained clear. The flower buds were left in the fixative solution at -20°C for one month, and then transferred to a 70% ethanol solution and left at -20°C until use.

Pachytene Slide preparation

Pachytene slides were prepared following the method described previously by Olga Kulikova (Kulikova et al. 2001), except that the flower buds were digested in an enzyme mixture consisting of 1% cytohelicase (Bio Sepra 249701), 1% pectolyase Y-23 (Sigma P-3026), and 1% cellulose RS (Yakult 203027) in 10mM sodium citrate buffer at pH4.5 for 3 hours.

BAC clone isolation and labeling

BAC DNA was isolated either according to the alkaline lysate method (Woo et al. 1994) or PureLink ™ HiPure Plasmid DNA purification kits for Midi preparation of Plasmid DNA (Invitrogen K2100-04). For indirect labelling, the isolated BAC DNA was labelled with either Biotin-Nick translation Mix or Dig-Nick translation Mix (Roche). The Biotin-labelled BAC probes were detected by Avidin-Texas Red and amplified by biotin-conjugated goat-anti-Avidin and Avidin-Texas Red (Roche). The Dig-labelled BAC probes were detected by sheep-antidigoxigenin-fluorescein (FITC) and amplified by rabbit-anti sheep-FITC (Roche). For direct labelling, the isolated BAC DNA was labelled with Cy3.5-dCTP (GE Healthcare), Fluorescein-12-dUTP (PerkinElmer) or CyTM3-dUTP (Amersham Biosciences), without further detection and amplification steps.

Fluorescence *in situ* hybridization (FISH)

FISH was performed according to the protocol previously described by Olga Kulikova with few adaptions (Kulikova et al., 2001). Cot-100 DNA was used as a competitor for blocking the repetitive sequences that appeared in the BAC clones. Cot-100 was prepared according to the protocol described by Michael S. Zwick. 2 ug Cot-100 DNA is sufficient to block 200 ng BAC DNA. For the directly labelled probes, the detection and amplification steps were omitted. After hybridization the slides were rinsed in 50% formamide / 2xSSC three times, 5 minutes each. The slides were washed three times in 70%, 90% and 100% ethanol, respectively, and were air dried afterwards. The images were captured using a Photometrics Sensys 1305 x 1024 CCD camera, and further improvement of the selected images was done in Adobe Photoshop.

Imaging and measurements

All the images used for the measurement were captured using a 100x objective and 1x optovar on a Zeiss Axioplan 2 Imaging Photomicroscope equipped with epifluorescence illumination. Chromosome lengths and gap sizes were measured in images of at least 5 different pachytene spreads, using the software package Image-Pro Plus.

CHAPTER 4

Integration genetic and cytogenetic maps of red clover

Chunting Lang[1], Leif Skøt[2], Charlotte Jones[2], Dave Kudrna[3], Michael Abberton[2], Hans de Jong[4], Ton Bisseling[1], René Geurts[1]

1. Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands
2. Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Gogerddan, Aberystwyth, Ceredigion SY23 3EB, Wales, UK
3. Arizona Genomics Institute, 1657 E Helen Street, Keating BLD, The University of Arizona, Tucson, AZ 85721, USA
4. Laboratory of Genetics, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, Radix West, Building 107, 6708 PB, Wageningen, the Netherlands

Abstract

A molecular cytogenetic map of the legume forage crop red clover (*Trifolium pratense L*.) (2n = 2x = 14) was constructed based on a DAPI stained pachytene karyogram. We described a pachytene karyogram in which all 7 chromosomes can be identified based on chromosome length, centromere index, diagnostic heterochromatin structures, and the positions of rDNA loci and telomeric tandem repeat sequences. Strikingly, a consistent divergence in length of the telomeric repeat unit between both chromosomal ends is observed in all 7 chromosomes. This difference in length of telomeric repeats can be used as a diagnostic tool to orient chromosomes, which is especially helpful for those chromosomes with a median centromere position. Furthermore, the correlation between genetic linkage groups and the individual chromosomes is determined using FISH mapping of bacterial artificial chromosome (BAC) clones with a known genetic map position. This integrated genetic and cytogenetic map can be used in comparative genomic studies using the legume *Medicago truncatula* as reference species.

Introduction

Red clover (*Trifolium pratense L.*) is a cultivated legume species that has a long history as grassland forage crop for livestock production. Also it is used as organic manure, mainly because of high nitrogen content due to its capability to live in symbiosis with nitrogen fixing rhizobium bacteria (Kolliker et al., 2003). As a biological nitrogen source, it is important for sustainable agriculture (Taylor and Quesenberry, 1996; Singh et al., 2007; Tejada et al., 2008). Furthermore, red clover is rich in secondary metabolites including isoflavones and phytoestrogens, and derived dietary supplements receive increasing attention (Beck et al., 2005; Saviranta et al., 2008; Taponen et al., 2010). All these properties make red clover an attractive crop for agricultural production and breeding, which justifies comprehensive – omics-based investments. Here we present a high resolution karyotype based pachytene bivalents that has been integrated with the red clover genetic map.

The *Trifolium* genus is among the largest genera in the Legume family encompassing over 230 species divided over two subgenera; *Chronosemium* and *Trifolium*, respectively (Gillett and Taylor, 2001; Lewis et al., 2005; Ellison et al., 2006). All species are herbaceous and 16 are cultivated as forage crops, including red clover (Gillett and Taylor, 2001). *Trifolium* is part of the *Trifolieae* tribe, which includes among others also the genus *Medicago* (Lavin et al., 2005). This suggests that species of both genera will display to a certain extent a conserved genome organization, including relative high levels of synteny.

For white clover and red clover, genetic maps have been constructed based on restriction fragment length polymorphism (RFLP), simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) markers (Isobe et al., 2003; Sato et al., 2005; Isobe et al., 2009). These genetic maps, which are anchored to a red clover physical map, facilitate comparative studies with model species. In legumes, two species are generally used as models, *Lotus japonicus* and *Medicago truncatula*, and for both species substantial genome sequence information has been generated (Young et al., 2005; Sato et al., 2008).

Furthermore, the genome sequence of soybean (*Glycine max*) has been determined (Schmutz et al., 2010). Of these three legume species with sequenced genomes, *M. truncatula* is phylogenetically closest to red clover. Both species belong to the *Trifolieae* tribe, of which the origin is dated to 24.7+/-2.3 million years ago (Lavin et al., 2005). In comparison, the last common ancestor with *L. japonicus* and soybean is estimated to have lived ~50 and ~54 million years ago (Lavin et al., 2005). This suggests that the *M. truncatula* genome would be most suited as reference for red clover.

Red clover has a relatively small diploid genome (430-470 Mb) divided over 7 chromosomes (2n = 2x = 14) (Sato et al., 2005). It is a cross-pollinated species with a self-incompatibility system (Townsend and Taylor, 1985). Therefore, red clover cultivars are highly heterogeneous and individual plants have high levels of heterozygosity, resulting in large genetic variation within and between red clover populations (Brook, 1991; Kongkiatngam et al., 1995). This provides genetic variation for breeding purposes, but also complicates genetic and genomic research because inbred lines cannot be developed. The usual strategy for the construction of genetic maps, starting with one mapping population obtained from two inbred parental lines, can therefore not be used in red clover research due to its allogamous nature. Instead, independent genetic maps of red clover were developed on two heterogeneous lines, and the two linkage maps were integrated using markers that showed heterozygosity in both lines (Isobe et al., 2003). All markers that were developed from cDNA are relatively conserved in sequence, which facilitates re-use of markers in other germplasms and subsequent comparative studies.

The first cytogenetic map for red clover is based on metaphase chromosomes. It shows that chromosome (Chr.) 1 corresponds to genetic linkage group (LG) 5; Chr.2 - LG2; Chr.3 - LG7; Chr.4 - LG1; Chr.5 - LG3; Chr.6 - LG6 and Chr.7 - LG4 (Sato et al., 2005). However, red clover metaphase chromosomes are very small, ranging in size between 1.9 - 2.9 µm (Kazimierski et al., 1972). Such small chromosomes cannot be observed in much detail using optical microscopes and thus most morphological structures in the chromosomes remain unclear. Chromosome bivalents at pachytene stage, as found during meiosis I, are generally 10-40 times longer chromosome length comparing with mitotic metaphase (de Jong et al., 1999). A karyotype based on pachytene chromosomes will show more detail when compared to metaphase karyotypes, and therefore is a desirable resource for studies in red clover.

In this study we describe a red clover pachytene karyotype that includes chromosome length (euchromatin and pericentromeric heterochromatin), centromere position, the position of 45S and 5S rDNA loci, and molecular organization of telomere repeats. Furthermore, the correlation between genetic linkage groups, physical map and the individual pachytene chromosomes is determined using fluorescent *in situ* hybridization (FISH) mapping of BAC clones with a known genetic map position. The resulting integration of genetic map and pachytene based cytogenetic map of red clover can be used in future sequencing projects, and as a reference for comparative genomics with *M. truncatula* or other legume species.

Results

Pachytene karyotype of *Trifolium pratense*

We selected red clover variety Milvus for karyotyping as this variety was used as reference for structural genomics. This variety Milvus was also used in a cross with variety Britta to construct a genetic map (Skøt et al., in preparation) (Winters et al., 2009).

To construct a pachytene karyotype we studied chromosomes in meiosis I pachytene stage from pollen mother cells (Fig. 4.1a). Five well spread pachytene preparations were selected in which all 7 pachytene bivalents could be traced. These were further characterized based on the length of chromosome arms, size of pericentromeric heterochromatin, centromere index (short arm / total individual chromosome length), and the length of euchromatic region (Table 4.1). The total length of the pachytene complement was 446.5 ± 21.9 µm (n = 5), and the variation in length between different spreads was less than 3% (Table 4.1). Below we described how we determined the correlation between pachytene chromosomes and genetic linkage groups. However, from here on, we will use correct chromosome numbering. By straightening the pachytene bivalents the variation in chromosome arm length and morphology was visualized (Fig. 4.1b). The lengths of individual chromosomes vary from 53.5 ± 4.0 μm for the shortest chromosome (#6) to 76.6 ± 4.3 μm for the longest chromosome (#2) (Table 4.1). In comparison, mitotic metaphase chromosomes range from 2.1 - 2.9 μm in size (Kazimierski et al., 1972), indicating a 25 - 30x increasing in resolution when using pachytene bivalents.

Within a pachytene cell complement, the two longest chromosomes (#2 and #7) could be discriminated on the basis of a difference in size of the pericentromeric heterochromatin. Chromosome 2 has clear pericentromeric heterochromatin (7.2 μm), such region is lacking in chromosome 7 (Fig. 4.1a). Four medium-sized chromosomes vary in length between 56 - 67 μm (# 1, 3, 4 & 5) (Table 4.1). Chromosome 1 can be easily distinguished from the other chromosomes because of its highly condensed pericentromeric heterochromatin (18 μm) (Fig. 4.1a). It comprises ~33.1% of the length of chromosome 1 and ~4.1% of the total length of all chromosome arms (Table 4.1). The remaining three medium-sized chromosomes (#3, 4 & 5) have a similar sized pericentromeric heterochromatin region varying in length between 7 - 8 μm. They were discriminated on basis of their centromere index (Table 4.1); chromosome 3 with the smallest centromere index (28) to the chromosome 5 with the largest centromere index (43) (Table 4.1).

Next, we compared the pachytene karyotype of Milvus to that of Britta. Highly condensed pericentromeric heterochromatin was observed in chromosomes 1, 3, 4 and 5, whereas remaining chromosomes had smaller blocks of pericentromeric heterochromatin (Fig. 4.1c). The less stained centromere is most easily observed in chromosome 1 due to its largest pericentromeric heterochromatin region. Compared with Milvus, no obvious morphological difference was observed in Britta pachytene karyotype (Fig. 4.1c).

In summary, the seven pachytene chromosomes of red clover can be distinguished based on chromosome length, centromere index, and size of pericentromeric heterochromatin and euchromatic region.

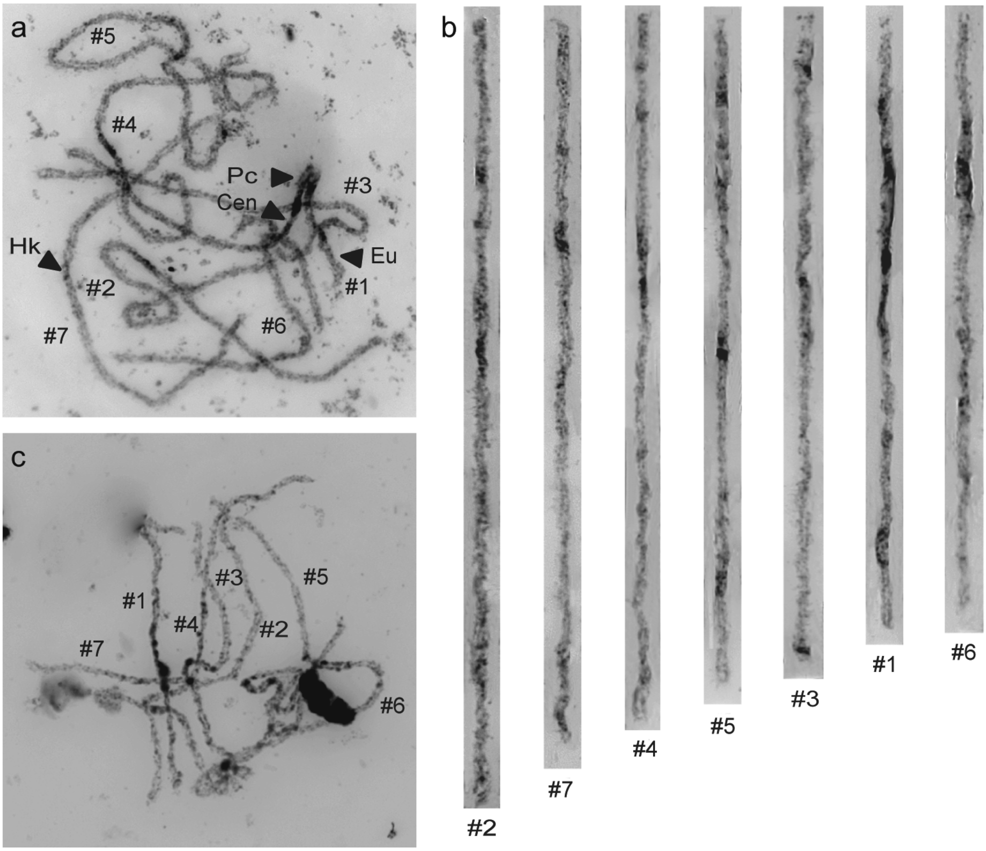


Figure 4.1: Pachytene chromosome morphology of red clover varieties Milvus and Britta. (a) DAPI-stained pachytene chromosomes of red clover variety Milvus. The image shows darkly stained pericentromeric heterochromatin (Pc), less darkly stained euchromatin (Eu) and less stained centromere (Cen) inside the pericentromeric heterochromatin. Heterochromatic knobs (Hk) are visible in the euchromatin of all chromosomes. The chromosome numbers are indicated. (b) Pachytene chromosomes of variety Milvus straightened using Image J software. (c) DAPI-stained pachytene chromosomes of variety Britta. The chromosome numbers are indicated.

Parameter	**Chromosome**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Average length [a]	55.5 ± 3.5	76.6 ± 4.3	59.9 ± 3.6	66.6 ± 4.0	63.4 ± 3.6	53.5 ± 4.0	71.0 ± 3.8	446.5 ± 21.9
Eu-short arm [b]	9.9 ± 1.3	27.1 ± 3.4	15.3 ± 4.7	22.4 ± 2.5	24.4 ± 3.0	16.3 ± 1.2	31.5 ± 0.7	142 ± 4.9
Eu-long arm [c]	27.5 ± 2.8	42.1 ± 2.6	37.1 ± 2.8	36.9 ± 2.5	31.4 ± 2.2	33.5 ± 3.0	35.3 ± 2.6	242 ± 11.4
Total EU [d]	37.4 ± 1.7	69.2 ± 3.1	52.4 ± 3.3	59.3 ± 3.8	55.8 ± 3.7	49.8 ± 3.8	66.8 ± 2.9	384 ± 8.2
% Hetero-chromatin [e]	33.1	9.4	12.8	10.8	11.1	7.1	5.6	12.4
Centromere index [f]	41	37	28	38	43	34	47	100
Total cell complement [g]	12.4	17.1	13.4	14.9	14.2	12.0	15.9	
45S rDNA [h]	S	-	-	-	-	S	-	
5S rDNA [i]	S	S	-	-	-	-	-	

Table 4.1: The characterization of individual pachytene chromosomes; variety Milvus. (a) The average length of individual pachytene chromosomes in µm ± SD. (b) The length of short arm of short arm euchromatin in µm ± SD. (c) The length of long arm euchromatin in µm ± SD. (d) The length of total euchromatin of individual chromosomes in µm ± SD. (e) The percentage of heterochromatin in cell complement. (f) Centromere index is percentage of short arm / average length of corresponding individual chromosome length. (g) Total cell complement is percentage of individual chromosome length / total length of all chromosomes. (h, i) Number of 45S rDNA and 5S rDNA loci and their physical position on the pachytene chromosomes, short arm (S).

Ribosome genes as diagnostic tool for chromosome identification

Ribosome genes are commonly used as a diagnostic tool for identification of individual chromosomes in a cell complement due to their different nuclear organization and loci number. We conducted FISH using 45S and 5S rDNA probes to determine whether it can be used as additional diagnostic tool to support chromosome identification in red clover. We used Milvus and Britta for this study. Two-color FISH experiment was applied on pachytene chromosomes using digoxigenin-dUTP labeled 45S rDNA (green) and biotin-dUTP labeled 5S rDNA (red) as probes. Two loci were found for both 45S rDNA and 5S rDNA (Fig. 4.2a, b). A strong 45S rDNA hybridization signal was present within pericentromeric heterochromatin region of chromosome 1, whereas a second 45S rDNA locus was positioned on the short arm of chromosome 6, also within the pericentromeric heterochromatin (Fig. 4.2a). 5S rDNA was localized in the pericentromeric region chromosome 1 in close vicinity of the 45S rDNA and a second smaller 5S rDNA locus was present on the short arm of chromosome 2, close to the border the heterochromatin and the euchromatin (Fig. 4.2a). As the rDNA loci are conserved in Milvus and Britta, chromosome identification can be simplified by using ribosome genes as diagnostic tool.

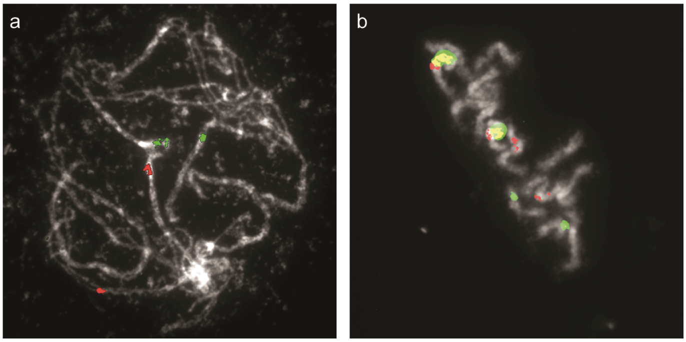


Figure 4.2: FISH visualizing the loci of 45S rDNA (green) and 5S rDNA (red) on red clover chromosomes. (a) Two loci for both 45S and 5S rDNA were observed on the variety Milvus. Two loci of 45S rDNA were positioned on pachytene chromosome 1 and 6. Two loci of 5S rDNA were localized on pachytene chromosome 1 and 2. (b) Two loci for both 45S and 5S rDNA were observed on variety Britta. The same localizations were observed on Britta as on Milvus.

Size divergence of telomeric repeat locus as diagnostic tool

The telomere repeated sequence is highly conserved among most of higher plant species, though the length of the repetitive block can vary. To characterize the telomere loci in red clover, we performed a FISH experiment by using a consensus telomeric repeat sequence as a probe. The *A. thaliana* telomeric clone pAtT4 containing the tandem repeated sequence 5'-CCCTAAA-3' was hybridized to the chromosomes of both varieties, Milvus and Britta. Bright fluorescent signals visualized seven loci, each on a distal end of a chromosome (Fig. 4.3a, b). Strikingly, the second telomeric signal of each chromosome was very weak, and difficult to detect. These observations hold for both varieties (Fig. 4.3a, b). This striking deviation in size of the repetitive telomeric locus for each individual chromosome provides a diagnostic tool for chromosome orientation, which especially can help to orient meta-centromeric chromosomes (#1, #5 and #7) (Fig. 4.3b). For instance, chromosome 1 has a telomeric signal on its short arm but not on its long arm, whereas chromosome 5 has a signal on its long arm only (Fig 4.3b).

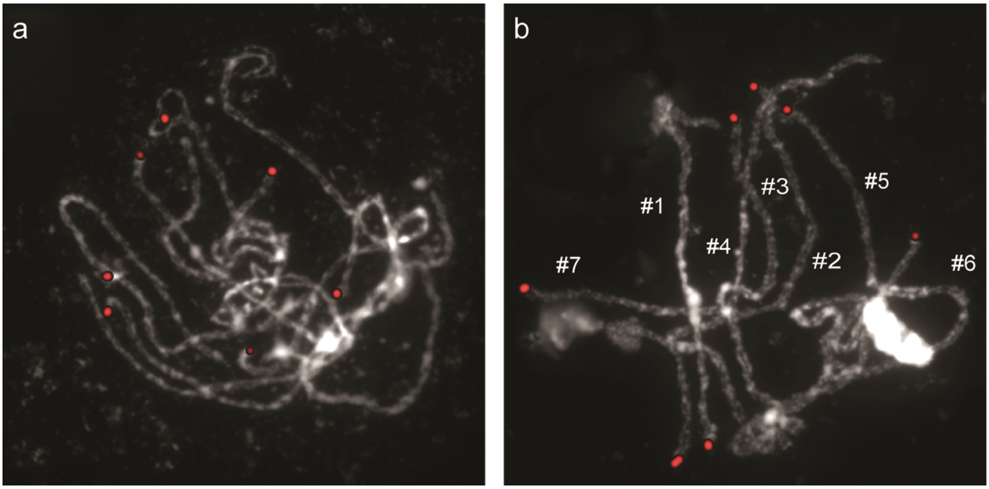


Figure 4.3: FISH visualizing telomere repeats on red clover pachytene chromosomes. (a) Seven loci were observed on variety Milvus, each locus on a distal end of individual chromosome. (b) Seven loci were visualized on variety Britta, each locus was hybridized on one end of each chromosome.

Integration genetic and cytogenetic maps of red clover

To integrate the pachytene based karyotype with the genetic linkage map of red clover and to display physical positions of molecular markers on chromosomes relative to the structural markers, molecular genetic markers of each linkage group were selected (Table 4.2) (Sato et al., 2005), (Skøt et al., in preparation). These markers were either simple sequence repeats (SSR) or single nucleotide polymorphisms identified in expressed sequence tags (ESTs) that were subsequently used to identify corresponding BAC clones. Of each linkage group two to five markers were converted into FISH probes. To prevent binding of the putative repetitive sequences in the BACs to the chromosomes, Cot-100 fraction was used to block the repetitive sequences in all BAC clones during FISH experiments (see material and methods; Fig. 4.4h). In this way the selected probes visualized only a single locus (Fig. 4.4a-g). Their genetic position along with the corresponding positions on pachytene chromosomes were subsequently integrated (Fig. 4.5). In the way, chromosomes 1, 2, 3, 4, 5, 6 and 7 could be linked to linkage groups LG5, LG2, LG7, LG1, LG3, LG6 and LG4, respectively (Table 4.2, Fig. 4.4i). All markers are located in euchromatic regions of the chromosomes, which is in line with the genic nature of the original genetic markers.

To summarize, the molecular markers selected from each linkage group allowed us to identify the corresponding pachytene bivalents. Subsequently, the genetic position of mapped molecular markers could be correlated to the physical position on pachytene chromosomes (Fig. 4.5).

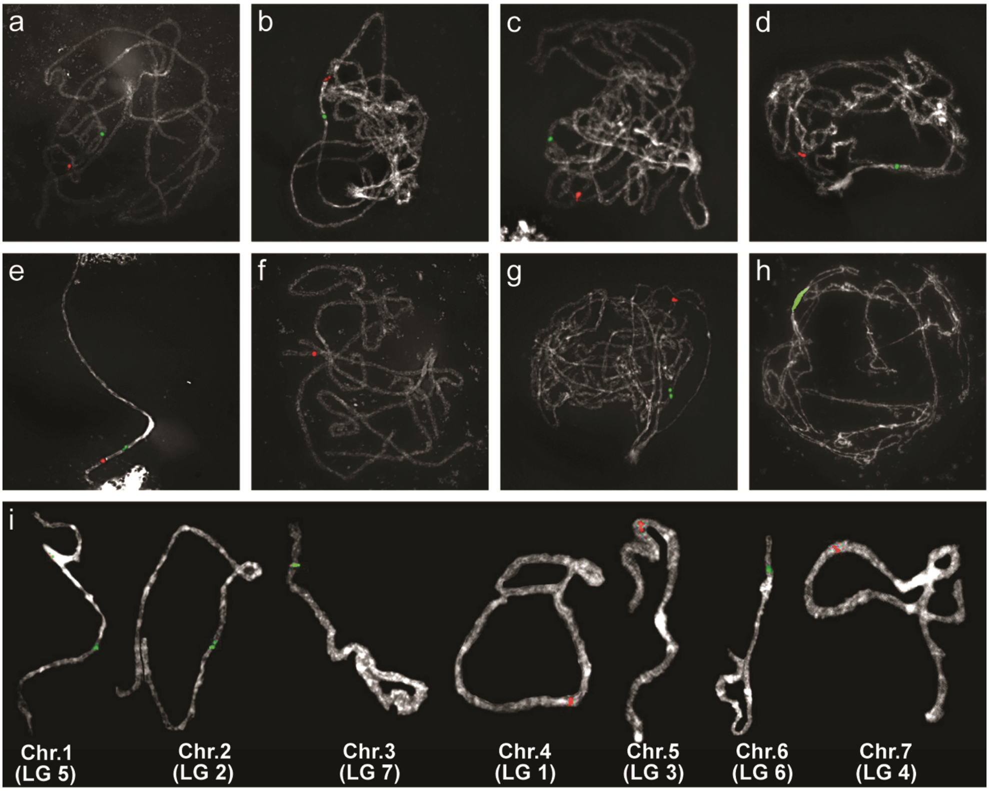


Figure 4.4: Integration of cytogenetic map and genetic linkage groups. (a) BACs of linkage group 1 mapped to chromosome 4: TA3816 (green) and RCS3555 (red). (b) BACs of linkage group 2 mapped to chromosome 2: TA11 (green) and Tp_GeM57 (red). (c) BACs of linkage group 3 mapped to chromosome 5: TA2729 (green) and TA1580 (red). (d) BACs of linkage group 4 mapped to chromosome 7: RCS0714 (green) and Tp_GeM04 (red). (e) BACs of linkage group 5 mapped to chromosome 1: RCS5643 (green) and RCS1762 (red). (f) BACs of linkage group 6 mapped to chromosome 6: RCS0069 (red). (g) BACs of linkage group 7 mapped to chromosome 3: TP_GeM35 (green) and RCS5570 (red). (h) FISH mapping of red clover Cot-100, used as blocking agent in all FISH experiments. (i) Assignment of linkage groups to pachytene chromosomes using the marker-based BAC clones. Chromosomes were digitally isolated from pachytene complements.

LG	**Genetic marker**	**BAC clone**	**Forward primer (5'-3')**	**Reverse primer (5'-3')**	**Genetic distance (cM)**	**Chr**
1	BB914499	c0031L22	AATGCCCTGCAATCTCTGAC	ATTAGGGCAACCCTTTCCAA	1.1	4
	TpGeM20	c0041D19	GAACACTTCAACCGCCAAAT	CCCCAAAACCCTAAGCTACC	42.3	
	BB916296	c0036M02	TTTGGTGCGTTGTGGAGTAG	GCATCAAAGCCAGGAAATGT	46.9	
	TA3816	c0031J19	TTCAACAACTCCCCTTTTCAA	GGAAACAACGAACTCCCAGA	50.3	
	RCS3555	c0024E03	GCAAAAGATCCCGTCACAGT	CTCAGTAGCAGCACCACCAA	85.6	
2	BB913735	c0044M18	TTGCAGAACCTTGCCTCTTT	GGAAAAGAACCCTTAATGGAAGA	31.4	2
	TA11	c0013E16	CCTCCCGTTGTACCAATAACA	TCATGGCCTTAATTTCTGGTG	63.5	
	TpGeM57	c0047E24	GGAAAGTAAGCCTGCACCTG	TGCATTTTTCTCCTGCCTCT	68.7	
	BB916818	c0015A04	GGAAACGCTAATCCAGGAGA	AGGCAACTCATGACGACAAA	75.3	
	RCS1864	c0026O21	AACCCAACAACACCAACACA	CGTTTCAGAAGTGGCACTGA	84.6	
3	TA1580	c0046H02	GTTGGAAAGGGTTGTGCTGT	AGCTGAGCAATGACCTGGTT	6.9	5
	TA3953	c0007P09	GGTGTTCCTTTTCAGCAAGG	TGGCTCAACAGGGTATTTCC	30.3	
	RCS1587	c0022A24	TTCACACCAATTCCTCCTCC	TTCCAACCAAAAACTCCGAC	51.2	
	TA2729	c0046H02	GCACCAAAAATTTCGAAAGAA	AAAAGGGTGGTCTTGAAACTGA	70.1	
4	Tp_GeM19	208C09	ACTAACGAAAACGCGACACC	AAGTGCTCATCCCCAATCAC	0	7
	RCS0714	c0037M21	AGGGTTTGGAATGTGTTGGT	TGGTTCAAGCTGTACAAAAGGA	6.7	
	RCS5390	c0017F18	CCACACCCTCTCTTCAATCA	ATTTGATCCAATCCAGCGAC	39.7	
	TA1636	c0022B11	CCACCTCGTCTTTCATCCAT	TCCTAGGCGACAGAAAATCG	73	
	Tp_GeM04	b0009D11	TTGTGCAAAACAGGATTGCT	GGTCTTTGCCAAAAGATCCA	102.5	
5	RCS1762	c0019N11	AAATGGCGCAAGAGAACAAT	ACCAACCCAAGCAGATGAAG	0	1
	RCS0131	c0013D10	ACGTGACGGA	AACCCTTCAAACC	25.6	

			GAGAGCTACG	CAAAACC		
	RCS5643	c0009E14	TGGGAATCGCT TAGTATCGG	TGTGTTTGGACTT CTTCAGGC	36.9	
	TA989	c0044G01	TCTGTGGCATC GAAATCAAA	CAGTGAAGGCTCC AATGTCA	77.3	
6	Tp_GeM09	c0005F02	ACGTTTCCAAT GCGAAAAAC	GTGCTGCTGGTGT CATGTTC	56.7	6
	RCS0069	c0023O07	ATTGCAAACCG AACCTGAAC	TACAATCCCTCGG TGCATTT	61.8	
7	RCS2866	c0004B16	CGGTTTGAATT TGAACATGG	TATGAAGGTTTAG GCGTGGC	11.6	3
	Tp_GeM35	0036M11	TGATTGGTGTT CTTGGTGGA	CTTGCATTCAAAG GGAGTGA	24	
	RCS5570	c0026H11	AATCCCCAAAA GCCATATCC	GGAAGATTGAGGT GGTCCAA	unknown	

Table 4.2: Red clover molecular markers distribution and characterization

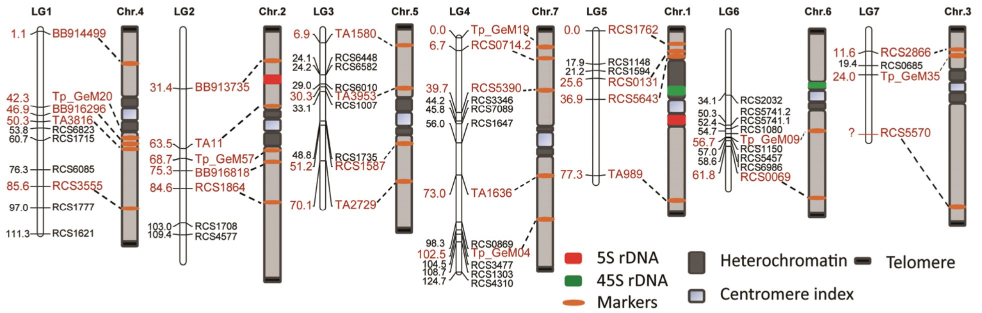


Figure 4.5: Integration of cytogenetic karyotype and genetic linkage groups. The molecular markers are mapped based on their genetic positions. BAC clones containing these markers are mapped to chromosomes based on FISH result.

Discussion

We constructed the first karyotype of red clover based on DAPI-stained pachytene chromosomes that enables differentiation between euchromatic and heterochromatic regions. Red clover contains 4 chromosomes (#1, 3, 4 and 5) with a large pericentromeric heterochromatin blocks. In case of chromosome 1, this block is around three times larger when compared to the other 3 chromosomes and this region encompasses the nuclear organizing region (NOR). In case of the remaining 3 chromosomes (# 2, 6 and 7) pericentromeric heterochromatin is not visible suggesting that these chromosomes have rather limited amounts of pericentromeric repeat sequences. In comparison, *M. truncatula* has large pericentromeric heterochromatic blocks that are equally distributed over all chromosomes (Kulikova et al., 2001). Since the difference in pericentromeric heterochromatin between red clover and *M. truncatula* (12.4% of the total complement for red clover versus 14.6% for *M. truncatula*) correlates with the size of their genomes, ~450 versus 560 Mb, respectively, it suggests that the difference in genome size is caused in part by clustered pericentromeric repetitive sequences. To suppress the amount of hybridization signal caused by repetitive sequences present in the BAC probes of red clover, a blocking treatment with Cot-100 fraction of genomic DNA needed to be applied. No such blocking step in FISH experiments used in *M.truncatula* (Kulikova et al., 2001). This suggests that in *M. truncatula* pericentromeric repeats are less dispersed and are less present in euchromatic regions. This is consistent with the observation that the euchromatic arms of red clover are rich in small heterochromatin knobs.

Red clover pachytene bivalents are 25 - 30 times larger when compared to the chromosomes in metaphase stage. This enables us to use morphological criteria like the distinction between euchromatin and heterochromatin, length of chromosome arms and centromere position to identify each individual pachytene chromosome. However, identification of individual chromosome can be simplified by using diagnostic probes, like 5S rDNA and 45S rDNA. 45S rDNA sequences are organized in tandem arrays within the nucleolar organizing region (NOR). In red clover two such loci are observed in the pericentromeric region of chromosomes 1 and 6. Both loci significantly differ in size. While on chromosome 6 the 45S rDNA locus is relatively small, the locus on chromosome 1 encompasses over half of the pericentromeric region, which means the two loci can be distinguished visually. Similar studies were done on a different clover species, which showed significant variation in 45S rDNA loci compared to red clover (Ansari et al., 1999). Variation in rDNA loci between closely related species, or even ecotypes have been observed in many plant species; e.g. Petunia, cereals, and Populus (Dubcovsky and Dvorak, 1995; Badaeva et al., 1996; Prado et al., 1996). In this respect it is noteworthy that the red clover varieties Milvus and Britta do not display a difference in rDNA loci.

The telomere repeats in plants were first characterized in *A. thaliana*, in which the consensus sequence is TTTAGGG. This sequence occurs in the majority of plant species, including the legumes *M. truncatula*, *L. japonicas* and soybean (Richards and Ausubel, 1988; Ganal et al., 1991; Burr et al., 1992; Gardiner et al., 1996; Zellinger and Riha, 2007; Schmutz et al., 2010). In this study, we provide evidence that red clover also has *Arabidopsis*-like telomere repeats. However, a remarkably, though consistent, variation in length of the telomeric repeat units between both ends of individual chromosomes is found. In different *Arabidopsis* ecotypes telomere length varied from 2.5 kb to 8 kb (Maillet et al., 2006), whereas in different maize inbred lines the range of telomere length is even broader, from 2.8 kb up to 40 kb (Burr et al., 1992). Also variation of telomere length among chromosomes within a karyotype was observed in rice, where FISH mapping of the telomeric repeat showed that the signal on chromosome 12 was more intense than on chromosome 6 due to a difference in copy number (Ohmido et al., 2001). In our study on red clover, telomeric repeat hybridization signals were observed only on one end of each chromosome. Theoretically, we cannot exclude that some chromosome ends may have different telomere repeats. Though as the 5′-CCCTAAA-3′ telomeric repeat is highly conserved, we anticipate this is very unlikely. Therefore, the difference in telomeric repeat loci in red clover could be caused by a difference in copy number of telomeric repeat units between both chromosome ends. The hybridization techniques we applied are not sensitive enough to detect low copy numbers of telomeric repeats.

We integrated the genetic map and cytogenetic maps of red clover using BAC clones that gave unique signal on the pachytene chromosomes with Cot-100 blocking DNA. This the fourth cytogenetic map of a legume species after *M. truncatula* (Kulikova et al., 2001), *L. japonica* (Pedrosa et al., 2002), and common bean (Pedrosa et al., 2003), and the first pachytene based cytogenetic map for the *Trifolium* genus. The integrated maps of red clover can be used in physical genomic studies as well as reference for comparative studies to legume model species like *M. truncatula*. These resources will be useful in future studies on the evolution of *Trifolium* species and for comparison of the red clover genome to the genomes of other legumes.

Materials and methods

Plant material

Plants of two red clover cultivars Milvus and Britta were grown under green house condition, which is 12 hours daylight and around 20-22°C. The flower buds were collected between 9 am to 11 am and fixed in acetic acid-ethanol (1:3). The fixative was replaced a few times until the solution remained clear. The flower buds were left in the fixative solution at 20°C for one month, and then transferred into 70% ethanol at -20°C until use.

Slide preparation of pachytene chromosomes

The method for slide preparation of red clover pachytene chromosomes was adapted from the *Medicago truncatula* slide preparation protocol described by Olga Kulikova (Kulikova et al., 2001). Flower buds of 2 mm in length were rinsed with Milli-Q water twice and citrate buffer (10 mM sodium citrate, pH 4.5 with citric acid) one time, and then incubated in 1% enzyme mixture-citrate buffer (1:2) at 37°C in a water-saturated atmosphere for 1 hour and 15 minutes for cell wall digestion. The enzyme mixture included 1% cytohelicase (Bio Sepra 249701), 1% pectolyase Y-23 (Sigma P-3026), and 1% cellulose RS (Yakult 203027), in citrate buffer. After the enzyme treatment, the soft flower buds were rinsed with sterile MQ water 3 times. One flower bud is sufficient for one slide preparation. The anthers were dissected from a flower bud and transferred to the middle of a cleaned grease-free slide with 2 µl MQ water, and then carefully dissected and homogenized the anthers with needle. 80 µl of 60% acetic acid was added, mixed well, the cells were spread on the slide. The slide was baked at 50°C hot plate for 2 minutes to remove the cytoplasm. The slide was taken off and ice-cold acetic acid: ethanol (1:3) was added in a circle around the suspension. The slide was left on the hot plate for drying. All slides were screened using phase contrast microscope for chromosome spreading and presence of pachytene stage nuclei. Selected slides were stored in a dust free microscopic box until use.

Probe isolation and labeling

Clone pTA71, which contains a 9.1 kb fragment of 45S rDNA of wheat (Gerlach and Bedbrook, 1979), clone pCT4.2, which contains a 5S rDNA of *Arabidopsis thaliana* (Campell et al., 1992), and clone pAtT4, which contains a telomeric repeat 5′-CCCTAAA-3′ of *Arabidopsis thaliana* (Richards and Ausubel, 1988) were isolated according to Easy Nucleic Acid Isolation Kit (Omgea Bio-tek) and labeled with either biotin-dUTP or digoxigenin-dUTP nick translation mix (Roche).

Three bacterial artificial chromosome (BAC) libraries were constructed with DNA from a genotype of the red clover variety Milvus as described previously (Farrar et al., 2007; Winters et al., 2009). The Tp_MBa library was made using genomic DNA partially cut with HindIII, and represented 6.5X coverage, while the Tp-ABa and Tp_ABb were made using EcoRI and BamHI, respectively. The latter two libraries were made and each represented 10X coverage (Ammiraju et al., 2006). BAC addresses of red clover molecular markers on the genetic map were identified using a PCR based approach (Farrar et al., 2007). Briefly, this involved a three-dimensional pooling strategy of the BAC library, and PCR amplification of these pools identified the individual plate(s) containing the marker in question. Amplification of the positive plates identified the individual clones. Their identities were confirmed by amplicon sequencing using an ABI 3730 Genetic Analyzer. The details of the BAC clones used in this work are shown in Table 2. BAC DNA was isolated according to the alkaline lysate method and labeled with either biotin-dUTP or digoxigenin-dUTP nick translation mix (Roche).

Genetic mapping

The genetic map was constructed from an F1 population of 188 genotypes which were progeny of a cross between a genotype of the variety Milvus with a genotype of the variety Britta. Both parents were highly heterozygous due to the allogamous nature of red clover. A total of 143 markers were mapped onto seven linkage groups. They were either microsatellite markers originally identified by Sato et al. (2005) or gene-based single nucleotide polymorphisms identified in this work by amplicon sequencing (Sato et al., 2005), and validated by progeny testing in this map. The JoinMap 4 programme was used to generate the linkage map with settings as described (Van Ooijen, 2006).

Cot-100 isolation and labeling

Red clover Cot-100 was prepared according to the protocol described by Michael S. Zwick with some modifications (Zwick et al., 1997). The total genomic DNA was isolated using the cetyltrimethylammonium bromide (CATB) method and sonicated to a fragment size of about 500 bp. The sheared genomic DNA was denatured at 95°C for 10 minutes, and then re-annealed at 65°C for 17 hours and 13 minutes in a rotating oven. The concentration of sheared genomic DNA was 0.5469 g/L, which was equivalent to 16.13 x 10-4 mol/L using an average molecular weight for a deoxynucleotide monophosphate of 339 g/mol (0.5469 g/L / 339 g/mol = 16.13 x 10-4 mol/L). To produce Cot-100, the re-annealing time was 100 / 16.13 x 10-4, which is 17 hours and 13 minutes. The reminding single strand DNA was removed using S1 endonuclease (Fermentas, final concentration 1 U/µg) at 37°C for 90 minutes. The reaction was stopped and extracted by adding equal volumes of chloroform-isoamylalcohol (24:1), mixed well and centrifuged at 4,000 rpm for 10 minutes. The upper layer was transferred into a new tube. DNA was precipitated with 2.5 volumes of ice cold 100% ethanol at -20°C, overnight, and then centrifuged at 4°C with 11,000 rpm for 30 minutes. DNA was air dried and dissolved in 20 µl HB50 (50% deionized formamide, 2 x SSC, 50 mM sodium phosphate, pH 7) and labeled with digoxigenin-dUTP nick translation mix (Roche).

Fluorescence *in situ* hybridization (FISH)

FISH was performed as described by Zhong et al. (1996), with an extra pachytene slide incubation step at 65°C for 30 minutes and without pepsin treatment. Cot-100 was added in a 50 to 1 ratio compared to labeled probe to block repetitive sequences. The hybridization mixture contained 20 ng labeled probes, 1 µg Cot-100, 10 µl 20% dextran sulfate in HB50 (50% deionized formamide, 2 x SSC, 50 mM 1M sodium phosphate pH7, 20% dextran sulfate), and was diluted with HB50 to a total volume of 20 µl. Probes labeled with digoxigenin-dUTP were detected with sheep-antidigoxigenin-fluorescein (FITC) and amplified with rabbit-anti-sheep-FITC, resulting in green signal. Probes labeled with biotin-dUTP were detected with streptavidin CY3 and amplified with streptavidin biotin, resulting in red signal. Chromosomes were counterstained with DAPI (4′, 6-diamidino-2-phenylindole) in Vectashield antifade solution (Vector Laboratories), 5µg/mL. Slides were checked under a Zeiss Axioplan 2 Imaging Photomicroscope equipped with epifluorescence illumination, using filter sets for DAPI, FITC, and Cy3 fluorescence. The images were captured by a Photometrics Sensys 1305 x 1024 pixel CCD camera and straighten in software Image J, and level improvement and sharpening of selected images were done in Adobe Photoshop.

CHAPTER 5

Red clover unique non-LTR retrotransposon (LINE) is associated with species specific chromosomal rearrangements

Chunting Lang[1], Leif Skøt[2], Dave Kudrna[3], Michael Abberton[2], Ton Bisseling[1], René Geurts[1]

1. Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands
2. Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Gogerddan, Aberystwyth, Ceredigion SY23 3EB, Wales, UK
3. Arizona Genomics Institute, 1657 E Helen Street, Keating BLD, The University of Arizona, Tucson, AZ 85721, USA

Abstract

The *Trifolium* genus is one of the top twenty largest genera within the legume family and contains several economically interesting clover species. Here we study the evolutionary history of the *Trifolium* genus using comparative chromosome painting with sequences from *M. truncatula* chromosome 5 as reference. We find that the species in the subgenus *Chronosemium* have high co-linearity with *M. truncatula*, but extensive chromosomal rearrangements are observed in the subgenus *Trifolium*, more specifically in the sections *Trichocephalum* and *Trifolium*. We also identified red clover-specific rearrangements. By zooming in on a chromosomal rearrangement breakpoint on the short arm of *M. truncatula* chromosome 5, we find that this breakpoint is associated with a red clover specific repetitive sequence. To identify this repetitive sequence, we sequenced BAC clones containing the region surrounding the breakpoint and find a red clover-specific LINE-like non-LTR retrotransposon at the breakpoint that we named Tp_LINE. Tp_LINE is highly repetitive in red clover, but absent in sister species *T. hirtum* and *T. diffusum*. These results are consistent with the hypothesis that there is a causal link between the presence of Tp_LINE and occurrence of chromosomal rearrangements in red clover.

Introduction

The legume family (*Fabaceae*), which contains 650 genera and ~ 18,000 species, is the third largest flowering plant family (Lewis et al., 2005). Many species of this family are important resources for human food and/or animal feed. Legumes have the unique capability to fix nitrogen in symbiosis with rhizobium bacteria, which is why they are often used as natural fertilizers. Natural fertilizers are particularly common in the *Trifolieae* tribe that includes alfalfa (*Medicago sativa*) and clover species; e.g. red clover (*Trifolium pratense*) and white clover (*Trifolium repens*). *Trifolieae* forms a morphologically distinctive tribe that encompasses 6 genera, together include ~485 species (Lewis et al., 2005). More than half of the species belong to the genus *Trifolium*, which is one of the top twenty largest legume genera. The *Trifolieae* tribe diverged 24.7 +/- 2.3 million years ago (mya) and encapsulate another important tribe namely *Fabeae* that harbors 3 important crops; namely pea (*Pisum sativum*), lentil (*Lens culinaris*) and faba bean (*Vicia faba*) (Lavin et al., 2005). Comparative genetic and genomic studies within the *Trifolieae* and *Fabeae* tribes as well with phylogenetically more distinct legume species revealed a high degree of co-linearity between species (Choi et al., 2004b; Choi et al., 2004a; Kalo et al., 2004; Zhang et al., 2007; George et al., 2008; Hand et al., 2010; Nayak et al., 2010). However, for one species, namely red clover, more extensive rearrangements have been found when compared with either *Medicago truncatula* or *Lotus japonicus* (Sato et al., 2005). We investigated the occurrence of chromosomal rearrangements in red clover through a comparison with legume model species *M. truncatula*. Here we show that red clover contains a unique LINE-like non-LTR retrotransposon, Tp_LINE, which is not present in related *Trifolium* species. By zooming in on a single chromosomal breakpoint we find it associated with Tp_LINE.

Red clover is a commercially important crop grown as pasture and natural fertilizer. It has a relatively small diploid genome (430-470 Mb) divided over 7 chromosomes (2n = 2x = 14) (Arumuganathan and Earle, 1991). Thereby it differs from the ancestral chromosome number of 8 (2n = 2x = 16) found in 80% of the *Trifolium* species, including the closest sister species *Trifolium pallidum*, *Trifolium diffusum*, and

Trifolium andricum (Ellison et al., 2006). Only 20% of the other species in the *Trifolium* genus display a deviating chromosome number, varying from n = 5 to 24 (Ellison et al., 2006). Similarly, most species belonging to the genus *Medicago* also have 8 as base chromosome number; including *M. truncatula*, a legume model species (2n = 16). The *Medicago* and *Trifolium* genera have diverged ~24 mya. As the genome sequence of *M. truncatula* has been unraveled, it can be used as reference in comparative studies on closely related species.

Comparative genetic and genomic studies between *M. truncatula*, soybean (*Glycine max*), pea, common bean (*Phaseolus vulgaris*), Alfalfa, white clover, *L. japonicus* and chickpea (*Cicer arietinum*) demonstrated a high co-linearity among these species (Choi et al., 2004b; Kalo et al., 2004; Zhang et al., 2007; George et al., 2008; Hand et al., 2010; Nayak et al., 2010). Interestingly, this is independent of the genome size. For example, there is a high level of synteny between the genomes of *M. truncatula* and pea, even though the pea genome is 10 times larger than the genome of *M. truncatula*, and has one chromosome less (2n = 14) (Choi et al., 2004b). Despite its dramatically larger genome, pea shows high levels of microsynteny with *M. truncatula* and the differences in chromosome number are caused by just a few chromosomal rearrangements (Delseny, 2004; Kalo et al., 2004). Similarly, white clover displays a high degree of macrosynteny when compared to *M. truncatula* (Zhang et al., 2007), suggesting that the genomes of the species in the *Trifolium* genus are still largely collinear with the genome of *M. truncatula*. However, alignment of a red clover physical map to the *M. truncatula* genome sequence displays numerous translocation events (Sato et al., 2005). The number of rearrangements in red clover thus seems strikingly higher than is reported for other legume species that have been aligned to the *M. truncatula* genome (Sato et al., 2005; George et al., 2008; Hand et al., 2010). Because of the high degree of macrosynteny in white clover, it has been suggested that the rearrangements have mainly occurred in the *Trifolium* section, one of five major sections in the *Trifolium* genus, to which red clover belongs (Ellison et al., 2006).

Large-scale chromosomal rearrangements, including inversion, duplication, deletion, and translocation are frequently found in pericentromeric heterochromatin regions, which contain various types of repeats and transposable elements (Sharma and Raina, 2005; Raskina et al., 2008). These transposable elements are grouped into two classes. Class I contains RNA-mediated transposable elements and consists of two major groups: long terminal repeats retrotransposons (LTR-retrotransposons) and non-long terminal repeat retrotransposons (Non-LTR retrotransposon). Class II contains DNA-mediated transposable elements (DNA transposons) (Berg and Howe, 1989; Finnegan, 1989; Flavell et al., 1994). Non-LTR retrotransposons are well characterized in animals, but in plant they have been examined only in few species. Non-LTR retrotransposons can be distinguished into two types, LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements). LINEs can be several kilobases long, have no long terminal repeats, and possess a poly(A) tail which defines the 3′ terminus of the element (Schmidt, 1999). In contrast to LINEs, SINEs are up to several hundred basepairs in length, have a similar 5′ region as tRNA genes and a 3′ end similar to LINEs. The first LINE recognized in a plant species was the cin4 element from maize (Zea mays) (Schwarz-Sommer et al., 1987a; Schwarz-Sommer et al., 1987b). del2 was identified in *Lilium speciosum* (Leeton and Smyth, 1993) and several LINE-like elements have been identified in *Arabidopsis thaliana* and *Beta* species (Kubis et al., 1998). LINEs contain open reading frames (ORFs) encoding a gag protein, and endonuclease and reverse transcriptase domains that perform retrotransposition (Schmidt, 1999). SINEs have been most extensively studied in mammals, only a few SINEs have been identified in plant species, such as rice (*Oryza sativa*), tobacco (*Nicotiana tabacum*) and *Brassica* (Mochizuki et al., 1992; Yoshioka et al., 1993; Deragon et al., 1996). The conserved tRNA-like sequence motifs found in SINEs serve similar function as in tRNA genes, for transcription by RNA polymerase III. SINEs do not encode their own reverse transcriptase and therefore do not have the capability to transpose autonomously like LINEs. However they can be transposed by enzymes transcribed from LINEs, having a similar 3′ end to LINEs, and they often generate short duplication sites.

The first TEs associated with chromosomal breaks were found in maize, of which the genome is 84% composed of transposable elements, mostly LTR retrotransposons (McDonald et al., 1997; Craig et al., 2001). Studies in Drosophila have showed that clusters of DNA transposons and non-LTR RNA retrotansposons can trigger chromosomal inversion (Cirera et al., 1995; Ranz et al., 2001; Gonzalez et al., 2002; Casals et al., 2003). So both classes of transposable elements can mediate chromosomal rearrangements. In line with this, we aim to investigate whether the arrangements in red clover are associated with a specific type of transposable element.

To locate chromosomal rearrangements, comparative chromosome painting (CCP) can be applied. CCP uses fluorescent labeled bacterial artificial chromosome (BAC) clones containing sequences from a donor species in fluorescent *in situ* hybridization (FISH) experiments. By pooling many BAC clones, each containing up to a few hundreds of kb of sequence, large parts of chromosomes can be painted in closely related species. A clever choice of BAC clones anchored on physical and genetic maps allows determining the relative order of sequences in the recipient genome and thus co-linearity among the species. CCP has been used to reveal extensive chromosome reshuffling between *A. thaliana* and *Arabidopsis lyrata* and reconstruct karyotype evolution in *Brassicaceae* (Berr et al., 2006; Lysak et al., 2006; Mandakova and Lysak, 2008). Also, this technology is successfully applied in vertebrate species (Balmus et al., 2007; Ferguson-Smith and Trifonov, 2007; Romanenko et al., 2007; Sitnikova et al., 2007).

In this chapter, we study chromosome co-linearity between red clover and *M. truncatula* using CCP with BAC clones covering short and long arms of *M. truncatula* chromosome 5. We find that *M. truncatula* chromosome 5 is translocated into two chromosomes in red clover with several inter-chromosomal rearrangements. To uncover the extent of the chromosome co-linearity between the *Trifolium* genus and *M. truncatula*, we also applied CCP to seven other *Trifolium* species with five, seven or eight chromosome pairs. We find that extensive inter-chromosomal rearrangements only occurred in red clover, and that these rearrangements are associated with red clover specific repeats. To determine the type of these repeats, we sequenced the red clover BAC clone that spans the rearrangement on the short arm of *M. truncatula* chromosome 5 and find a red clover specific non-LTR retrotransposon; namely Tp_LINE that is positioned at the chromosomal breakpoint.

Results

The *M. truncatula* chromosome 5 homologous region is separated into six blocks in red clover

To visualize chromosomal rearrangements that have occurred in red clover and other species within the *Trifolium* genus, we apply comparative chromosome painting and focus on a single *M. truncatula* chromosome, namely chromosome 5 (Mt5). Both euchromatic arms of this chromosome are completely covered by a physical tiling path of BAC clones, which subsequently have been sequenced (see Chapters 2 & 3). To conduct chromosome painting, BAC clones were selected along the physical map of both arms with a spacing of ~400 kb on average. In this way, a total of 87 *M. truncatula* BACs were picked; 43 of the short arm and 44 of the long arm, respectively. For both arms the BACs were divided in 2 pools of 21-22 clones. BACs of each pool were labeled either with digoxigenin-dUTP (for green fluorescent detection) or biotin-dUTP (for red fluorescent detection).

First, we hybridized the labeled BAC pools to *M. truncatula* pachytene chromosomes. This resulted in an alternating red and green signal that formed a single continuous region in both arms of Mt5 (Fig. 5.1a, b). This is in agreement with the physical map data. Next, we hybridized the BAC pools of Mt5 to red clover pachytene chromosomes. Thereby, we conducted separate hybridizations for the BAC pools of either both arms of Mt5. Two pools of the short arm of Mt5 hybridized to the long arm of an individual pachytene chromosome of red clover. This visualized 3 rearranged chromosomal blocks (Fig. 5.1c). The terminal part of the long arm of red clover corresponds to the terminal part of the

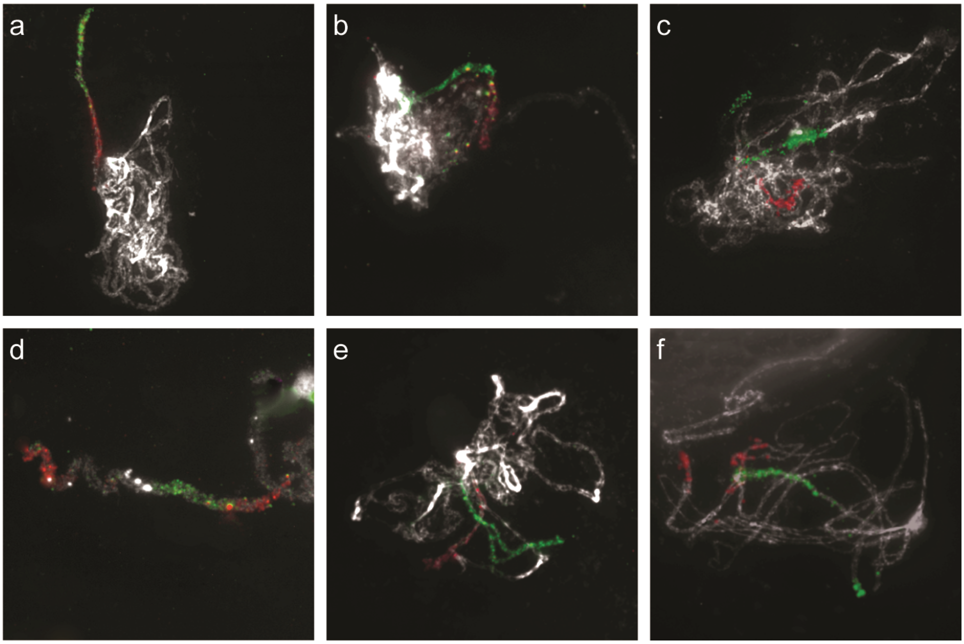


Figure 5.1: Genome comparison between *M. truncatula* and red clover using chromosome painting. (a) BACs covering *M. truncatula* chromosome 5 short arm were divided in two pools, each covering half of the arm. One BAC pool was detected with green fluorescent and another one with red fluorescent, and then painted on to *M. truncatula*, showing a continuous region on the short arm of *M. truncatula*. (b) The same method was used on the long arm of *M. truncatula* chromosome 5, showing a continuous region from the pericentromere until the telomere ends. (c) BACs covering *M. truncatula* chromosome 5 short arm hybridized on red clover, showing a translocation of part of the arm. (d) BACs covering *M. truncatula* chromosome 5 long arm also showed a translocation of the arm on red clover. (e) BAC pools of *M. truncatula* chromosome 5 short arm were labeled with green and the long arm was labeled with red, showing the clear signals on the euchromatic portion of *M. truncatula* chromosome 5. (f) BACs covering *M. truncatula* chromosome 5 short (green) and long arm (red) showed it was rearranged into different chromosomes on red clover with inter-chromosomal break.

short arm of Mt5. The remaining part of the short arm of Mt5 was translocated to the pericentromere-adjoining part of the long arm of this red clover chromosome (Fig. 5.1c). The same approach was applied for the two BAC pools originating from the long arm of Mt5. Hybridizing the pooled BACs to red clover pachytene chromosomes again revealed 3 blocks with rearrangements (Fig. 5.1d). These comprised the terminal part of the short arm and the pericentromere-adjoining part of the long arm of one red clover pachytene chromosome (Fig. 5.1d). To determine whether both arms of Mt5 map to the same red clover chromosome, a FISH experiment was performed in which BAC pools represented two arms of Mt5, in which the short arm was labeled with digoxigenin-dUTP and long arm was labeled with biotin-dUTP. This showed that two arms of Mt5 mapped to two different chromosomes in red clover, and underwent several chromosome breaks and translocations (Fig. 5.1e, f). Taken together, comparative chromosome painting confirmed that red clover experienced several chromosomal rearrangements when compared to *M. truncatula* (Fig. 5.2).

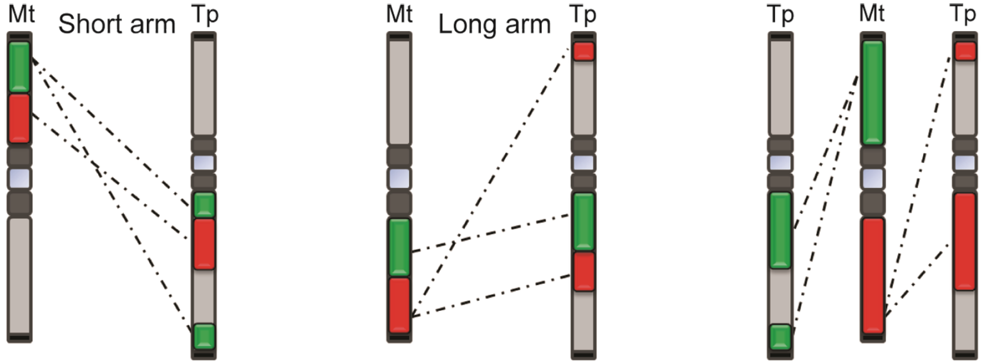


Figure 5.2: Schematic representation of comparative chromosome painting using *M. truncatula* chromosome 5 BAC pools as probes on *M. truncatula* (Mt) and red clover (Tp).

Trifolium* species are generally highly collinear with *M. truncatula

While complex chromosomal rearrangements were found between *M. truncatula* and red clover, the question remains at which point in evolution these rearrangements occurred. Based on comparative genetic mapping *M. truncatula* and white clover where found to be mostly co-linear (George et al., 2008; Hand et al., 2010). However, this study was based on 159 genetic markers across all chromosomes and thus cannot provide conclusive proof that no major rearrangements have occurred between Mt5 and its white clover counterpart. To determine whether white clover and *M. truncatula* are truly collinear, and therefore the red clover rearrangements occurred within the *Trifolium* subgenus, we applied comparative chromosome painting. White clover is an out-breeding, allotetraploid species (2n = 4x = 32) for which both diploid progenitors have been identified; *Trifolium pallescens* and *Trifolium occidentale*, respectively (Ellison et al., 2006). Instead of using allotetraploid white clover, we selected *T. pallescens* for the chromosome comparison with *M. truncatula*. The BAC pools representing the short and long arms of Mt5 were detected with green and red fluorescent labeling and used as probes on *T. pallescens* pachytene chromosomes. This showed that both arms of Mt5 were hybridized on a single chromosome of *T. pallescens* (Fig. 5.3a), indicating that this species shares a high level of macrosynteny with *M. truncatula*. This is in contrast to red clover. Therefore, we hypothesized that the chromosomal rearrangements in red clover have occurred within the *Trifolium* subgenus. To confirm this hypothesis, we studied two species of the *Chronosemium* subgenus; namely *Trifolium campestre* (2n = 14) and *Trifolium* patens (2n = 16). Comparative chromosome painting was conducted on metaphase chromosomes of both species using the labeled BAC pools of Mt5 as probes. The result showed that the two arms of Mt5 correspond to the both arms of a single chromosome in *T. campestre* and *T. patens* (Fig. 5.3b, c). Therefore we concluded that the chromosomal rearrangements we identified in red clover have occurred after the split of the *Trifolium* and *Chronosemium* subgenera.

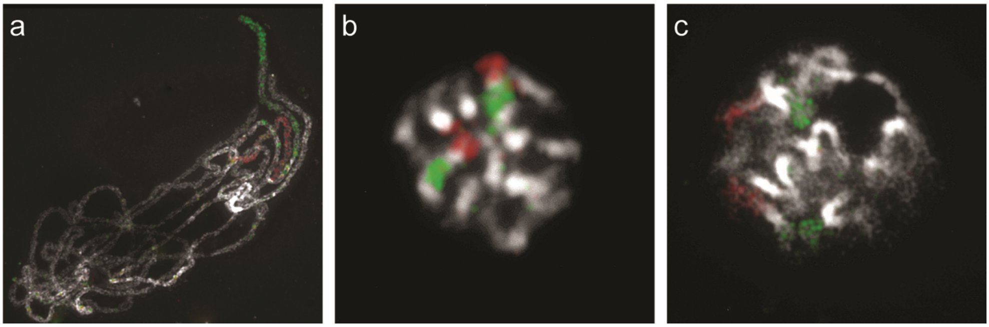


Figure 5.3: Chromosome painting using BAC pools covering the short arm (green) and the long arm (red) of *M. truncatula* chromosome 5 as probes on *T. pallescens*, *T. campestre*, and *T. hirtum*. (a) Hybridization to *T. pallescens* showed that two arms of *M. truncatula* chromosome 5 corresponded to two arms of one chromosome. (b) Hybridization showed that two arms of *M. truncatula* chromosome 5 represent two arms of one chromosome in *T. campestre*. (c) The same was observed in *T. patens*; both arms of *M. truncatula* chromosome 5 were hybridized to one chromosome.

Red clover chromosomes are more extensively rearranged than in related species

To determine at which point in the evolution the observed rearrangements have occurred in the *Trifolium* subgenus, we conducted comparative chromosome painting studies in the species that are more closely related to red clover than white clover. The *Trifolium* subgenus is divided in 5 main sections (*Trichocephalum*, *Trifolium*, *Vesicastrum*, *Trifoliastrum* and *Involucrarium*) (Fig. 5.4).

First we studied the Mt5 homologous region in *Trifolium medusem*, which is part of another section than red clover; namely *Trichocephalum* (Ellison et al., 2006) (Fig. 5.4). *T. medusem* was selected, because it has the same base chromosome number as red clover (2n =14), whereas most of the species in this section have 8 as base chromosome number (Ellison et al., 2006). We applied chromosome painting on *T. meduseum* pro-metaphase chromosomes, in which BAC pools represented two arms of Mt5. This revealed that both arms of Mt5 were represented by two arms of different chromosomes in *T. meduseum* (Fig. 5.5a). However, we did not find evidence for further inter-chromosomal rearrangements. This suggests that inter-chromosomal rearrangements have occurred in the *Trifolium* section.

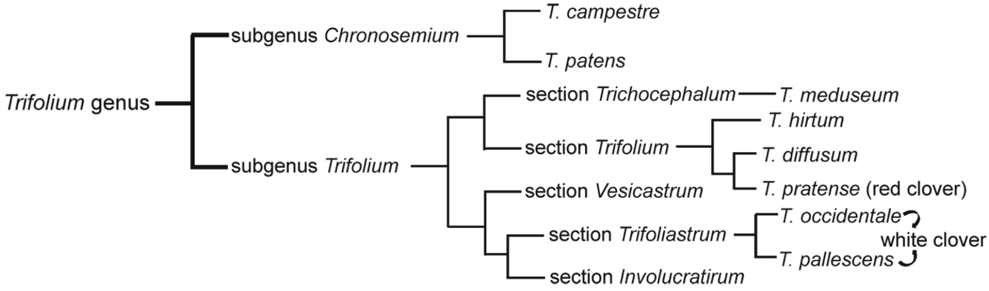


Figure 5.4: The *Trifolium* genus is divided in two subgenera, *Chronosemium* and *Trifolium*. Subgenus *Trifolium* includes five main sections.

Next we zoomed in the section *Trifolium*, by studying two species closely related to red clover; namely *Trifolium diffusum* and *Trifolium hirtum* (Fig. 5.4). *T. diffusum* (2n = 16) is a sister species of red clover, whereas *T. hirtum* (2n =10) has a different chromosome number (Ellison et al., 2006). Comparative chromosome painting was applied to these two species. The BAC pools representing the short and long arms of Mt5 were used as probes on *T. diffusum* and *T. hirtum* pro-metaphase chromosomes, and results were compared to those on red clover. The hybridization showed that both arms of Mt5 were represented by two arms of different chromosomes in *T. diffusum* and *T. hirtum* (Fig. 5.5b, c), similar as observed in *T. medusem*. On both species the hybridization signal was continuous from telomere to pericentromeric region, suggesting that no large chromosomal rearrangements occurred when compared to *M. truncatula* chromosome 5.

No evidence was found that multiple chromosomal breaks and translocations as observed in red clover occurred in a common ancestor with *T. hirtum*, *T. diffusum* and *T. meduseum*. Based on these results, we concluded that inter-chromosomal rearrangements occurred most likely within the red clover lineage. This means that the driving force of chromosomal rearrangement must be present in red clover, but absent

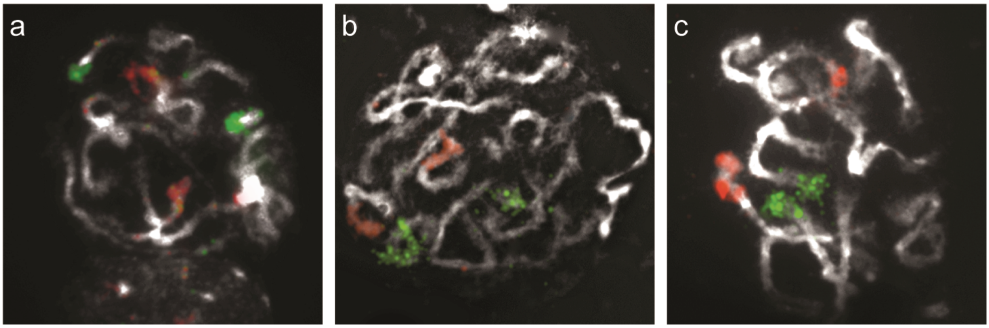


Figure 5.5: Chromosome painting using BAC pools covering short arm (green) and the long arm (red) of *M. truncatula* chromosome 5 as probes on *T. meduseum*, *T. diffusum*, and *T. hirtum* pro-metaphase chromosomes. (a) Hybridization to *T. meduseum* showed that two arms of *M. truncatula* chromosome 5 were translocated to two arms of different chromosomes. (b) Hybridization also showed that both arms of *M. truncatula* chromosome 5 represent two arms of different chromosomes on *T. diffusum*. (c) The same case was observed in *T. hirtum*; both arms of *M. truncatula* chromosome 5 were hybridized on two arms of different chromosomes.

in other *Trifolium* species. We hypothesize that red clover contains a DNA element that enhances heterologous recombination events. To provide support for this hypothesis, we studied the sequences around a breakpoint of one red clover specific translocation.

Chromosome breakpoints are marked with a red clover specific repetitive sequence

We showed the occurrence of several red clover specific inter-chromosomal rearrangements. The molecular mechanisms underlying chromosome evolution are largely unknown, though genome analysis have shown that breakpoints often are associated to repetitive sequence elements, including retrotransposons (Casals and Navarro, 2007). To determine whether the chromosomal rearrangements in red clover are associated with a specific repeat sequence we focused on a single breakpoint. Thereby we aim to pinpoint the recombined region and characterize the associated sequences. We hybridized to red clover pachytene chromosomes ever smaller *M. truncatula* BAC pools surrounding the approximate breakpoint region, determined based on linear length of the hybridized BAC pool signal. In this way we are able to pinpoint the break on a *M. truncatula* contig of 4 BACs; mth2-42c9 (CU326392), mth2-54e16 (CU137658), mth4-37c5 (CT963108) and mth2-181d5 (CT573076) spanning a 450 kb region (Fig. 5.6a-c). Next, we identified corresponding red clover BACs using BLAST based on red clover BAC end sequences. The identified red clover BACs where subsequently used for chromosome painting on red clover pachytene chromosome spreads to determine whether they co-localize with either of corresponding *M. truncatula* BACs mth2-42c9, mth2-54e16, mth4-37c5 and mth2-181d5. In this way a red clover BAC could be identified, Tp_c0045O12, that showed co-localization with *M. truncatula* BAC clone mth4-37c5 (Fig. 5.7a). This suggests that red clover BAC Tp_c0045O12 represents, at least in part, the region orthologous to *M. truncatula* BAC clone mth4-37c5.

Exploiting the red clover physical map we identified the corresponding contig that contained Tp_c0045O12. This contig (#419) contains 8 overlapping BACs (Tp_a0008F04, Tp_c0045O12, Tp_c0045C24, Tp_b0019N02, Tp_c0012C10, Tp_0021K17, Tp_b0001L17 and Tp_b0001K24) and spans ~200 kb (Fig. 5.7b) (Skot et al., in preparation). We selected three red clover BACs (Tp_c0045C24, Tp_c0012C10, Tp_b0001L17) that formed a tilling path and used these as probe to study co-localization with *M. truncatula* BAC clone mth4-37c5 (Fig. 5.7b). Red clover clone Tp_c0045C24 co-localized with *M. truncatula* clone mth4-37c5, as it formed a single focus near the telomere end in red clover (Fig. 5.7c). In contrast, red clover clones Tp_c0012C10 and Tp_b0001L17 mapped at several locations, resulted in a high background (Fig. 5.7d). Therefore, we conclude that a repeat sequence is present in the red clover clones Tp_c0012C10 and Tp_b0001L17.

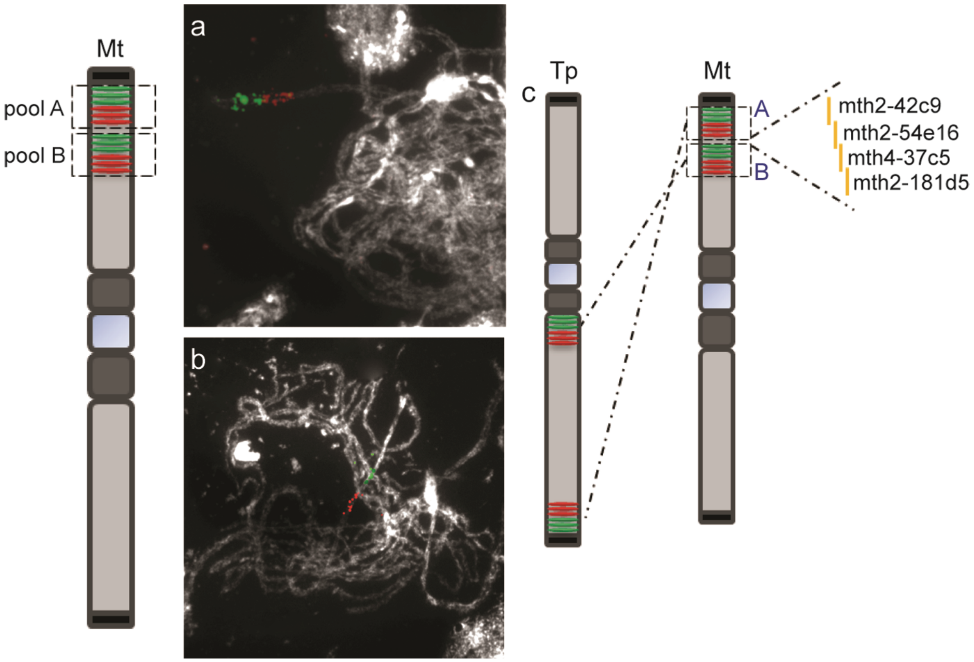


Figure 5.6: Pinpointing the chromosomal break in red clover (Tp). (a) Pool A of six *M. truncatula* BACs that hybridize near the telomere end of the short arm of *M. truncatula* chromosome 5 was divided into two smaller pools; three Mt BACs were labeled with green fluorescent tags and three with red. They were used as probes for the hybridization on red clover, showing that green detected BAC pools hybridized near the telomere ends and red detected BAC pools adjacent to the green signals on the centromeric side. (b) Pool B of six *M. truncatula* BACs that hybridize closer to the pericentromeric region in the short arm of *M. truncatula* chromosome 5, used as probes in red clover, showing that the three green tagged *M. truncatula* BACs hybridized near the pericentromere and are flanked by the red tagged *M. truncatula* BACs. (c) Schematic representation of summarizing figure (a) and (b) shows the breakpoint region in red clover is pinpointed between pool A and pool B, which contain 4 *M. truncatula* BACs.

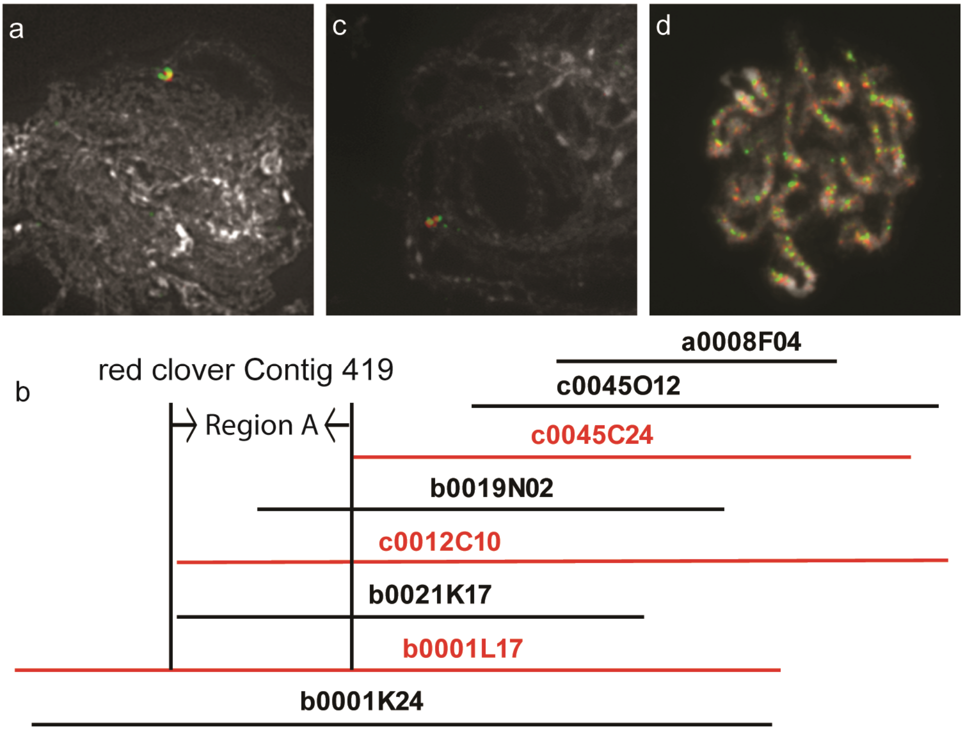


Figure 5.7: Identification of the red clover BAC clones associated with the red clover chromosomal rearrangement. (a) Red clover BAC clone Tp_c0045O12 (red signal) is co-localized with *M. truncatula* BAC clone mth4-37c5 (green signal), near the telomere end in red clover. (b) Total 8 red clover BACs were identified in contig 419. Three BAC clones (Tp_c0045C24, TP_c0012C10 and Tp_b0001L17) were picked as a tilling path. (c) Red clover BAC clone Tp_c0045C24 (red signal) was hybridized at the same position as *M. truncatula* BAC clone mth4-37c5 (green). (d) Red clover BAC clones Tp_c0012C10 (green) and Tp_b0001L17 (red) both mapped at multiple locations. Yellow signals indicated overlap between BACs.

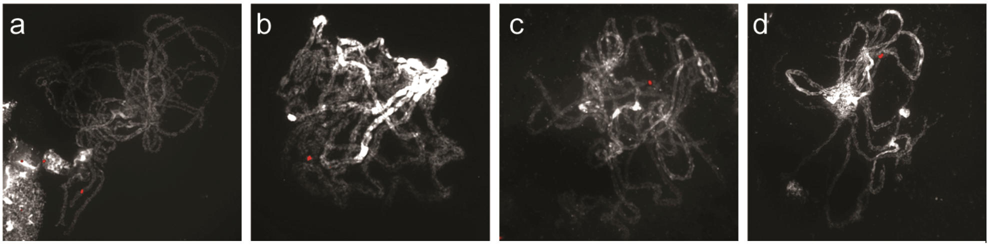


Figure 5.8: FISH mapping of repeat containing red clover BAC clone Tp_b0001L17 on (a) *T. pallescens* (b) *T. hirtum* (c) *T. diffusum* (d) *M. truncatula* showied a single focus in all four species.

Red clover specific non-LTR retrotransposon element is associated with an inter-chromosomal break

To determine whether the repeat sequence present in clone Tp_b0001L17 is red clover specific, we conducted FISH experiments in related species *T. pallescens*, *T. hirtum*, and *T. diffusum*, as well as in *M. truncatula*. These experiments revealed that Tp_b0001L17 mapped to a single position in all four species (Fig. 5.8a-d). This strongly suggests that the repeat sequence present in Tp_b0001L17 did not occur in related *Trifolium* species, and possibly is unique to red clover.

To identify the repeat sequence causing non-specific hybridization, Next-Generation sequencing was applied to clone Tp_b0001L17 and Tp_c0045C24. Around 11x coverage was obtained using 50 bp paired end reads. The obtained sequence could be assembled in 12 large contigs (>5 kb) plus 33 smaller ones (<5 kb) representing 213,731 kb unique red clover sequence. We focused on the repeat sequences present in the region unique to red clover clone Tp_b0001L17 when compared to Tp_c0045C24 (region A, Fig. 5.7b). Within this region we identified a LINE-like non-LTR retrotransposon using the repeatmasker software (www.repeatmasker.org). The identified repeat sequence shared 33% identity with *M. truncatula* reverse transcriptase and 32% identity with *A. thaliana* non-LTR retrotransposon transposase, according to BLAST (Fig. 5.9a). We named this element *Trifolium pratense* LINE-retrotransposon (Tp_LINE). Strikingly, the sequences

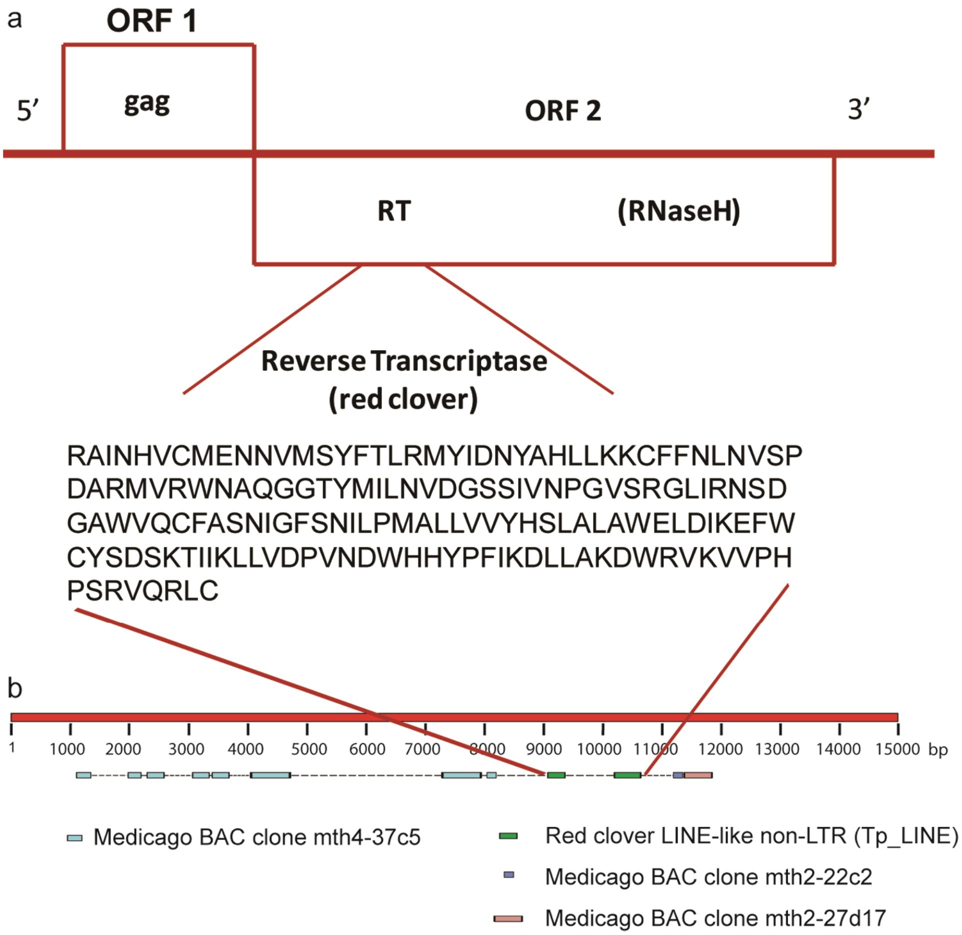

a
ORF 1
gag
5’
ORF 2
3’
RT
(RNaseH)
Reverse Transcriptase
(red clover)
RAINHVCMENNVMSYFTLRMYIDNYAHLLKKCFFNLNVSP
DARMVRWNAQGGTYMILNVDGSSIVNPGVSRGLIRNSD
GAWVQCFASNIGFSNILPMALLVVYHSLALAWELDIKEFW
CYSDSKTIIKLLVDPVNDWHHYPFIKDLLAKDWRVKVVPH
PSRVQRLC
b
1
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
bp
Medicago BAC clone mth4-37c5
Red clover LINE-like non-LTR (Tp_LINE)
Medicago BAC clone mth2-22c2
Medicago BAC clone mth2-27d17

Figure 5.9: Schematic representation of red clover LINE-like non-LTR retrotransposon. (a) The structure of the non-LTR retrotransposon and red clover reverse transcriptase was characterized as a 166 amino acid fragment. (b) One side of Tp_LINE showed homology with *M. truncatula* BAC Clone mth4-37c5 and another side showed homology with *M. truncatula* BAC clone mth2-22c2 and mth2-27d17.

flanking Tp_LINE showed at one side homology with *M. truncatula* BAC clone mth4-37c5 (chromosome 5), whereas the other side is homologous to *M. truncatula* BAC clone mth2-22c2 and mth2-27d21 (chromosome unknown) (Fig. 5.9b). Subsequent annotation using FGENESH predicted that the SNARE associated Golgi protein was present at the side of *M. truncatula* BAC clone mth4-37c5. However, in the side of clones mth2-22c2 and mth2-27d17 we did not find any putative genes. *M. truncatula* clone mth4-37c5 originates from the short arm of Mt5, whereas mth2-22c2 and mth2-27d17 are not mapped. Because Mt5 is sequenced near to completion (chapter 3) and mth2-22c2 and mth2-27d17 are not part of Mt5, it strongly suggests that both BACs originated from a different chromosome. This is in agreement with a chromosomal rearrangement in red clover having occurred at the position of Tp_LINE. To determine whether Tp_LINE represents repetitive sequence present in BAC clone Tp_b0001L17 we applied FISH on red clover chromosomes using Tp_LINE and clone Tp_b0001L17 as probe. This showed that Tp_LINE (green) hybridized at multiple locations on different chromosomes and overlapped with the signal of clone Tp_b0001L17 (red) (Fig. 5.10).

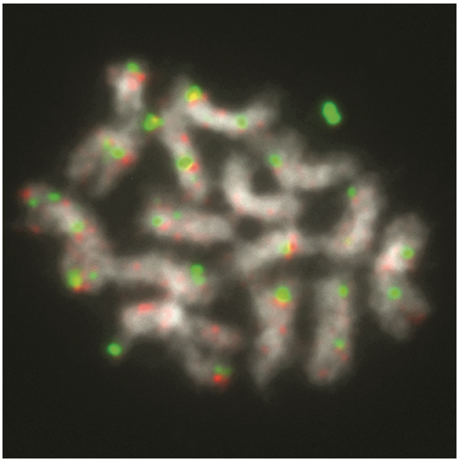


Figure 5.10: Digoxigenin-dUTP labeled Tp_LINE (green) hybridized at multiple locations and overlapped with biotin-dUTP labeled red clover clone Tp_b0001L17 (red).

Discussion

Here we show that comparative chromosome painting (CCP) can be applied to analyze the amount of genome co-linearity among plant species even when no comparative genetic and cytogenetic maps are available. By performing CCP with *M. truncatula* BAC probes on red clover and other *Trifolium* species, chromosome translocation events have been detected that contributed to the evolution of the karyotypes of species in the *Trifolium* genus.

An important finding in our study was that chromosomal rearrangements were more extensively observed in *Trifolium* and *Trichocephalum* sections than in the other three sections (*Vesicastrum*, *Trifoliastrum*, and *Involucrarium*) of the *Trifolium* subgenus. Based on current understanding of genome evolution in the *Trifolium* genus, an ancestral chromosome number of 8 is most common. Species with deviating chromosome numbers occur mostly in sections *Trifolium* and *Trichocephalum*, in which most species are diploid and annual. This observation is supported by our finding that two arms of *M. truncatula* chromosome 5 were translocated to two different chromosomes in *Trifolium* species belonging to the two sections *Trifolium* and *Trichocephalum*. This was not the case for *T. pallescens* belonging to the section *Trifoliastrum*. Chromosome number variations are indeed rare in the three large sections *Vesicastrum*, *Trifoliastrum*, and *Involucrarium*, in which most species are polyploid and perennial (Ellison et al., 2006). The presence of extra chromosomes may impede the formation or fixation of rearranged chromosome complements, and as polyploidy is associated with a perennial lifestyle (Taylor and Smitht, 1979), perhaps it contributes to a higher rate of evolution in annual species as compared to perennial species.

While some rearrangements are shared by all species that are tested in the *Trifolium* genus, additional rearrangements have occurred in red clover, causing reduction of the chromosome number from the ancestral complement of 8 to 7. Theoretically the reduction in chromosome number could be achieved by simple translocations involving only two chromosomes. However, in case of red clover the evolutionary history of the observed rearrangements are more complex. In *T. hirtum* and *T. meduseum*, the two arms orthologous to *M. truncatula* chromosome 5 are represented by single arms, albeit on different chromosomes. In red clover, however the two arms of *M. truncatula* chromosome 5 are underwent further rearrangements, including pericentromeric and paracentromeric inversions. A comparison between red clover and *T. medusem*, which have the same chromosome number but are not closely related, shows that the reduction in chromosome number indeed occurred independently. Also, comparison between red clover (2n = 14) and *T. hirtum* (2n = 10) shows that chromosome number reduction occurred several times independently within the same section.

The red clover specific inter-chromosomal breakpoints are associated with clusters of tandem repeats and transposable elements most commonly found in pericentromeric regions. Chromosomal rearrangements frequently occur in pericentromeric regions, most likely by ectopic homologous recombination between repetitive elements. Especially, transposable elements have been long recognized as a cause of chromosomal rearrangements. We found that an inter-chromosomal breakpoint in red clover is associated with LINE-like non-LTR retrotransposons. The LINE element we discovered is located right at the breakpoint of a chromosomal rearrangement between red clover and does not seem to occur in closely related clover species. Therefore, it is likely that this LINE-like non-LTR retrotransposon is the cause of the inter-chromosomal rearrangement in red clover. To our knowledge, this is the first time that a direct correlation has been found between a LINE-like non-LTR retrotransposons and a chromosomal rearrangement.

In contrast to LTR retrotransposons, which are well characterized and omnipresent in plant genomes, the occurrence of non-LTR retrotransposons is poorly described in plants. In general, LINEs are variable in structure and sequence in different plant species. Although they contain seven conserved domains, the sequence of the encoded reverse transcriptase can be greatly diverged. For instance, in the *Beta* species, homology between the sequences of reverse transcriptase ranged between 43% and 96% (Schmidt et al., 1995; Kubis et al., 1998), in agreement with observations in different lines of *Arabidopsis* (Wright et al., 1996). This is consistent with our finding that red clover contains a unique LINE, of which the sequence has apparently diverged quickly in the short time since red clover speciation. This may due to accumulation of mutations in integrated LINE copies over evolutionary history.

Concluding, in this chapter we have shown that chromosomal rearrangements have occurred at several times in the evolutionary history of the *Trifolium* genus. Some chromosomal rearrangements date back to the split between sections *Trifolium* and *Trichocephalum*, while others are specific for red clover. More generally we show that using comparative chromosome painting, it is possible to study the evolutionary history of plant families without resorting to extensive genomic studies.

Material and methods

Plant material

All plants were grown from seeds under green house condition, which is 12 hours daylight and 20-22°C. The seeds of red clover variety Milvus, *Trifolium* species *T. diffusum*, *T. hirtum*, *T. pallescens*, *T. campestre* and *T. patens* were obtained from Aberystwyth Genetic Resources Information System, IBERS, Aberystwyth Univeristy. The seeds of *T. meduseum* were obtained from U. S. national Plant Germplasm System (NPGS), USDA, ARS, WRPIS, Washington State University. *M. truncatula* genotype Jemalong A17 was used. The flower buds were collected between 9 to 11 AM and fixed in acetic acid-ethanol (1:3). The fixative was replaced a few times until the solution remained clear. The flower buds were left in the fixative solution at 20°C for one month, and then transferred into 70% ethanol at -20°C until use.

Slide preparation

The method for *M. truncalua* slide preparation was performed according to the protocol described by Olga Kulikova (Kulikova et al., 2001). The method for *Trifolium* species slide preparation was adapted based on *M. truncatula* slide preparation. Flower buds of 2 mm in length for red clover, *T. diffususm*, *T. hirtum*, and *T. pallescens*, 3 mm in length for *T. meduseum*, *T. campestre* and *T. patens* were rinsed twice with Milli-Q water, once in citrate buffer (10 mM sodium citrate, pH 4.5 with citric acid), and then incubated in 1% enzyme mixture-citrate buffer (1:2) at 37°C in a water-saturated atmosphere for 1 hour and 15 minutes for cell wall digestion. The enzyme mixture included 1% cytohelicase (Bio Sepra 249701), 1% pectolyase Y-23 (Sigma P-3026), and 1% cellulose RS (Yakult 203027), in citrate buffer. After the enzyme treatment, the soft flower buds were rinsed with sterile Milli-Q water 3 times. The anthers were dissected from a flower bud. One anther suffices for one slide preparation. The anther was transferred to the middle of a cleaned grease-free slide with 2 µl Milli-Q water, and then carefully dissected and homogenized with a needle. 80 µl of 60% acetic acid was added, and individual cells were spread on the slide. The slide was baked at 50°C hot plate for 2 minutes to remove cytoplasm. The slide was taken off and ice-cold acetic acid: ethanol (1:3) was added in a circle around the suspension. The slide was left on the hot plate for drying. All slides were screened using phase contrast microscope for chromosome spreading and presence of pachytene stage nuclei. Selected slides were stored in a dust free microscope slide box until use.

Probe isolation and labeling

Bacterial artificial chromosome (BAC) clones of *M. truncatula* for comparative chromosome painting were identified based on their physical map position (http://medicagohapmap.org/) with a spacing of ~400 kb on average between BAC clones. The corresponding red clover BAC clones were identified using BLAST based on red clover BAC end sequences. Both *M. truncatula* and red clover BAC clones were isolated according to PureLink TM HiPure Plasmid DNA Purification Kits (Invitrogen). BAC clones were labeled with either biotin-dUTP or digoxigenin-dUTP nick translation mix (Roche).

Cot-100 isolation

Red clover Cot-100 was prepared according to the protocol described in chapter 4.

Fluorescence *in situ* hybridization (FISH)

FISH mapping of individual *M. truncatula* BAC clones was performed according to the protocol described by Olga Kulikova (Kulikova et al., 2001). FISH mapping of individual red clover BAC clones was performed according to the protocol described in chapter 4. To perform cross-species FISH using *M. truncatula* BAC clones on red clover and other *Trifolium* species *T. diffusum*, *T. hirtum*, *T. pallescens*, *T. campestre*, *T. meduseum* and *T. patens*, 35% formamide (35% deionized formamide, 2 x SSC, pH7) instead of 50% formamide (50% deionized formamide, 2 x SSC, pH7) was used in the post-hybridization washing.

Comparative chromosome painting

Comparative chromosome painting was performed according to the protocol described by Mandakova and Lysak (2008) with some modifications. Pepsin treatment was omitted. Biotin or digoxigenin-dUTP labeled BAC clones were pooled and precipitated through addition of one-tenth volume of 4M Lithium Chloride, and 2.5 volume of ice-cold absolute ethanol, incubated at -80°C for 30 minutes or -20°C for overnight. The BAC DNA was collected by centrifuging at 13,000g, 4°C for 30 minutes. The pellet was air dried and suspended in 20 µl HB50 (50% deionized formamide, 2 x SSC, 50 mM 1M sodium phosphate pH7, and 20% dextran sulfate). 40 µl hybridization mix was applied on one slide. Hybridization mix contained 20 µl precipitated BAC DNA, 50 to 1 ratio of labeled BAC DNA and Cot-100, up to total 40 µl with 20% dextran sulfate in HB50 (50% deionized formamide, 2 x SSC, 50 mM 1M sodium phosphate pH7, 20% dextran sulfate). The probe and chromosomes were denatured on a hot plate at 80°C for 3 minutes and incubated in a moist chamber at 37°C for at least 20 hours.

Post-hybridization washing was performed in 50% formamide (*M. truncatula* BAC clones on its own chromosomes) or 35% formamide (*M. truncatula* BAC clones on the chromosomes of *Trifolium* species) in 2x SSC at 42°C for three times, each five minutes. The detection and amplification steps were followed as described above for the FISH. The images were captured in fluorescence microscope and Photometrics Sensys 1305x1024 pixel CCD camera (Zeiss).

Sub-cloning

Red clover BAC clone c0012C10 was used as template DNA. 5 µg BAC DNA was digested separately using restriction enzyme Pvu II, EcoR V and Swa I with appropriate restriction buffers in total 40 µl sterilized Milli-Q water at 37°C for 18 hours. The digested BAC DNA was purified through PCR purification kit (Roche). pJET1.2/blunt cloning vector kit (Fermentas) was used for cloning, after which plasmids were introduced into E. coli.

Colony hybridization

Plates with sub cloned colonies were replicated by striping 90 colonies per new plate, and incubated overnight at 37°C. Nytran-Plus Membrane filter (Dassel) was placed on the top of the colonies for 2 minutes. The membrane filter with colonies side up was transferred to denaturation buffer (0.5M sodium hydroxide, 1.5M sodium chloride) for 5 minutes, followed by neutralization buffer (1.5M sodium chloride, 0.5M Tris, pH 8.0) for 5 minutes. The neutralized membrane filter was rinsed in buffer (0.2M Tris in 2x SSC, pH 8.0) for 2 minutes. The filter with colonies side up was baked at 80°C for 1 hour. Post washing was performed in rub buffer (0.2M Tris, 1M sodium chloride, 0.1% SDS, in 2x SSC, pH 8.0) at 42°C for 15 minutes, dried at room temperature. The hybridization was carried out according to standard procedure (Sambrook, Fritsch, and Maniatis 1989). The BAC probe c0045C24 was labeled with biotion-dUTP nick translation mix. Hybridization was performed overnight in standard buffer with 50% formamide at 42°C. Stringency washes were carried out with 0.1% SDS in 2x SSC and 0.1% SDS in 1x SSC at 68°C, each 15 minutes. The colonies without hybridization signals were picked up and isolated according to the Easy Nucleic Acid Isolation Kit (Omgea Biotek). The colonies with positive insert were sequenced and assembled with software DNA star.

BAC sequencing

BAC DNA clones were end-sequenced on a Illumina Genome Analyzer II (Illumina Inc., San Diego, USA) platform and assembled using CLC Genomics Workbench software (CLC bio, Aarhus, Denmark). This resulted in approximately 11 x 10^6 paired trimmed reads with an average length of 202 nucleotides. The total amount of sequenced DNA from all assembled contigs was 213,731 kb. Genes were predicted using FGENESH (softberry). Artemis genome browser was used to view and edit the data (Rutherford et al., 2000).

CHAPTER 6

General Discussion

Application of fluorescent *in situ* hybridization (FISH) as a validation tool in whole genome sequencing projects

In the last few years, whole-genome sequencing has revolutionized biological sciences by making available the genome sequences of many species. While in the last decade a few genomes of model organisms were laboriously sequenced by large sequencing consortia over the course of many years, nowadays many eukaryote sequences are available with more being produced every month. In the field of plant sciences, many genomes such as *Arabidopsis thaliana*, rice (*Oryza sativa*), soybean (*Glycine max*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), poplar (*Populus trichocarpa*), strawberry (*Fragaria vesca*), domesticated apple (*Malus x domestica Borkh*) and cucumber (*Cucumis sativus*) are available already (The Arabidopsis Genome, 2000; Yu et al., 2002; Tuskan et al., 2006; Huang et al., 2009b; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010; Velasco et al., 2010; Shulaev et al., 2011), and almost any crop species of great economic importance is either being sequenced or has been completed. The generation of gigabases of sequence has now become routine.

However, while sequencing itself has become very effective and relatively easy, sequence assembly has not become easier. The most accurately assembled sequences available today are those where an accurate physical map was produced based on clones containing hundreds of kbs of sequences, often in bacterial artificial chromosomes (BACs). These BAC clones were sub-cloned and sequenced individually using Sanger sequencing, which results in long reads of up to a kb. Assembly is therefore relatively easy, because relatively few long sequences need to be assembled. Even so, many gaps can remain in assembled sequence because long stretches of repetitive sequences cannot be spanned by individual sequence reads or BAC clones are absent in the libraries. Therefore pieces of unique sequence on either end of a large repetitive sequence block cannot be linked together, creating a gap. For further assembly genetic maps are required, which contain the positions of polymorphisms that can be identified in the assembled sequence. Such maps can then be used to order sequenced fragments into linkage groups, but physical sizes of gaps cannot be easily identified from genetic maps because recombination frequencies are not equal across the genome. Also, not every sequence fragment contains polymorphisms which can be used to tether the fragment to a genetic map. Therefore, to improve the quality of sequence assemblies, other techniques are required.

One of these techniques is fluorescent *in situ* hybridization (FISH). In FISH, DNA from individual clones within a sequence assembly is fluorescently labeled and used as probe to identify the location of the done on a karyogram. Thus an approximate genomic position can be determined for labeled clones. The probe DNA can vary widely in size, from large labeled fragments of 50 - 150 kb, which can often be observed directly, to as small as a 500 bp fragment, which can be detected indirectly through application of amplification using specific antibodies. Unique sequences are readily detected in this way, but the presence of large numbers of tandem or dispersed repeats in the BAC clones, which hybridize on multiple loci, can cause the mis-localization of BACs. To suppress repeat hybridization, repeats can be masked by providing an excess amount of isolated repeat sequence, such as a Cot-100 preparation. We showed that in medium-sized plant genomes, repeats can be effectively suppressed in the labeled BAC clones, allowing the localization of almost any BAC clone. We then used these data to improve and validate the sequence assembly in various ways.

In this thesis, FISH was used to validate and improve the assembly of the genome sequence of *Medicago truncatula*. A major question in the *M. truncatula* genome sequencing project was how much euchromatin was covered by the assembled sequence and how large remaining gaps were. To visualize the locations of gaps on the chromosomes, the microscopic distance between clones on opposing ends of the gap is directly measured in karyograms. The best resolution is obtained in FISH on pachytene chromosome spreads, which are 20-30 times longer than metaphase chromosomes, and allow direct observation of euchromatin and heterochromatin. Through the assumption that chromatin condensation is reasonably uniform throughout euchromatin, microscopic distances can be converted into physical distances. Thus FISH can be used to measure gap sizes, as we show in chapter 3 of this thesis. Similarly, the distances between the terminal sequences of an assembly and the actual end of chromosome arms can be measured. In this way, accurate assessment of the progress of sequencing is possible. While such assessments can also be derived from sequencing statistics, such as through k-mer analysis (Chor et al., 2009), FISH still provides a much more direct and unambiguous measurement of sequence coverage.

Application of FISH in modern shotgun sequencing

In contrast to the old, slow method of BAC-by-BAC sequencing, current shotgun sequencing produces billions of sequence reads often no longer than 50bp, which need to be assembled into a single gigabase-long assembly. This complicates the assembly problem enormously. As a result, usually hundreds of thousands of individual contiguous sequences (contigs) are produced. Assembly is made slightly easier by sequencing both ends of clones of various sizes. Paired-end information can then be used to assemble contigs into scaffolds, ordered sets of contigs for which paired-end clone size provides a measure of gap sizes. Even then still thousands of scaffolds remain, which need to be placed in order using genetic methods. The resulting genome sequence is obviously of much lower quality than a BAC-by-BAC sequenced genome. However, shotgun genome sequences can be obtained for a fraction of the cost of higher-quality sequences and are thus far more economical than high-quality sequences.

The question is therefore what FISH can contribute to improve the assembly of shotgun-sequenced genomes. FISH can be used for gap-sizing and determining of sequence coverage by using as probes large paired-end sequenced clones that are assembled into scaffolds, just like in BAC-by-BAC sequencing projects. But some of the assembly problems typical for shotgun sequencing can also be solved using FISH. Pooled-clone FISH could be used to assemble scaffolds which do not carry any genetic markers into linkage groups. Due to the relatively short lengths of contigs in shotgun sequencing, relatively more contigs do not carry genetically mapped polymorphisms than in BAC-by-BAC sequencing. FISH may also be used to verify unreliable assemblies which have low confidence scores deriving from the assembly method. In such cases, a number of adjacent contigs or scaffolds could be labeled with alternating colors, which should result in an unbroken chain of alternating red and green signals if the assembly is correct.

Use of FISH for curation of annotation in shotgun-sequenced genomes

After sequencing and assembly, annotation is necessary to find genes and correct sequencing errors. However, costs for thorough annotation of genome sequences are now exceeding costs for sequencing and assembly of shotgun sequenced genomes. While annotation pipelines exist that greatly speed up annotation (Elsik et al., 2007; Cantarel et al., 2008), extensive manual curation of every scaffold is still required to correct errors in gene model predictions. Often costs for such annotation are prohibitive and no or limited manual curation is performed (Al-Dous et al., 2011; Dassanayake et al., 2011). Because many genes in plants occur in families, limited curation of assembly and annotation can very easily lead to collapse of homologous sequences. In gene cloning projects a missed homologue can easily result in months of wasted work. Therefore, when working with sequences from any poorly annotated genome, great care should be taken to verify whether assembly and annotation are correct. FISH can then be a very useful tool to screen for presence of homologues independently of the assembly, and to verify that genomic positions of sequences are correct. Also, the order of syntenic sequences between compared species, as determined using FISH, can be used to determine whether two genes are orthologs or homologs.

Application of FISH in comparative genomics

The question how species evolve and how new species originate, is still and important topic of research. Therefore it is important from an evolutionary perspective to understand what chromosomal rearrangements have occurred during speciation. While comparison between sequenced species is a powerful way to find many chromosomal rearrangements (Hu et al., 2011) and while such comparisons have identified genome duplication events as important drivers of evolution (Flagel and Wendel, 2010), there is no yet plant family in which sufficient members have been sequenced to map individual rearrangements to a phylogeny. In chapter 5 of this thesis we used comparative chromosome painting FISH to compare the genomes of several *Trifolium* species and *M. truncatula*, in order to find the evolutionary history of chromosomal rearrangements between these species. By careful selection of species across the different clades, we managed to date individual chromosomal rearrangements to several points within the evolutionary history of the *Trifolium* genus. The main advantage of using FISH over other techniques such as sequence alignments is that cross-species FISH can be applied to a large number of species in a family without the need to sequence them all. Another advantage of using FISH for comparative genomics is that FISH is relatively insensitive to small deletions and insertions, such as commonly occur between species. When comparative studies are performed based on alignments of sequenced data or on genetic map comparisons, many observed rearrangements are caused by transposon-associated small insertions, deletions or duplications, resulting in spurious signals that need to be removed through heavy filtering. For instance, when *Arabidopsis* and *Brassica rapa* were compared, 50% of genes in a region needed to correspond between the species for synteny to be observed (Mun et al., 2009). Such high levels of synteny become less common when genetic distance between species increases. In contrast, cross-species FISH is less sensitive to small rearrangements and transposon action, and because FISH probes can be detected with great sensitivity, probe alignment may be observed even when only a small fraction of synteny between the species is retained. Thus, cross-species FISH remains the preferred method for the localization of chromosomal rearrangements.

The future of FISH in the genomic era

As described in the previous paragraph, FISH is a great tool to resolve problems that occur in genome sequencing and comparative genomics, which are difficult to solve with sequencing or genetic map based methods. The main reason why FISH is not used more often is that it has relatively low throughput. While the FISH protocol itself is relatively straight-forward and could be automated using robots, preparation of good pachytene spreads on slides is relatively difficult and only results in a few well-spread nuclei at best. As a result, microscopy is labor-insensitive. For large genome sequencing projects resulting in high-quality sequence assemblies, it is feasible to close the few remaining sequence gaps using FISH. But for shotgun-sequenced genomes containing many gaps, labor cost and effort are prohibitive. Therefore, a method to improve the efficiency of FISH is urgently needed.

The best way to parallelize preparation of biological materials is to use protoplasts or isolated nuclei. These can be kept in solution, meaning that experiments can be easily scaled up to the numbers required in modern genomics using pipetting robots or fluorescence assisted cell sorters (FACS). Unfortunately, it is difficult if not impossible to induce protoplasts to perform meiosis, meaning that pachytene spreads are not available. Therefore, other cell stages need to be used in FISH.

One solution could be to lyse cells and nuclei, and gently wash them across a surface, resulting in stretched fibers as in existing fiber FISH protocols (Parra and Windle, 1993). Fiber FISH has even higher resolution than pachytene FISH and pairs of labeled clones could be visualized on fibers relatively easily. Microscopic distances between the clones can then be measured like in pachytene FISH. While slide preparation for fiber FISH is also technically challenging, it could be parallelized more easily than pachytene spread production, and automated microscopy and image analysis would be more easy to implement. A disadvantage of this technique is that fiber FISH is not suitable for measuring long distances of several megabases between clones. Another method to parallelize FISH would be to use metaphase FISH, possibly on protoplasts. The main drawback of metaphase FISH is that chromosomes are very condensed during this stage of mitosis, which means it is difficult to obtain spatial resolution to clearly observe and measure separation between signals. One solution could be the use of super resolution microscopy (Huang et al., 2009a; Huang, 2010), which is however still under development.

The most promising method to parallelize FISH is to perform FISH on interphase nuclei. While interphase chromosomes are folded into three-dimensional domains in which genetic distances do not directly relate to microscopic distances, there is a statistical correlation between genetic and microscopic distance across an ensemble of nuclei (Bohn et al., 2007). Therefore, for two clones separated by up to several Mb, the physical distance could be calculated from a sample of several hundred distances measured in individual nuclei. Interphase FISH is relatively easy to perform and automate using cell sorting machines (Trask, 2002; Watanabe et al., 2009). Measurements could be performed on an automated microscope. Thus a fully automated FISH machine could be constructed capable of determining physical distances between many pairs of clones per day. Such an approach could result in a revival of the use of FISH in genomic research.

Conclusions

In this thesis, FISH has been employed as a tool to solve many questions originating from the *Medicago* genome sequencing project, and to perform comparative genomics between *Medicago* and many *Trifolium* species. This thesis thus shows that FISH is an important research method that complements whole genome sequencing very well by providing an independent way to verify sequencing results and extend the use of reference genomes towards research in genomes of species that have not been sequenced. Moreover, with some modifications the throughput of FISH techniques could be improved, which would lead to many more applications of FISH in the genomic era.

Summary

The Legume family contains a number of important crops that provide a major source of protein and lipids for both human and animals. More than 30% of dietary protein for human consumption and more than 35% of industrial vegetable oil comes from legumes. Legumes are also natural fertilizers due to their unique capability to symbiotically interact with nitrogen-fixing rhizobium bacteria. Bacteria and legumes can fix atmospheric nitrogen to provide ammonium for plant use. This process is the most important biological source of organic nitrogen compounds. Therefore, a great amount of scientific study on legumes is dedicated to the elucidation of this nitrogen fixation symbiosis.

For the advancement of nitrogen fixation studies, information derived from genome sequencing is essential. *Medicago truncatula* has been selected as a legume model species for legume genome sequencing. In chapter 2, we describe the sequence of the euchromatic region of the *M. truncatula* genome. A BAC-by-BAC sequencing approach in combination with Next-Generation sequencing is used in this genome sequencing project. The genome sequence is assembled and annotated, capturing ~94% of all *M. truncatula* genes. This provides valuable insights into legume specific traits. Analysis of the sequence has shown that a whole genome duplication has occurred in the legume lineage approximately 58 million years ago. The genome sequence of *M. truncatula* was also compared with another two sequenced legume genomes, *Lotus japonicus* and *Glycine max*, which shows significant macrosyteny between these species. Sometimes the syntenic blocks are as large as entire chromosome arms. The resulting genome sequence of *M. truncatula* is not only used for the genome comparison to the species with known genetic and genomic information, but also for comparison with species with complex genetics and limited genomic tools such as *Medicago sativa* and pea (*Pisum sativum*). Thus, the *M. truncatula*

genome sequence provides great opportunities to expand knowledge about the genomes of other species.

In chapter 3, we describe how the *M. truncatula* genome sequencing project was supported by fluorescent *in situ* hybridization (FISH) experiments. The *M. truncatula* genome sequencing project aims at sequencing the euchromatic regions of all 8 chromosomes. We estimated sequence coverage of the euchromatic sequence using FISH. We have developed and improved FISH techniques suitable for cytogenetic mapping of BAC clones on *M. truncatula* pachytene chromosomes. The most distal telomere end BACs and pericentromeric heterochromatin border BACs of each chromosome were mapped. Based on physical measurements of BAC positions on each chromosome, we calculate that 94% of the euchromatin is located between terminal BACs of chromosome sequence assemblies, but gaps remain within the assembly. Mapping of gap sizes indicates that more than 99% of *M. truncatula* chromosome 5 has been covered and assembled. In contrast to chromosome 5, chromosome 6 still has some large gaps that cannot be closed even by expanding BAC contigs. We have identified the approximately physical positions of these large gaps on chromosome 6, and provide estimates of gap sizes. Finally, we used FISH to resolve inconsistencies between the sequence and the genetic map regarding the correct assembly of ends of chromosomes 4 and 8.

The genome sequence of *M. truncatula* can be used as a reference in studies of other legume genomes. The *Trifolium* genus contains several economically and agriculturally important crops, like red clover, white clover and other cultivated clover species. *M. truncatula* is the phylogenetically closest sequenced species to the *Trifolium* genus, which both belong to the *Trifolieae* tribe. The origin is dated 24.7+/-2.3 million years ago, long after the divergence of *L. japonicus* and soybean ~50 and ~54 million years ago, respectively. Therefore the *M. truncatula* genome is most suited as reference to study *Trifolium* genus. We performed a comparative genomic study. An important crop in *Trifolium* is red clover (*T. pratense*). It has relatively small diploid genome of ~430-470 Mb divided over 7 chromosomes. In chapter 3, we describe a cytogenetic map of red clover based on a DAPI stained pachytene karyogram, in which all 7 pachytene chromosomes can be identified. We measure pachytene chromosome lengths and describe the centromere, heterochromatin structures, and positions of rDNA loci and telomeric tandem repeat sequences. A consistent divergence in length of the telomeric repeat sequences between the two arms of each chromosome is found, showing FISH signal on only one end of each pachytene chromosome. This difference in length of telomeric repeats can be used to orient chromosomes, which is otherwise hard due to their median centromere positions. With well characterized pachytene complements, we are able to further integrate the genetic map into the cytogenetic map of red clover. The genetic map of red clover has been constructed based on 143 genetic markers mapped onto seven linkage groups. We identify BACs containing these markers and use these as probes in FISH experiments. Thus, the correlation between genetic linkage groups and chromosomes is determined.

The integrated genetic and cytogenetic map can be used in comparative genomic studies using the legume model species *M. truncatula* as reference, which we describe in chapter 5. In this chapter we study the evolutionary history of the *Trifolium* genus through a genome comparison with *M. truncatula*. We use the sequences covering the short and long arms of *M. truncatula* 5 as probes to hybridize on red clover and other *Trifolium* species. The result reveals that more extensive chromosomal rearrangements have occurred in red clover than other *Trifolium* species. We identify a red clover specific chromosomal rearrangement that is associated with a red clover-specific repetitive sequence. This repetitive sequence is identified as a LINE-like non-LTR retrotransposon through sequencing of the breakpoint region.

Finally, we discuss the future perspectives of the FISH technique in modern genome sequencing methods. We describe the possibilities offered by FISH to validate and correct sequencing, assembly and annotation problems, and discuss possible improvements to increase the throughput of FISH. We conclude that FISH is still a valuable independent tool for verification of sequencing results, and for performing genomic research in species without having to sequence the genome. The research described in this thesis has helped to advance legume genomics both through sequencing of a reference genome and through comparative studies between several legume species in the genus *Trifolium*. It thus provides important resources to the legume research community, which we expect will lead to many exciting new studies in the future.

Samenvatting

De Vlinderbloemigen-familie bevat een aantal gewassen die een belangrijke bron van eiwitten en lipiden vormen voor zowel mens als dier. Meer dan 30% van de eiwitbehoefte voor menselijke consumptie en meer dan 35% van de industriële plantaardige olie wordt geleverd door Vlinderbloemige-gewassen. Daarnaast worden Vlinderbloemigen ook gebruikt als groenbemesting, dit door hun unieke vermogen een symbiose aan te kunnen gaan met stikstofbindende Rhizobium bacteriën. In samenwerking met Vlinderbloemige-planten zijn deze bacteriën in staat atmosferische stikstof om te zetten in ammonium, dat vervolgens door de plant gebruikt kan worden. Dit proces van stikstoffixatie is de belangrijkste biologische bron van organische stikstofverbindingen. Wetenschappelijk onderzoek aan Vlinderbloemige planten is daarom ook voor een belangrijk deel gewijd aan deze stikstofbindende symbiose.

Moleculair genetische studies vereisen genoom informatie. *Medicago truncatula* is geselecteerd als een Vlinderbloemige-modelplant. In hoofdstuk 2 beschrijven we de de sequentie van het euchromatische deel van het *M. truncatula* genoom. Om deze te verkrijgen is een gecombineerde strategie gebruikt van een BAC-by-BAC aanpak in combinatie met next-generation sequencing. De zo verkregen genoomsequentie representeert ~94% van alle *M. truncatula* genen. Dit biedt een waardevol inzicht in de onderliggende genoom-informatie van vlinderbloemige specifieke eigenschappen. Analyse van de sequentie heeft aangetoond dat er ongeveer 58 miljoen jaar geleden, kort na het ontstaan van de vlinderbloemige familie, een genoom verdubbeling is opgetreden. Door het genoom van *M. truncatula* te vergelijken met het genoom van twee andere vlinderbloemigen, namelijk *Lotus japonicus* en soja (*Glycine max*), is een aanzienlijk niveau van geconserveerde volgorde van genen (macrosyteny) tussen deze soorten zichtbaar gemaakt. In enkele gevallen omspant de geconserveerde regio een volledig chromosoom arm. Het *M. truncatula* genoom wordt niet alleen gebruikt in vergelijking met andere soorten met gesequenced genoom, maar ook voor de vergelijking met soorten met complexere genetica of beperkte genomische tools zoals lucerne (*Medicago sativa*) en erwt (*Pisum sativum*). Het *M. truncatula* genoom kan daarom gezien worden als referentie voor de Vlinderbloemige-plantenfamilie.

In hoofdstuk 3 wordt beschreven hoe het *M. truncatula* sequencing-project werd ondersteund door middel van fluorescentie *in situ* hybridisatie (FISH) experimenten. Het sequencing project richtte zich op de euchromatische regio's van alle 8 chromosomen. Met behulp van FISH is een schatting gemaakt van de genoom coverage die is bereikt in het sequencing project. Hier voor is het FISH protocol voor *M. truncatula* pachytene chromosomen verder geoptimaliseerd. Om inzicht te krijgen in de grootte van het euchromatische deel van het genoom zijn merker BAC clones gezocht die specifiek zijn voor het distale uiteinde van elk van de 16 chromosoom-armen, alswel voor de grensgebieden met het pericentromerisch heterochromatine. Deze merker BACs zijn gebruikt als probe in FISH experimenten op pachytene chromosomen. Voor elk chromosoom-arm is de fysieke afstand tussen de merker BACs gemeten. Op basis hiervan is berekend dat 94% van het feitelijke euchromatine wordt afgebakend door de door ons geïdentificeerde merker BACs. Deze merkers zijn ook gepositioneerd op de genetische kaart. Echter in de genoom sequentie van dit euchromatische deel bevinden zich ook nog niet ontrafelde delen. Hiervoor is nog geen sequentie voor bepaald. Wij hebben een aantal chromosomen geanalyseerd om inzicht te krijgen in de grootte van deze zogenaamde gaps in de genoomsequentie. Voor *M. truncatula* chromosoom 5 blijken de 7 aanwezige gaps relatief klein, een de gegenereerde sequentie omspant dan ook meer dan 99% van beide chromosoom-armen. In tegenstelling tot chromosoom 5, heeft chromosoom 6 nog steeds een aantal aanzienlijke gaps die niet gecoverd worden door de gegenereerde sequentie. Door middel van FISH is de grootte en fysieke posities van deze gaps bepaald. Tot slot is FISH gebruikt om inconsistenties tussen de genetische en fysieke kaart van chromosomen 4 en 8 op te lossen.

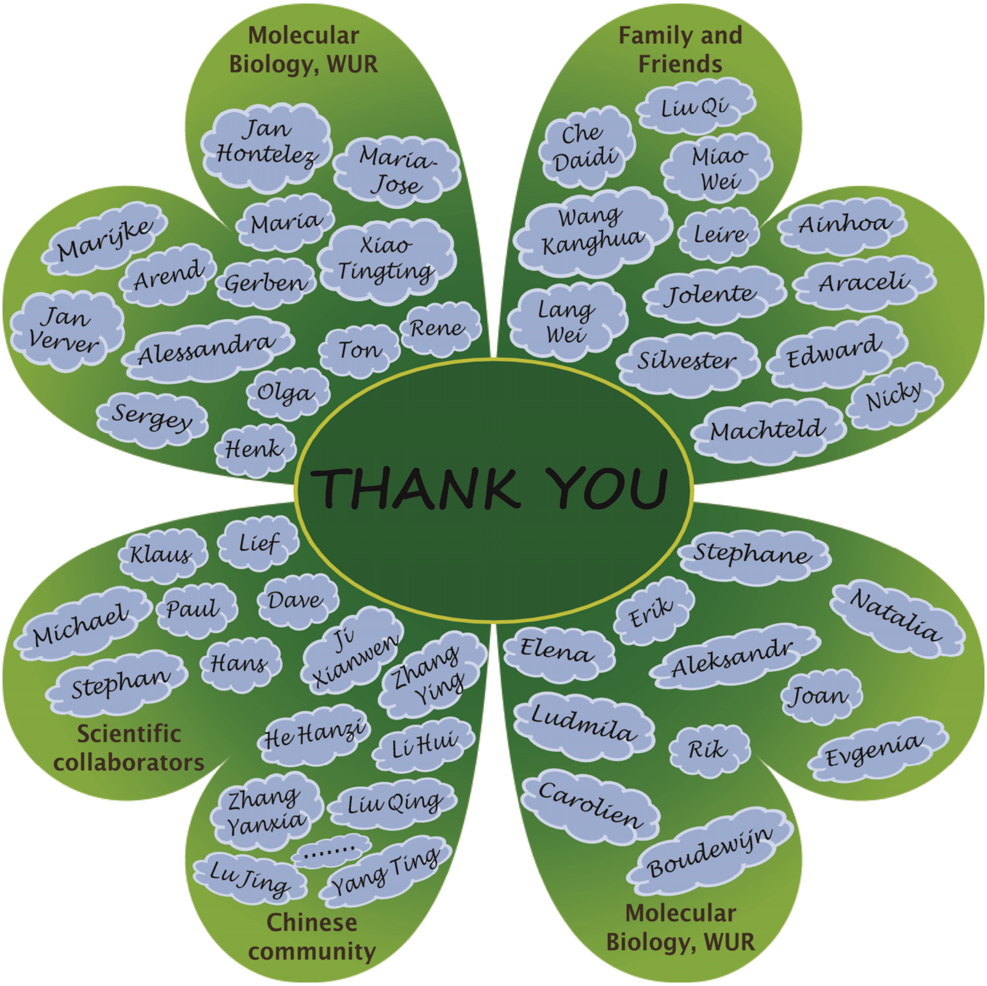
De genoomsequentie van *M. truncatula* kan worden gebruikt als een referentie in studies met andere Vlinderbloemige-genomen. Het geslacht *Trifolium* bevat een aantal economisch en landbouwkundig belangrijke gewassen, zoals rode en witte klaver, en andere gecultiveerde klaversoorten. Fylogenetisch gezien is, van de soorten met een gesequenced genoom, *M. truncatula* het nauwst verwant aan het *Trifolium* geslacht. Beide maken deel uit van de Trifolieeae groep. De oorsprong van de Trifolieeae groep is gedateerd 24,7±2,3 jaar geleden, wat veel recenter is dan de splitsing met andere model-Vlinderbloemigen zoals L. japonicus en soja die 50 en 54 miljoen jaar geleden zijn gedivergeerd. Daarom is van deze drie model-Vlinderbloemigen, het *M. truncatula* genoom het meest geschikt als referentie voor studies aan *Trifolium* soorten. Wij hebben een vergelijkende studie uitgevoerd tussen *M. truncatula* en rode klaver (*Trifolium pratense*). Rode klaver heeft een relatief klein diploïd genoom van ~ 430 tot 470 Mb verdeeld over 7 chromosomen. In hoofdstuk 3 beschrijven we de constructie van een cytogenetische kaart van rode klaver op basis van een DAPI gekleurde pachytene karyogram, waarin alle zeven pachytene chromosomen worden geïdentificeerd. We hebben de lengte van de pachytene chromosoom gemeten en beschrijven de centromeerpositie, heterochromatine structuren, posities van rDNA loci en telomeer tandem repeat sequenties. Opmerkelijk is de vinding dat er een consistente divergentie is in de lengte van de telomeer repeat cluster tussen beide uiteinden van van elk individueel chromosoom. FISH met de een geconserveerde telomeer probe geeft slechts signaal op één uiteinde van elke pachytene chromosoom. Dit verschil in lengte van de telomeer repeat cluster kan gebruikt worden om chromosomen oriënteren. Dit is met name handig voor chromosomen met een mediaane centromeer positie. Met goed gekarakteriseerde pachytene chromosomen is het vervolgens mogelijk om de genetische kaart en de cytogenetische kaart van rode klaver te integreren. De genetische kaart van rode klaver is gebouwd op basis van 143 genetische merkers verspreid over zeven linkage-groepen. Om tot integratie van beide kaarten te komen zijn BAC clones gebruikt als probe in FISH studies die ook verankerd zijn op de genetische kaart.

De geïntegreerde genetische en cytogenetische kaart van rode klaver kan gebruikt worden in vergelijkende genomische studies met het *M. truncatula* als referentie. In hoofdstuk 5 beschrijven we zo'n studie. In dit hoofdstuk bestuderen we de evolutionaire geschiedenis van het *Trifolium* geslacht door middel van een genoom vergelijking met *M. truncatula*. Daarbij richten we ons op *M. truncatula* chromosoom 5. BAC clones die de korte en lange arm van dit chromosoom representeren zijn gebruikt als probes in FISH studies met rode klaver en andere *Trifolium* soorten. Deze studie laat zien dat aanzienlijk meer chromosomale herschikkingen hebben plaatsgevonden in rode klaver dan in vergelijking met andere *Trifolium* soorten. Door in te zoomen op een chromosomaalbreekpunt in rode klaver is er een, voor deze soort, specifieke repetitieve sequentie geïdentificeerd. Deze sequentie heeft homologie met een LINE none-LTR retrotransposon.

In het laatste hoofdstuk bespreek ik de perspectieven van FISH technologie voor moderne (next generation) genoom sequencing methoden. Ik beschrijf de mogelijkheden om door middel van FISH genoom-assemblies te valideren en mogelijke manieren om de throughput te verbeteren. Ik concludeer dat FISH nog steeds een waardevol hulpmiddel is voor onafhankelijke verificatie van sequencing resultaten en voor het uitvoeren van genoom onderzoek in soorten waarvan geen genoom-sequentie bekend is.

Het onderzoek beschreven in dit proefschrift heeft bijgedragen aan het genoom research bij vlinderbloemige-soorten. Dit is gedaan door middel van ondersteuning van een groot internationaal programma dat als doel had het ontrafelen van het genoom van de modelplant *M. truncatula* als wel door middel van vergelijkende studies tussen klaversoorten en het *M. truncatula* referentiegenoom. Het uitgevoerde onderzoek omvat dus een belangrijke bijdrage aan vlinderbloemige-onderzoeksgeneemschap wat hopelijk zal leiden tot veel opwindende toekomstige studies.

Acknowledgements



Curriculum vitae

Chunting Lang was born on 14 February 1981, in Qiqihar, China. She has attended a bachelor study in horticulture in Northeast Agricultural University, Harbin, China, in 1999. She graduated in 2003, with a specialization in architecture and landscape designing. After working at Qiqihar University and a medicinal herb farm, she started her master study in Wageningen University in the department of plant breeding in 2005. She performed her thesis research project at the cytogenetics group of prof. dr. Hans de Jong, which has resulted in a publication. Afterwards, she spent her internship in the same group in cooperation with Rijk Zwaan breeding company. In 2007, she successfully completed the MSc curriculum, specializing in “plant breeding and genetic resources”.

After finishing her master, she decided to continue as PhD student in the laboratory of molecular biology, Wageningen University, which resulted in this thesis. In October 2011 Chunting started working as a Kalanchoë breeder in the breeding company Fides.

References

Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., Debarry, J., Arondel, V., Ohlrogge, J., Saie, I.J., Suliman-Elmeer, K.M., Bennetzen, J.L., Kruegger, R.R., and Malek, J.A. (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat Biotechnol **29,** 521-527.

Ammiraju, J.S.S., Luo, M.Z., Goicoechea, J.L., Wang, W.M., Kudrna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N.B., Blackmon, B., Fang, E., Tomkins, J.B., Brar, D., MacKill, D., McCouch, S., Kurata, N., Lambert, G., Galbraith, D.W., Arumuganathan, K., Rao, K.R., Walling, J.G., Gill, N., Yu, Y., SanMiguel, P., Soderlund, C., Jackson, S., and Wing, R.A. (2006). The Oryza bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. Genome Research **16,** 140-147.

Ansari, H.A., Ellison, N.W., Reader, S.M., Badaeva, E.D., Friebe, B., Miller, T.E., and Williams, W.M. (1999). Molecular Cytogenetic Organization of 5S and 18S-26S rDNA Loci in White Clover (*Trifolium repens L.*) and Related Species. Annals of Botany **83,** 199-206.

Arrighi, J.F., Barre, A., Ben Amor, B., Bersoult, A., Soriano, L.C., Mirabella, R., de Carvalho-Niebel, F., Journet, E.P., Gherardi, M., Huguet, T., Geurts, R., Denarie, J., Rouge, P., and Gough, C. (2006). The Medicago truncatula lysin [corrected] motif-receptor-like kinase gene family includes NFP and new nodule-expressed genes. Plant Physiol **142,** 265-279.

Arumuganathan, K., and Earle, E.D. (1991). Nuclear DNA Content of Some Important Plant Species. Plant Molecular Biology Reporter **9,** 11.

Badaeva, E.D., Friebe, B., and Gill, B.S. (1996). Genome differentiation in Aegilops .1. Distribution of highly repetitive DNA sequences on chromosomes of diploid species. Genome **39,** 293-306.

Balmus, G., Trifonov, V.A., Biltueva, L.S., O'Brien, P.C., Alkalaeva, E.S., Fu, B., Skidmore, J.A., Allen, T., Graphodatsky, A.S., Yang, F., and Ferguson-Smith, M.A. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative Cetartiodactyla ancestral karyotype. Chromosome Res **15,** 499-515.

Bancroft, I. (2001). Duplicate and diverge: the evolution of plant genome microstructure. Trends Genet **17,** 89-93.

Beck, V.B., Rohr, U., and Jungbauer, A. (2005). Phytoestrogens derived from red clover: An alternative to estrogen replacement therapy? Journal of Steroid Biochemistry and Molecular Biology **94,** 499-518.

Bennett, M.D., and Smith, J.B. (1976). Nuclear dna amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci **274,** 227-274.

Bennetzen, J.L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115,** 29-36.

Berg, D.E., and Howe, M.M. (1989). Mobile DNA. (Washington: American Society for Microbiology).

Berr, A., Pecinka, A., Meister, A., Kreth, G., Fuchs, J., Blattner, F.R., Lysak, M.A., and Schubert, I. (2006). Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between Arabidopsis thaliana and Arabidopsis lyrata. Plant J **48,** 771-783.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12,** 1093-1101.

Bohn, M., Heermann, D.W., and van Driel, R. (2007). Random loop model for long polymers. Phys Rev E Stat Nonlin Soft Matter Phys **76,** 051805.

Brook, G.M. (1991). Chloroplast DNA diversity within and among populations of *Trifolium pratense*. Current Genetics **19,** 411-416.

Burr, B., Burr, F.A., Matz, E.C., and Romero-Severson, J. (1992). Pinning down loose ends: mapping telomeres and factors affecting their length. Plant Cell **4,** 953-960.

Campell, B.R., Song, Y.G., Posch, T.E., Cullis, C.A., and Town, C.D. (1992). Sequence and Organization of 5s Ribosomal Rna-Encoding Genes of Arabidopsis-Thaliana. Gene **112,** 225-228.

Cannon, S.B., Ilut, D., Farmer, A.D., Maki, S.L., May, G.D., Singer, S.R., and Doyle, J.J. (2010). Polyploidy did not predate the evolution of nodulation in all legumes. PLoS One **5,** e11630.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18,** 188-196.

Casals, F., and Navarro, A. (2007). Chromosomal evolution: inversions: the chicken or the egg? Heredity **99,** 479-480.

Casals, F., Caceres, M., and Ruiz, A. (2003). The foldback-like transposon Galileo is involved in the generation of two different natural chromosomal inversions of Drosophila buzzatii. Mol Biol Evol **20,** 674-685.

Chang, S.B., Anderson, L.K., Sherman, J.D., Royer, S.M., and Stack, S.M. (2007). Predicting and testing physical locations of genetically mapped loci on tomato pachytene chromosome 1. Genetics **176,** 2131-2138.

Chang, S.B., Yang, T.J., Datema, E., van Vugt, J., Vosman, B., Kuipers, A., Meznikova, M., Szinay, D., Lankhorst, R.K., Jacobsen, E., and de Jong, H. (2008). FISH mapping and molecular organization of the major repetitive sequences of tomato. Chromosome Res **16,** 919-933.

Cheng, Z., Buell, C.R., Wing, R.A., Gu, M., and Jiang, J. (2001). Toward a cytological characterization of the rice genome. Genome Res **11,** 2133-2141.

Choi, H.K., Mun, J.H., Kim, D.J., Zhu, H., Baek, J.M., Mudge, J., Roe, B., Ellis, N., Doyle, J., Kiss, G.B., Young, N.D., and Cook, D.R. (2004a). Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci U S A **101,** 15289-15294.

Choi, H.K., Kim, D., Uhm, T., Limpens, E., Lim, H., Mun, J.H., Kalo, P., Penmetsa, R.V., Seres, A., Kulikova, O., Roe, B.A., Bisseling, T., Kiss, G.B., and Cook, D.R. (2004b). A sequence-based genetic map of Medicago truncatula and comparison of marker colinearity with M. sativa. Genetics **166,** 1463-1502.

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. Genome Biol **10,** R108.

Cirera, S., Martin-Campos, J.M., Segarra, C., and Aguade, M. (1995). Molecular characterization of the breakpoints of an inversion fixed between Drosophila melanogaster and D. subobscura. Genetics **139,** 321-326.

Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. (2001). Mobile DNA II. (Washington: American Society for Microbiology).

Cremer, T., Lichter, P., Borden, J., Ward, D.C., and Manuelidis, L. (1988). Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. Hum Genet **80,** 235-246.

Dassanayake, M., Oh, D.H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.J., Bressan, R.A., Zhu, J.K., Bohnert, H.J., and Cheeseman, J.M. (2011). The genome of the extremophile crucifer Thellungiella parvula. Nat Genet **43,** 913-918.

de Jong, J.H., Fransz, P., and Zabel, P. (1999). High resolution FISH in plants - techniques and applications. Trends in Plant Science **4,** 258-263.

Delseny, M. (2004). Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement. Curr Opin Plant Biol **7,** 126-131.

Deragon, J.M., Gilbert, N., Rouquet, L., Lenoir, A., Arnaud, P., and Picard, G. (1996). A transcriptional analysis of the S1Bn (Brassica napus) family of SINE retroposons. Plant Mol Biol **32,** 869-878.

Doyle, J.J., and Luckow, M.A. (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol **131,** 900-910.

Dubcovsky, J., and Dvorak, J. (1995). Ribosomal-Rna Multigene Loci - Nomads of the Triticeae Genomes. Genetics **140,** 1367-1377.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L., Shiloff, B.A., and Bennetzen, J.L. (2001). Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol **125,** 1342-1353.

Ellison, N.W., Liston, A., Steiner, J.J., Williams, W.M., and Taylor, N.L. (2006). Molecular phylogenetics of the clover genus (Trifolium--Leguminosae). Mol Phylogenet Evol **39,** 688-705.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., and Weinstock, G.M. (2007). Creating a honey bee consensus gene set. Genome Biol **8,** R13.

Fajkus, J., Kovarik, A., Kralovics, R., and Bezdek, M. (1995). Organization of telomeric and subtelomeric chromatin in the higher plant Nicotiana tabacum. Mol Gen Genet **247,** 633-638.

Farrar, K., Asp, T., Lubberstedt, T., Xu, M.L., Thomas, A.M., Christiansen, C., Humphreys, M.O., and Donnison, I.S. (2007). Construction of two Lolium perenne BAC libraries and identification of BACs containing candidate genes for disease resistance and forage quality. Molecular Breeding **19,** 15-23.

Ferguson-Smith, M.A. (1997). Genetic analysis by chromosome sorting and painting: phylogenetic and diagnostic applications. Eur J Hum Genet **5,** 253-265.

Ferguson-Smith, M.A., and Trifonov, V. (2007). Mammalian karyotype evolution. Nat Rev Genet **8,** 950-962.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. Trends Genet **5,** 103-107.

Flagel, L.E., and Wendel, J.F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol **186,** 184-193.

Flavell, A.J., Pearce, S.R., and Kumar, A. (1994). Plant transposable elements and the genome. Curr Opin Genet Dev **4,** 838-844.

Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H. (2000). Integrated cytogenetic map of chromosome arm 4S of A. thaliana: structural organization of heterochromatic knob and centromere region. Cell **100,** 367-376.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16,** 805-814.

Gamas, P., Debellé, F., Hélène Berges, Godiard, L., Niebel, A., Journet, E.-P., and Gouzy, J. (2006). *Medicago truncatula* cDNA and genomic libraries.

Ganal, M.W., Lapitan, N.L., and Tanksley, S.D. (1991). Macrostructure of the tomato telomeres. Plant Cell **3,** 87-94.

Gardiner, J.M., Coe, E.H., and Chao, S. (1996). Cloning maize telomeres by complementation in Saccharomyces cerevisiae. Genome **39,** 736-748.

George, J., Sawbridge, T.I., Cogan, N.O., Gendall, A.R., Smith, K.F., Spangenberg, G.C., and Forster, J.W. (2008). Comparison of genome structure between white clover and Medicago truncatula supports homoeologous group nomenclature based on conserved synteny. Genome **51,** 905-911.

Gerlach, W.L., and Bedbrook, J.R. (1979). Cloning and Characterization of Ribosomal-Rna Genes from Wheat and Barley. Nucleic Acids Research **7,** 1869-1885.

Gill, K.S., Gill, B.S., Endo, T.R., and Boyko, E.V. (1996a). Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics **143,** 1001-1012.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. (1996b). Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics **144,** 1883-1891.

Gillett, J.M., and Taylor, N.L. (2001). The world of clovers. (Ames : Iowa State University Press, Ames, Iowa, USA).

Gonzalez, J., Ranz, J.M., and Ruiz, A. (2002). Chromosomal elements evolve at different rates in the Drosophila genome. Genetics **161,** 1137-1154.

Graham, P.H., and Vance, C.P. (2003). Legumes: Importance and Constraints to Greater Use. Plant Physiology **131,** 872–877.

Grant, D., Cregan, P., and Shoemaker, R.C. (2000). Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci U S A **97,** 4168-4173.

Hand, M.L., Cogan, N.O., Sawbridge, T.I., Spangenberg, G.C., and Forster, J.W. (2010). Comparison of homoeolocus organisation in paired BAC clones from white clover (*Trifolium repens L.*) and microcolinearity with model legume species. BMC Plant Biol **10,** 94.

Harper, M.E., Ullrich, A., and Saunders, G.F. (1981). Localization of the human insulin gene to the distal end of the short arm of chromosome 11. Proc Natl Acad Sci U S A **78,** 4458-4460.

Hegnauer, R., and Grayer-Barkmeijer, R.J. (1993). Relevance of seed polysaccharides and flavonoids for the classification of the leguminosae: A chemotaxonomic approach. Phytochemistry **34,** 3-16.

Herendeen, P.S., Crepet, W.L., and Dilcher, D.L. (1992). The fossil history of the Leguminosae: phylogenetic and biogeographical implications. (Royal Botanic Gardens, Kew, UK).

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D., and Guo, Y.L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet **43,** 476-481.

Huang, B. (2010). Super-resolution optical microscopy: multiple choices. Curr Opin Chem Biol **14,** 10-14.

Huang, B., Bates, M., and Zhuang, X. (2009a). Super-resolution fluorescence microscopy. Annu Rev Biochem **78,** 993-1016.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E.A., Wu, Y., Guo, J., He, J., Jia, Z., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W.K., Kim, J.Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Fang, L., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Li, S., Zhang, X., Wang, J., Sun, R., Zhang, B., Jiang, S., and Du, Y. (2009b). The genome of the cucumber, Cucumis sativus L. Nat Genet **41,** 1275-1281.

Isobe, S., Klimenko, I., Ivashuta, S., Gau, M., and Kozlov, N.N. (2003). First RFLP linkage map of red clover (*Trifolium pratense L.*) based on cDNA probes and its transferability to other red clover germplasm. Theor Appl Genet **108,** 105-112.

Isobe, S., Kolliker, R., Hisano, H., Sasamoto, S., Wada, T., Klimenko, I., Okumura, K., and Tabata, S. (2009). Construction of a consensus linkage map for red clover (*Trifolium pratense L.*). BMC Plant Biol **9,** 57.

Jiang, J., and Gill, B.S. (1994). Nonisotopic in situ hybridization and plant genome mapping: the first 10 years. Genome **37,** 717-725.

Kalo, P., Seres, A., Taylor, S.A., Jakab, J., Kevei, Z., Kereszt, A., Endre, G., Ellis, T.H., and Kiss, G.B. (2004). Comparative mapping between *Medicago sativa* and *Pisum sativum*. Mol Genet Genomics **272,** 235-246.

Kato, T., Kawashima, K., Miwa, M., Mimura, Y., Tamaoki, M., Kouchi, H., and Suganuma, N. (2002). Expression of genes encoding late nodulins characterized by a putative signal peptide and conserved cysteine residues is reduced in ineffective pea nodules. Mol Plant Microbe Interact **15,** 129-137.

Kazimierski, T., Kazimierska, E.M., and Strzyzewska, C. (1972). Species crossing in the genus *Trifolium*. Genetica Polonica **13,** 11-32.

Kennedy, A.R. (1995). The evidence for soybean products as cancer preventive agents. J Nutr **125,** 733S-743S.

Kistner, C., and Parniske, M. (2002). Evolution of signal transduction in intracellular symbiosis. Trends Plant Sci **7,** 511-518.

Kolliker, R., Herrmann, D., Boller, B., and Widmer, F. (2003). Swiss Mattenklee landraces, a distinct and diverse genetic resource of red clover (*Trifolium pratense L.*). Theoretical and Applied Genetics **107,** 306-315.

Kongkiatngam, P., Waterway, M.J., Fortin, M.G., and Coulman, B.E. (1995). Genetic-Variation within and between 2 Cultivars of Red-Clover (*Trifolium-Pratense L.*) - Comparisons of Morphological, Isozyme, and Rapd Markers. Euphytica **84,** 237-246.

Koo, D.H., Jo, S.H., Bang, J.W., Park, H.M., Lee, S., and Choi, D. (2008). Integration of cytogenetic and genetic linkage maps unveils the physical architecture of tomato chromosome 2. Genetics **179,** 1211-1220.

Ku, H.M., Vision, T., Liu, J., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A **97,** 9121-9126.

Kubis, S.E., Heslop-Harrison, J.S., Desel, C., and Schmidt, T. (1998). The genomic organization of non-LTR retrotransposons (LINEs) from three *Beta* species and five other angiosperms. Plant Mol Biol **36,** 821-831.

Kulikova, O., Gualtieri, G., Geurts, R., Kim, D.J., Cook, D., Huguet, T., de Jong, J.H., Fransz, P.F., and Bisseling, T. (2001). Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. Plant J **27,** 49-58.

Kulikova, O., Geurts, R., Lamine, M., Kim, D.J., Cook, D.R., Leunissen, J., de Jong, H., Roe, B.A., and Bisseling, T. (2004). Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. Chromosoma **113,** 276-283.

Landegent, J.E., Jansen in de Wal, N., van Ommen, G.J., Baas, F., de Vijlder, J.J., van Duijn, P., and Van der Ploeg, M. (1985). Chromosomal localization of a unique gene by non-autoradiographic in situ hybridization. Nature **317,** 175-177.

Langer-Safer, P.R., Levine, M., and Ward, D.C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci U S A **79,** 4381-4385.

Lavin, M., Herendeen, P.S., and Wojciechowski, M.F. (2005). Evolutionary rates analysis of *Leguminosae* implicates a rapid diversification of lineages during the tertiary. Syst Biol **54,** 575-594.

Leeton, P.R., and Smyth, D.R. (1993). An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol Gen Genet **237,** 97-104.

Lewis, G., Schrire, B., Mackinder, B., and Lock, M. (2005). Legumes of the world. (Royal Botanic Gardens, Kew, UK).

Liehr, T., Starke, H., Weise, A., Lehrer, H., and Claussen, U. (2004). Multicolor FISH probe sets and their applications. Histol Histopathol **19,** 229-237.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Lysak, M.A., Fransz, P.F., Ali, H.B., and Schubert, I. (2001). Chromosome painting in Arabidopsis thaliana. Plant J **28,** 689-697.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related *Brassicaceae* species. Proc Natl Acad Sci U S A **103,** 5224-5229.

Maillet, G., White, C.I., and Gallego, M.E. (2006). Telomere-length regulation in inter-ecotype crosses of Arabidopsis. Plant Molecular Biology **62,** 859-866.

Mandakova, T., and Lysak, M.A. (2008). Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (*Brassicaceae*). Plant Cell **20,** 2559-2570.

McDonald, J.F., Matyunina, L.V., Wilson, S., Jordan, I.K., Bowen, N.J., and Miller, W.J. (1997). LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica **100,** 3-13.

McKnight, T.D., and Shippen, D.E. (2004). Plant telomere biology. Plant Cell **16,** 794-803.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell **15,** 809-834.

Middleton, P.H., Jakab, J., Penmetsa, R.V., Starker, C.G., Doll, J., Kalo, P., Prabhu, R., Marsh, J.F., Mitra, R.M., Kereszt, A., Dudas, B., VandenBosch, K., Long, S.R., Cook, D.R., Kiss, G.B., and Oldroyd, G.E. (2007). An ERF transcription factor in *Medicago truncatula* that is essential for Nod factor signal transduction. Plant Cell **19,** 1221-1234.

Mochizuki, K., Umeda, M., Ohtsubo, H., and Ohtsubo, E. (1992). Characterization of a plant SINE, p-SINE1, in rice genomes. Jpn J Genet **67,** 155-166.

Molteni, A., Brizio-Molteni, L., and Persky, V. (1995). In vitro hormonal effects of soybean isoflavones. J Nutr **125,** 751S-756S.

Mun, J.H., Kim, D.J., Choi, H.K., Gish, J., Debelle, F., Mudge, J., Denny, R., Endre, G., Saurat, O., Dudez, A.M., Kiss, G.B., Roe, B., Young, N.D., and Cook, D.R. (2006). Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. Genetics **172,** 2541-2555.

Mun, J.H., Kwon, S.J., Yang, T.J., Seol, Y.J., Jin, M., Kim, J.A., Lim, M.H., Kim, J.S., Baek, S., Choi, B.S., Yu, H.J., Kim, D.S., Kim, N., Lim, K.B., Lee, S.I., Hahn, J.H., Lim, Y.P., Bancroft, I., and Park, B.S. (2009). Genome-wide comparative analysis of the Brassica rapa gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol **10,** R111.

Nagaki, K., Kashihara, K., and Murata, M. (2005). Visualization of diffuse centromeres with centromere-specific histone H3 in the holocentric plant Luzula nivea. Plant Cell **17,** 1886-1893.

Nayak, S.N., Zhu, H., Varghese, N., Datta, S., Choi, H.K., Horres, R., Jungling, R., Singh, J., Kishor, P.B., Sivaramakrishnan, S., Hoisington, D.A., Kahl, G., Winter, P., Cook, D.R., and Varshney, R.K. (2010). Integration of novel SSR and gene-based SNP marker loci in the chickpea genetic map and establishment of new anchor points with *Medicago truncatula* genome. Theor Appl Genet **120,** 1415-1441.

Ohmido, N., Kijima, K., Ashikawa, I., de Jong, J.H., and Fukui, K. (2001). Visualization of the terminal structure of rice chromosomes 6 and 12 with multicolor FISH to chromosomes and extended DNA fibers. Plant Mol Biol **47,** 413-421.

Oldroyd, G.E., and Downie, J.A. (2008). Coordinating nodule morphogenesis with rhizobial infection in legumes. Annu Rev Plant Biol **59,** 519-546.

Op den Camp, R., Streng, A., De Mita, S., Cao, Q., Polone, E., Liu, W., Ammiraju, J.S., Kudrna, D., Wing, R., Untergasser, A., Bisseling, T., and Geurts, R. (2011). LysM-type mycorrhizal receptor recruited for rhizobium symbiosis in nonlegume Parasponia. Science **331,** 909-912.

Pardue, M.L., and Gall, J.G. (1975). Nucleic acid hybridization to the DNA of cytological preparations. Methods Cell Biol **10,** 1-16.

Parra, I., and Windle, B. (1993). High resolution visual mapping of stretched DNA by fluorescent hybridization. Nat Genet **5,** 17-21.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F., Messing, J., and Rokhsar, D.S. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457,** 551-556.

Pedrosa, A., Vallejos, C.E., Bachmair, A., and Schweizer, D. (2003). Integration of common bean (*Phaseolus vulgaris L.*) linkage and chromosomal maps. Theoretical and Applied Genetics **106,** 205-212.

Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D., and Bachmair, A. (2002). Chromosomal map of the model legume Lotus japonicus. Genetics **161,** 1661-1672.

Peters, N.K., Frost, J.W., and Long, S.R. (1986). A plant flavone, luteolin, induces expression of Rhizobium meliloti nodulation genes. Science **233,** 977-980.

Peters, S.A., Datema, E., Szinay, D., van Staveren, M.J., Schijlen, E.G., van Haarst, J.C., Hesselink, T., Abma-Henkens, M.H., Bai, Y., de Jong, H., Stiekema, W.J., Klein Lankhorst, R.M., and van Ham, R.C. (2009). *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. Plant J **58,** 857-869.

Peterson, D.G., Stack, S.M., Price, H.J., and Johnston, J.S. (1996). DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. Genome **39,** 77-82.

Pfeil, B.E., Schlueter, J.A., Shoemaker, R.C., and Doyle, J.J. (2005). Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol **54,** 441-454.

Pich, U., and Schubert, I. (1998). Terminal heterochromatin and alternative telomeric sequences in *Allium cepa*. Chromosome Res **6,** 315-321.

Polhill, R.M., Raven, P.H., and Stirton, C.H. (1981). Evolution and systematics of the *Leguminosae*. In Advances in legume systematics, P.R. RM Polhill, ed (Royal Botanic Gardens, Kew, UK), pp. 1-26.

Prado, E.A., FaivreRampant, P., Schneider, C., and Darmency, M.A. (1996). Detection of a variable number of ribosomal DNA loci by fluorescent *in situ* hybridization in Populus species. Genome **39,** 1020-1026.

Rabin, M., Uhlenbeck, O.C., Steffensen, D.M., and Mangel, W.F. (1984). Chromosomal sites of integration of simian virus 40 DNA sequences mapped by in situ hybridization in two transformed hybrid cell lines. J Virol **49,** 445-451.

Ranz, J.M., Casals, F., and Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res **11,** 230-239.

Raskina, O., Barber, J.C., Nevo, E., and Belyayev, A. (2008). Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. Cytogenet Genome Res **120,** 351-357.

Richards, E.J., and Ausubel, F.M. (1988). Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. Cell **53,** 127-136.

Romanenko, S.A., Sitnikova, N.A., Serdukova, N.A., Perelman, P.L., Rubtsova, N.V., Bakloushinskaya, I.Y., Lyapunova, E.A., Just, W., Ferguson-Smith, M.A., Yang, F., and Graphodatsky, A.S. (2007). Chromosomal evolution of *Arvicolinae* (*Cricetidae*, *Rodentia*). II. The genome homology of two mole voles (genus Ellobius), the field vole and golden hamster revealed by comparative chromosome painting. Chromosome Res **15,** 891-897.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics **16,** 944-945.

Sankoff, D. (2001). Gene and genome duplication. Curr Opin Genet Dev **11,** 681-684.

Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., Kaneko, T., Sakurai, N., Okumura, K., Klimenko, I., Sasamoto, S., Wada, T., Watanabe, A., Kohara, M., Fujishiro, T., and Tabata, S. (2005). Comprehensive structural analysis of the genome of red clover (*Trifolium pratense L.*). DNA Res **12,** 301-364.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., Fujishiro, T., Katoh, M., Kohara, M., Kishida, Y., Minami, C., Nakayama, S., Nakazaki, N., Shimizu, Y., Shinpo, S., Takahashi, C., Wada, T., Yamada, M., Ohmido, N., Hayashi, M., Fukui, K., Baba, T., Nakamichi, T., Mori, H., and Tabata, S. (2008). Genome structure of the legume, Lotus japonicus. DNA Res **15,** 227-239.

Saviranta, N.M., Anttonen, M.J., von Wright, A., and Karjalainen, R.O. (2008). Red clover (*Trifolium pratense L.*) isoflavones: determination of concentrations by plant stage, flower colour, plant part and cultivar. Journal of the Science of Food and Agriculture **88,** 125-132.

Schmidt, T. (1999). LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol **40,** 903-910.

Schmidt, T., Kubis, S., and Heslop-Harrison, J.S. (1995). Analysis and chromosomal localization of retrotransposons in sugar beet (*Beta vulgaris L.*): LINEs and Ty1-copia-like elements as major components of the genome. Chromosome Res **3,** 335-345.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., and Jackson, S.A. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463,** 178-183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee,

S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326,** 1112-1115.

Schwarz-Sommer, Z., Leclercq, L., Gobel, E., and Saedler, H. (1987a). Cin4, an insert altering the structure of the A1 gene in Zea mays, exhibits properties of nonviral retrotransposons. EMBO J **6,** 3873-3880.

Schwarz-Sommer, Z., Shepherd, N., Tacke, E., Gierl, A., Rohde, W., Leclercq, L., Mattes, M., Berndtgen, R., Peterson, P.A., and Saedler, H. (1987b). Influence of transposable elements on the structure and function of the A1 gene of Zea mays. EMBO J **6,** 287-294.

Sharma, S., and Raina, S.N. (2005). Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. Cytogenet Genome Res **109,** 15-26.

Shirley, B.W., Kubasek, W.L., Storz, G., Bruggemann, E., Koornneef, M., Ausubel, F.M., and Goodman, H.M. (1995). Analysis of Arabidopsis mutants deficient in flavonoid biosynthesis. Plant J **8,** 659-671.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P., Burns, P., Davis, T.M., Slovin, J.P., Bassil, N., Hellens, R.P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O.R., Jensen, R.V., Allan, A.C., Michael, T.P., Setubal, J.C., Celton, J.M., Rees, D.J., Williams, K.P., Holt, S.H., Ruiz Rojas, J.J., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S.A., Troggio, M., Viola, R., Ashman, T.L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H.D., Bryant, D.W., Jr., Fox, S.E., Givan, S.A., Wilhelm, L.J., Naithani, S., Christoffels, A., Salama, D.Y., Carter, J., Lopez

Girona, E., Zdepski, A., Wang, W., Kerstetter, R.A., Schwab, W., Korban, S.S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J.L., Salzberg, S.L., Dickerman, A.W., Velasco, R., Borodovsky, M., Veilleux, R.E., and Folta, K.M. (2011). The genome of woodland strawberry (Fragaria vesca). Nat Genet **43,** 109-116.

Singer, S.R., Maki, S.L., Farmer, A.D., Ilut, D., May, G.D., Cannon, S.B., and Doyle, J.J. (2009). Venturing beyond beans and peas: what can we learn from Chamaecrista? Plant Physiol **151,** 1041-1047.

Singh, R.J., Chung, G.H., and Nelson, R.L. (2007). Landmark research in legumes. Genome **50,** 525-537.

Sitnikova, N.A., Romanenko, S.A., O'Brien, P.C., Perelman, P.L., Fu, B., Rubtsova, N.V., Serdukova, N.A., Golenishchev, F.N., Trifonov, V.A., Ferguson-Smith, M.A., Yang, F., and Graphodatsky, A.S. (2007). Chromosomal evolution of *Arvicolinae* (*Cricetidae*, *Rodentia*). I. The genome homology of tundra vole, field vole, mouse and golden hamster revealed by comparative chromosome painting. Chromosome Res **15,** 447-456.

Soltis, D.E., Soltis, P.S., Morgan, D.R., Swensen, S.M., Mullin, B.C., Dowd, J.M., and Martin, P.G. (1995). Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proc Natl Acad Sci U S A **92,** 2647-2651.

Sykorova, E., Fajkus, J., Meznikova, M., Lim, K.Y., Neplechova, K., Blattner, F.R., Chase, M.W., and Leitch, A.R. (2006). Minisatellite telomeres occur in the family *Alliaceae* but are lost in *Allium*. Am J Bot **93,** 814-823.

Szinay, D., Bai, Y., Visser, R., and de Jong, H. (2010). FISH applications for genomics and plant breeding strategies in tomato and other solanaceous crops. Cytogenet Genome Res **129,** 199-210.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008a). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res **18,** 1944-1954.

Tang, X., de Boer, J.M., van Eck, H.J., Bachem, C., Visser, R.G., and de Jong, H. (2009). Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology. Chromosome Res **17,** 899-915.

Tang, X., Szinay, D., Lang, C., Ramanna, M.S., van der Vossen, E.A., Datema, E., Lankhorst, R.K., de Boer, J., Peters, S.A., Bachem, C., Stiekema, W., Visser, R.G., de Jong, H., and Bai, Y. (2008b). Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. Genetics **180,** 1319-1328.

Taponen, J., Mustonen, E.A., Kontio, L., Saastamoinen, I., Vanhatalo, A., Saloniemi, H., and Wahala, K. (2010). Red Clover Derived Isoflavones: Metabolism and Physiological Effects in Cattle and Sheep and their Concentrations in Milk Produced for Human Consumption. Recent Advances in Polyphenol Research, Vol 2**,** 238-254.

Taylor, N.L., and Smitht, R.R. (1979). Red clover breeding and genetics. Advances in agronomy **31,** 125.

Taylor, N.L., and Quesenberry, K.H. (1996). Red Clover Science. (Dordrecht, The Netherlands: Kluwer Academic).

Tejada, M., Gonzalez, J.L., Garcia-Martinez, A.M., and Parrado, J. (2008). Effects of different green manures on soil biological properties and maize yield. Bioresour Technol **99,** 1758-1767.

The Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

The Rice genome, T.m.-b.s.o. (2005). The map-based sequence of the rice genome. Nature **436,** 793-800.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res **16,** 934-946.

Thoquet, P., Gherardi, M., Journet, E.P., Kereszt, A., Ane, J.M., Prosperi, J.M., and Huguet, T. (2002). The molecular genetic linkage map of the model legume *Medicago truncatula*: an essential tool for comparative legume genomics and the isolation of agronomically important genes. BMC Plant Biol **2,** 1.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. (1999). Colinearity and its exceptions in orthologous adh regions of maize and sorghum. Proc Natl Acad Sci U S A **96,** 7409-7414.

Townsend, C.E., and Taylor, N.L. (1985). Incompatibility and plant breeding, N.L. Taylor, Clover Science and Technology (Agronomy No.25), ed (Madison, Wisconsin: ASA, CSSA, SSSA), pp. 616.

Trask, B.J. (2002). Human cytogenetics: 46 chromosomes, 46 years and counting. Nat Rev Genet **3,** 769-778.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L.,

Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313,** 1596-1604.

Van de Velde, W., Zehirov, G., Szatmari, A., Debreczeny, M., Ishihara, H., Kevei, Z., Farkas, A., Mikulass, K., Nagy, A., Tiricz, H., Satiat-Jeunemaitre, B., Alunni, B., Bourge, M., Kucho, K., Abe, M., Kereszt, A., Maroti, G., Uchiumi, T., Kondorosi, E., and Mergaert, P. (2010). Plant peptides govern terminal differentiation of bacteria in symbiosis. Science **327,** 1122-1126.

Van Ooijen, J.W. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations (Kyazma B. V., Wageningen, the Netherlands).

Vance, C.P., Graham, P.H., and Allan, D.L. (2000). Biological nitrogen fixation. Phosphorus: a critical future need. In Nitrogen Fixation: From Molecules to Crop Productivity, M.H. FO Pedrosa, MG Yates, WE Newton, ed (Dordrecht: Kluwer Academic Publishers), pp. 506-514.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T.,

Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagne, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., and Viola, R. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). Nat Genet **42,** 833-839.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in Arabidopsis. Science **290,** 2114-2117.

Wang, H., Moore, M.J., Soltis, P.S., Bell, C.D., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R., and Soltis, D.E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A **106,** 3853-3858.

Watanabe, K., Pacher, M., Dukowic, S., Schubert, V., Puchta, H., and Schubert, I. (2009). The STRUCTURAL MAINTENANCE OF CHROMOSOMES 5/6 complex promotes sister chromatid alignment and homologous recombination after DNA damage in Arabidopsis thaliana. Plant Cell **21,** 2688-2699.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol Biol **42,** 225-249.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiol **126,** 485-493.

Winters, A., Heywood, S., Farrar, K., Donnison, I., Thomas, A., and Webb, K.J. (2009). Identification of an extensive gene cluster among a family of PPOs in *Trifolium pratense L*. (red clover) using a large insert BAC library. Bmc Plant Biology **9,** -.

Wojciechowski, M.F. (2003). Reconstructing the phylogeny of legumes (Leguminosae): An early 21ST century perspective. In Advances in Legume Systenatics, part 10, Higher Level Systematics, B.B. Klitgaard and A. Bruneau, eds (Royal Botanic Gargens, Kew, UK).

Wright, D.A., Ke, N., Smalle, J., Hauge, B.M., Goodman, H.M., and Voytas, D.F. (1996). Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana. Genetics **142,** 569-578.

Yang, S., Zhang, X., Yue, J.X., Tian, D., and Chen, J.Q. (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. Mol Genet Genomics **280,** 187-198.

Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N., and Machida, Y. (1993). Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. Proc Natl Acad Sci U S A **90,** 6562-6566.

Young, N.D., Mudge, J., and Ellis, T.H. (2003). Legume genomes: more than peas in a pod. Curr Opin Plant Biol **6,** 199-204.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A., and Tabata, S. (2005). Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol **137,** 1174-1181.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., and Yang, H. (2002). A draft sequence of the rice genome (*Oryza sativa L*. ssp. indica). Science **296,** 79-92.

Zellinger, B., and Riha, K. (2007). Composition of plant telomeres. Biochim Biophys Acta **1769,** 399-409.

Zhang, Y., Sledge, M.K., and Bouton, J.H. (2007). Genome mapping of white clover (*Trifolium repens L.*) and comparative analysis within the Trifolieae using cross-species SSR markers. Theor Appl Genet **114,** 1367-1378.

Zhong, X.B., Hans de Jong, J., and Zabel, P. (1996). Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence *in situ* hybridization (FISH). Chromosome Res **4,** 24-28.

Zhou, T., Wang, Y., Chen, J.Q., Araki, H., Jing, Z., Jiang, K., Shen, J., and Tian, D. (2004). Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics **271,** 402-415.

Zwick, M.S., Hanson, R.E., Islam-Faridi, M.N., Stelly, D.M., Wing, R.A., Price, H.J., and McKnight, T.D. (1997). A rapid procedure for the isolation of Cot-1 DNA from plants. Genome **40,** 138-142.

Education Statement of the Graduate School
Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Chunting Lang**
Date: **31 January 2012**
Group: **Laboratory of Molecular Biology, Wageningen University & Research centre**

1) Start-up phase	*date*
► **First presentation of your project**	
Morphology of *Trifolium pratense L.* pachytene chromosomes and integration of the cytogenetic karyotype and the genetic linkage map	Mar 07, 2008
► **Writing or rewriting a project proposal**	
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Radiation course	Nov 04-06, 2008
Subtotal Start-up Phase	*3,0 credits**

2) Scientific Exposure	*date*
► **EPS PhD Student Days**	
EPS PhD student day, Wageningen University	Sep 13, 2007
Joint international PHD retreat (SDV, IMPRS and EPS), Wageningen, The Netherlands	Oct 02-03, 2008
EPS PhD student day, Leiden University	Feb 26, 2009
2nd Joint international PHD retreat, Max-Planck-Institute for Plant Breeding Research, Cologne, Germany	Apr 15-17, 2010
► **EPS theme symposia**	
EPS Theme 4 Symposium, Leiden University	Dec 07, 2007
EPS Theme 4 Symposium, Wageningen University	Dec 12, 2008
EPS Theme 4 Symposium, Radboud University Nijmegen	Dec 11, 2009
EPS Theme 4 Symposium, Wageningen University	Dec 10, 2010
► **NWO Lunteren days and other National Platforms**	
NWO-ALW Lunteren days 2008 EPS	Apr 07-08, 2008
NWO-ALW Lunteren days 2009 EPS	Apr 06-07, 2009
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 19-20, 2010
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 04-05, 2011
► **Seminars (series), workshops and symposia**	
Flying seminar Scott Poethig	Sep 21, 2007
Seminar Greg Amoutzias	Oct 22, 2007
Flying seminar Hiroo Fukuda	Nov 26, 2007
Flying seminar Richard Vierstra	Apr 14, 2008
Flying seminar Simon Gilroy	May 18, 2008
Flying seminar ZhenBiao Yang	Jun 23, 2008
Seminar Valerie Williamson	Oct 23, 2009
Seminar Series Plant Sciences	Oct 13, 2009
Seminar Series Plant Sciences	Jan 12, 2010
Seminar Series Plant Sciences	Dec 08, 2009
Seminar Dr. Matteo Brilli	Apr 29, 2010
Seminar Prof. David Baulcombe	Sep 27, 2010
Seminar Prof. Peter Cook	Oct 27, 2010
Seminar Dr. Kirsten Bomblies	Nov 18, 2010
► **Seminar plus**	
seminar plus Simon Gilroy	May 18, 2008
► **International symposia and congresses**	
International chromosome conference 2007 (Amsterdam)	Aug 25-27, 2007
Project meeting 2008 (Amsterdam)	Mar 07, 2008
Project meeting 2009 (London)	Nov 09, 2009
Vth International Congress on Legume Genetics and Genomics (ICLGG)	Jul 02-08, 2010
► **Presentations**	
Oral: Morphology of *Trifolium pratense L*. pachytene chromosomes and integration of the cytogenetic karyotype and the genetic linkage map	Mar 07, 2008
Poster: Morphology of *Trifolium pratense L*. pachytene chromosomes and integration of the cytogenetic karyotype and the genetic linkage map	Oct 02-03, 2008
Oral: project meeting presentation	Nov 09, 2009
Oral: Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*	Dec 12, 2008
Poster: Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*	Apr 06-07, 2009
Oral: Alignment of *Medicago truncatula* and Legume *Trifolium pratense* genome	Dec 10, 2009
Oral: Translational genomics from model species *Medicago truncatula* to crop legume *Trifolium pratense*	Apr 15, 2010
Poster: The evolution history of the chromosome rearrangement between *Medicago truncatula* and *Trifolium pratense*	Jul 02-08, 2010
Oral: The evolutionary history of the chromosomer rearrangement between *Medicago truncatula* and *Trifolium pratense*	Dec 19, 2010
► **IAB interview**	Dec 02, 2009
► **Excursions**	
Subtotal Scientific Exposure	*20,8 credits**

3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	Plant Genetics: Natural Variation	Aug 25-29, 2008
	Molecular Phylogenies: Reconstruction and Interpretation	Oct 13-17, 2008
►	**Journal club**	
	Literature discussion MolBi	2007-2011
►	**Individual research training**	
	Subtotal In-Depth Studies	*6,0 credits**

4) Personal development		*date*
►	**Skill training courses**	
	Advanced Course Guide to Scientific Artwork	Dec 15-16, 2008
	Techniques for Writing and Presenting a Scientific Paper	Aug 31-Sep 03, 2010
	Career Perspectives	Oct 19,26-Nov 02, 09, 16, 2010
	Dutch Language Level 2	Jul 19-29, 2010
►	**Organisation of PhD students day, course or conference**	
►	**Membership of Board, Committee or PhD council**	
	Subtotal Personal Development	*5,2 credits**

TOTAL NUMBER OF CREDIT POINTS*	**35.0**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*