UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

The *Medicago truncatula* genome and analysis of nodule-specific genes

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

Jing Yi

Norman, Oklahoma

2009

THE *MEDICAGO TRUNCATULA* GENOME AND ANALYSIS OF NODULE-SPECIFIC GENES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

_____
Dr. Bruce A. Roe, Chair

_____
Dr. Paul F. Cook

_____
Dr. Ann H. West

_____
Dr. George Richter-Addo

_____
Dr. Jia Li

Acknowlegements

I would like to sincerely thank my major advisor, Dr. Bruce A. Roe, for the inspiration by his brilliant thoughts and his hard work, and for his constant help and encouragement. I also would like to express my deep appreciation to my other committee members, Drs Paul F Cook, Ann H West, George B Richter-Addo and Jia Li for their continued support, advice, and encouragement during my general exam and annual evaluation.

I thank all the members in Dr. Roe's lab, for their advice, help and support with special thanks to Hongshing Lai, Dr. Fares Najar, and Dr. Axin Hua for their help on my analysis; Drs Ruihua Shi, Jianfeng Li, Shweta Deschpande, Majesta O'Bleness, Iryna Sanders who as my colleagues I could discuss everything from personal life to scientific problems; as well as Shaoping Lin, Fu Ying, Yanbo Xing, Chunmei Qu, Ping Wang, Liping Zhou, Ziyun Yao, Baifang Qing, Keqing Wang, Limei Yang, Junjie Wu, Sulan Qi, Jiaxi Quan, Graham Wiley, Simone Macmil, Doug White, Steve Kenton, Jim White, Mounir Elharam, MaryCatherine Pottorff, Kay Lnn, Dixie Wishnuck and Drs Shelly Oommen, Christopher Lau, Jiaxin Qu, Hung-chun Yu, Leo Sukharnikov.

I would like to thank my parents, Dexin Yi and Biqin Wang, my sister, Ping Yi, for their support, love, and patience since they have been waiting for so long to see me.

I dedicate this dissertation to my husband Dr. Xiaoping Gou for his love, support, encouragement and patience for so many years and to my lovely daughter Zoe who is the motivation and inspiration for my life.

Table of contents

List of tables

List of figures

Abstract

The nitrogen-fixing plant *Medicago truncatula* is an important model system for identifying legume genes and determining their functions. With over 255 megabases of the genome, representing about 85% of the euchromatic regions, having been sequenced, my analysis reveals 50,540 predicted protein-encoding genes, 632 tRNA genes, 45 miRNA precursor candidates, and repetitive elements covering 11% of the sequence. ~ 50% predicted genes are supported by ESTs or TCs. About 40% of the predicted genes are intronless and there is evidence for 55% of them being expressed. A comparison of the *Medicago truncatula*, *Oryza sativa*, *Arabidopsis thaliana, Lotus japonicus, Glycine max,* and *Populus trichocarpa* genomes shows that the Medicago genome uniquely contains a high number of very short genes encoded by predicted genes with fewer than 99 nucleotides. The Gene Ontology (GO) annotation of the predicted genes showed that the nucleic acid binding domains are the most abundant domains in *M.truncatula*. The comparison between GO the annotation of *M.truncatula*, *O. sativa*, *A.thaliana, L. japonicus*, *G. max*, and *P. trichocarpa* reveals that all the six genomes have similar percentage of each of the major functional domains. The comparison of the top 40 Interpro domains in *M. truncatula* with the corresponding domains in the other five plants also indicates that most of the overrepresenting domains are overrepresenting in all the six genomes although some species-specific domains, such those for late nodulation, only are present in *M. truncatula*.

The *in silico* analysis of the Medicago Gene Index 9.0 revealed that 191 genes only are expressed in root nodules, with100 of them similar to known GenBank sequences. Of the several gene familys, my analysis of 50 nodule-specific cysteine-rich peptides (NCR) indicates that they have a conserved signal peptide, a conserved cysteine motif, and a highly divergent remaining sequence. Many of the NCR genes are clustered while others are dispersed throughout the Medicago genome, suggesting that they have ungone a recent tandem or segmental gene duplication. A Ka/Ks analysis of NCR genes indicates that although some NCR genes underwent positive selection, others underwent purifying selection. That the NCR intron sequence is highly conserved suggests it may act as an enhancer for nodule-specific NCR gene expression in combination with the conserved upstream cis-acting motifs. The phylogenetic tree of both defensin and NCR genes reveals that after gene duplication, some of the defensin genes still remained defensins as is the case with the five medicago defensins, while the other duplicated defensin genes mutated such that they now seem to function in symbiosis as NCR genes.

I also analyzed the three members of the glycine-rich peptide (GRP) gene family that are encoded on chromosome 2. These studies reveal that the GRP1 gene that is 5.2 Mb from the GRP2-GRP3 cluster arose from tandem gene duplication followed by either a deletion or an insertion from a common ancestor, an idea that is supported by sequence conservation in the signal peptide, the glycine-motif, and the 200bp upstream DNA sequence. In addition, my Ka/Ks analysis indicates that positive selection played an

important role during GRP gene evolution. Furthermore, the Leghemoglobin (Lb) genes that originated from nonsymbiotic hemoglobins (Hb) seem to have undergone a purifying selection that has preserved their ability to function during oxygen transport. Finally, the nodule-specific genes all seem to contain one or both of two nodule-specific motifs (CTCCT and AAAGAT) in their promoter regions suggesting that nodule-specific gene expression likely is co-regulated and about 50% of the predicted nodule-specific genes are clustered and have corresponding EST, indicating that they are expressed in root nodules.

1. Introduction

1.1 The importance of legumes

The legume family (Fabaceaea or Leguminosae) contains over 700 genera and 20,000 species, making it the third largest family of flowering plant, after orchids and sunflowers (Doyle et al. 2003), and an important food source for humans and domestic livestock (Gepts et al. 2005). Legumes differ from most other plants in that they have the ability to fix atmospheric nitrogen in symbiosis with rhizobial bacteria and thus do not require nitrogen fertilizer for their growth and development. Grain legumes provide one third of the world's dietary protein nitrogen and one third of its edible vegetable oil (Graham et al. 2003). Additionally, dietary legumes can reduce cholesterol levels in humans (Andersen et al. 1984) and help prevent cancer because they synthesize secondary compounds, including isoflavoids and triterpene sapoinins (Grusak et al. 2002, Madar et al. 2002), which also protect plants from pathogens and pests (Dixon et al. 2002). Lastly, legumes also are used world-wide in soil replenishment via crop rotation to maintain agricultural sustainability.

The legume family is a member of the dicot Eurosid clade that comprises three subfamilies: Caesalpinieae, Mimosoideae, and Papilionoideae (Gepts et al. 2005). These comprise economically important (Doyle et al. 2003), nutritious and versatile plants such as the cool-season legume *Medicago truncatula* (barrel medic), *Lotus japonicus*, *Pisum sativum* (pea), *Medicago sativa* (alfafa), and *Vicia arietinum*

(chickpea) and *Glycine max* (Soybean), *Phaseolus spp.* (common bean), and *Vigna radiate* (mungbean), the warm-season legumes.

1.2 *Medicago truncatula* as a model plant for legumes

Arabidopsis thaliana and *Oryza sativa* (rice) are model plants for dicotyledon and monocotyledon, respectively as their features are shared among a wide range of related taxa (Eckardt et al. 2001) . Although many of the agronomically and economically important genes in plants can be identified via homology with their counterparts in *A. thaliana* or *O. sativa*, since neither of these two model plants fixes nitrogen, they cannot be used as a model system to study this symbiotic process. In contrast, legumes do have the unique ability to fix nitrogen that the plant needs for the synthesis of its protein through symbiosis with rhizobia and in return, the host plant provides the rhizobia with all the nutrients they require, including a rich supply of reduced carbon (Fang et al. 1998). Unfortunately, it is difficult to use most cultivated legumes (including pea, alfalfa, and soybean) in genomic studies because of their large genomes (the 4 billion base pair (bp) pea genome is slightly larger than the human genome), genome complexity (alfalfa is a tetraploid, soybean is a polyploid), large seeds, and long life cycles. Therefore, among legume species, *Medicago truncatula* recently has emerged as a model plant for legume genetics and genomics. It has a small diploid genome, good genetic and physical map resources, fast generation time, high transformation

efficiency, excellent collections of phenotypic mutants (especially in nodule formation and symbiotic nitrogen fixation), diverse, naturally occurring ecotypes, and extensive synteny with the genomes of larger legumes (Cook et al. 1999, Cannon et al. 2005). By comparing the existing previously small amount of genomic data from legumes, it has been observed that the *M. truncatula* genome is highly conserved with that of alfalfa and pea (Gualtieri et al. 2002, Endre et al. 2002), and moderately conserved with that of soybean, at both micro-syntenic and macro-syntenic levels (Yan et al. 2003). Consequently, knowledge of the genome organization and structure of *M. truncatula* will be useful to explore conserved genes, gene order and gene orientation of the other crop legume genomes through comparative genomics as, for example, microsyteny was utilized to clone *M. truncatula* DMI2 and its ortholog, NIN1 from *M. sativa* (Endre et al. 2002, Stracke et al. 2002), which are expressed in the early stages of rhizobial symbiosis.

1.2.1 Nodule formation

Much already is known about legume-rhizobia symbiosis. The procedure of nodule formation is schematically shown in Figure 1.1. The rihizobia receive the signal of flavonoid compounds released by legumes and then release lipochito-oligosacharrides (Lerouge et al. 1990), termed *nodulation factors* (Nod) (Peters et al. 1986). The curling of root hairs to trap the rhizobia, the first step of nodulation, is induced by nodulation

factors and facilitates the rhizobia binding to the root hairs, as a prelude to them entering the plant root through a special structure called the infection thread. The infection thread grows through the root towards the nodule primordia within the root cortex (Brewin et al. 1998). When the rhizobia reach the nodule primordia, they are released from the infection thread and surrounded by a plant-derived membrane known as symbiosome membrane (Brewin et al. 1998). The bacteria in these compartments continue to divide and differentiate into nitrogen fixing bacteroids (Vasse et al. 1990). Cell division continues in the root tissue, and eventually the nodule containing the infected plant cells will form. Photosynthetic products that provide the energy source for the bacteria as well as C-backbone during nitrogen fixation and assimilation are transported to nodules. In the bacteroids, N_2 is reduced to ammonium by the nitrogenase complex in the following reaction (Scholte et al. 2002):

$$N_2 + 8H^+ + 8e^- + 16ATP + 16H_2O \rightarrow 2NH_3 + H_2 + 16ADP + 16Pi$$

Ammonia then is transported to the plant cell cytoplasm, where it is assimilated by a glutamine synthetase/glutamate synthase cycle to form glutamine (Scholte et al. 2002). Glutamine, in turn, is converted to the N-containing amides or ureides that are transported via the xylem to the other organs of the plant.

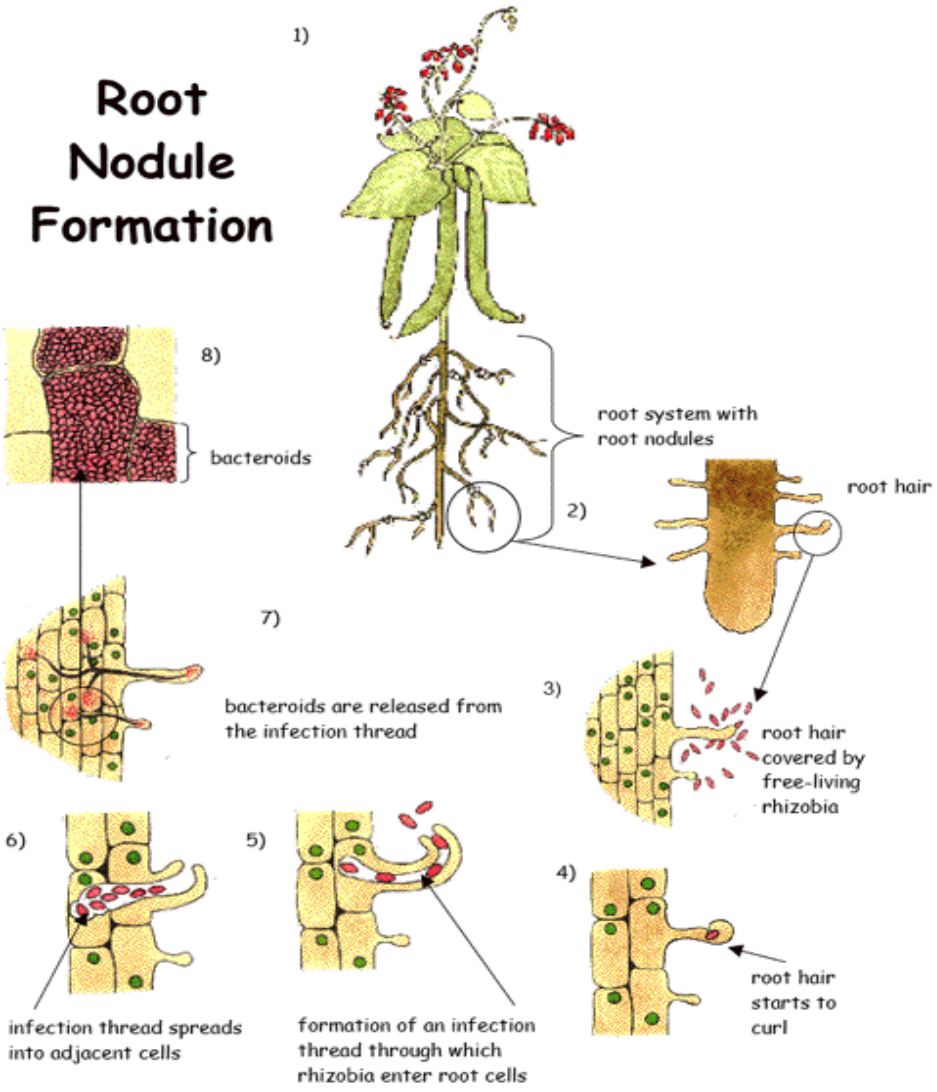


Figure 1.1 Schematic representation of nodule formation. Taken from Rhizobium, Root Nodules & Nitrogen Fixation, A post-16 resource from the Society for General Microbology, United Kingdom (2002)

1.2.2. Nodule-specific genes

All of the bacterial genes involved in symbiotic nitrogen fixation have been identified

(Gresshoff et al. 2003). However, our progress in understanding plant genes required for nodule development is slow because of the eukaryotic genome complexity and the absence of the plant genome sequence. For example, although it is well known that nodule specific genes were recruited from genes involved in plant growth and development, it is not known if these nodulation genes have common cis-regulatory elements or if they are co-regulated. Also, little is known about the genome organization of these genes, how they evolved and their ancestral origin. However, using differential screening, subtractive hybridization, and differential display, it has been possible to find nodule-specific related genes because their expression is specific for or greatly enhanced during symbiosis (Crespi et al. 2000). Nodulin genes can be divided into two groups: those induced at an early stage (such as ENOD2, ENOD12, ENOD16, ENOD20, and ENOD4 genes) and those expressed late and associated with nitrogen fixation (genes encode leghemoglobin, glutamine synthetase, and uricase) (Charon et al. 1999). Nodulin genes especially, the early nodulin genes were thought to be expressed exclusively in nodules. However, it is now clear that many of these genes also are expressed in nonsymbiotic tissues at low levels (Charon et al. 1999). Hirsch *et al* suggested many of the genes required in nodule development and nitrogen fixation were recruited from their original task in plant growth and development to function in the nodule (Hirsch et al. 2001). It now is apparent that while many nodulin genes have more than one function, some only are expressed in nodules. The nodulation gene, *nin-1*, isolated from *L. japonicus* using transposon mutagenesis (Hirsch et al. 1999),

encodes a transcription factor that functions at the infection-thread stage. Recently, more than 300 putative nodule-specific genes have been identified *in silico* by the analysis of the appearance and frequency of ESTs from different cDNAs (Fedorova et al. 2002). However, if the nodule-specific genes have common cis-regulatory elements and how they evolved remain unclear. It is clear from gene prediction results in completely sequenced organisms such as human and *A. thaliana* that only about 60% of expressed genes can be discovered through cDNA sequencing (Eckardt et al. 2001). This is because many genes are expressed at very low levels and /or in tissues, developmental stages, or in response to external stimuli that have not yet been examined and thus are underrepresented in EST libraries. ESTs only provide a partial DNA sequence, as they lack information about promoters, introns, or other regulatory elements. cDNA sequencing also fails to reveal genome context, such as nearby repetitive elements or other features. Moreover, EST projects cannot provide any information about gene order that is important for comparative genomics. Only through genomic sequencing can we obtain data necessary for comparative mapping, determine macro- and micro-synteny, and study the gene evolutionary history among different organisms. Since genes with the similar functions often are clustered in plants, for example, many of the *M. truncatula* disease-resistant genes are clustered on chromosome 4 and 6 (Zhu et al. 2002), while many *M. truncatula* nodule-specific genes are clustered on chromosome 8 (Thoquet et al. 2002, Huguet et al. 2000), obtaining the complete sequence of the euchromatic region of this model legume genome will provide

information directly related to the makeup of these gene clusters and duplicated genes in genome.

Just as the complete genome sequencing and gene tagging revealed the bacterial genes needed for both nodulation and nitrogen fixation (Gresshoff et al. 2003), the genomic sequence of *Medicago truncatula,* in combination with other molecular genetic technologies, also may help us find additional genes involved in the nodulation process.

1.3 The *Arabidopsis thaliana, Oryza sativa and Populus trichocarpa, Glycine max*, and *Lotus japonicus* genomes

Arabidopsis thaliana (mouse-ear cress or mustard weed)*, Oryza sativa* (rice), and *Populus trichocarpa* (black cottonwood) were the first sequenced dicot, monocot, and tree respectively. *Lotus japonicus* and *Glycine max* are other legumes being sequenced besides *Medicago truncatula*. Their analysis provides insight to the organization and structure of the plant genome and genes and also facilitates research into the genetic systems of the plants.

1.3.1 The *Arabidopsis thaliana* genome

Arabidopsis thaliana, the first plant to have its genome completely sequence, is considered a model for dicotyledonous plants because it is a small plant with a short generation time and a small genome size (125Mb) (The Arabidopsis Genome Initiative, 2000). The Arabidopsis Genome Initiative (AGI) began sequencing the *A. thaliana*

genome in 1996, and in 2000 the complete sequence was published (AGI, 2000). With a total length of sequenced genome from the telomeric region (rich in rDNA repeats) to centromeric region (180bp repeats) of about 115Mb, and unsequenced regions estimated at approximately 10 Mb, the total size of the *A. thaliana* genome is estimated as 125Mb. About 25,000 genes that code for proteins were predicted and grouped into approximately 11,000 families. Sixty-nine percent of the genes could be categorized into functional groups based on their sequence similarity to proteins of known function in other organisms. About 30% of the gene products were either plant-specific proteins or proteins similar to proteins from other organisms with unknown function. At that time, 35% of the classified genes were unique and 65% belonged to gene families, many of which were conserved in all eukaryotes. Approximately 150 gene families were annotated as plant-specific, including transcription factors, structure proteins, enzymes, and proteins of unknown function. Lateral gene transfer from a cyanobacterial-like plastid ancestor also was detected in the *A. thaliana* genome. One thousand five hundred twenty eight tandem arrays containing 17% of the genes were observed, along with 24 large duplicated segments of 100 kb or larger, that made up 58% of the genome. Many of the segments seem to have undergone further shuffling. It was suggested that the *A. thaliana* genome evolved from a tetraploid ancestor ~112Myr ago, because majority of the genome occurs in duplicated segments. Approximately 10% of the genome consists of transposons, with both class I (replicate trough an RNA intermediate) and class II (move directly through a DNA form) elements. The

transposon-rich regions are poor in genes and have lower rates of recombination and EST matches, indicating a correlation of high transposon density with low gene expression and low recombination. Heterochromatic regions surrounding the centromere are rich in transposons and other repetitive sequences, whereas the euchromatic arms are not. *A. thaliana* will be very important in the study of epigenetic inheritance and gene regulation, as it is the first sequenced methylated genome. As the first fully sequenced flowering plant, the analysis of the Arabidopsis genome has provided a better understanding of plant development and environmental responses, and the organization and dynamics of plant genomes.

A. thaliana and *Medicago truncatula* belong to the dicot subclass Rosidae (Zhu, 2003). A comparative genome analysis between Arabidopsis and *M. truncatula* revealed that a degenerate network of microsynteny was observed between these two genomes. However, the macrosynteny was not obvious using genetic map-based and bacterial artificial chromosome (BAC) sequence-based methods (Zhu et al. 2003). In addition, genetically linked loci in *M. truncatula* often have several syntenic regions in Arabidopsis genome that is consistent with the conclusion that the Arabidopsis genome consists of a large number of segmental duplications. The degenerate microsynteny was suggested partially due to the lack of a small number of *M. truncatula* homologs in *A. thaliana*.

1.3.2 The *Oryza sativa* genome

Oryza sativa, or rice, one of the most important food plants in the world, is an ideal model plant for the grasses. It has the smallest genome size of the major cereals, dense genetic maps, ease of transformation, and synteny among the cereals (Sasaki et al. 2002). The International Rice Genome Sequencing Project (IRGSP), established in 1998, sequenced 95% of the 389 Mbp genome of a single inbred cultivar, Oryza sativa ssp.japonica cv. Nipponbare (IRGSP, 2005). This included all of the euchromatic regions and two complete centromeres. In total, 37,544 non-transposable-element-related protein-coding genes were predicted, 71% of which were homologous to genes in *A. thaliana*. Twenty nine percent of the total predicted genes were from clustered families. About 60% of the 37,544 predicted genes had EST, or full-length cDNA support. The gene density is 9.9 kb/gene. 2,859 genes which can be transcribed were only found in rice and the other cereals. 0.38%-0.43% rice nuclear genome consists of organellar DNA fragments, indicating widespread and repeated DNA transfer from the organelles to the nuclear chromosomes. About 35% of the rice genome consists of transposons from all known transposon superfamilies. As with other eukaryotic centromeres, those of rice are rich in repetitive sequences that include satellite DNA at the center and retrotransposons and transoposons in the flanking regions. The centromeres of all rice chromosomes contain the highly repetitive 155-165 bp CentO satellite DNA, flanked by centromere-specific retrotransposons, as was demonstrated by

completely sequencing chromosomes 4 and 8 to reveal their 59 kb and 69 kb of clustered CentO repeats (Wu et al. 2004, Zhang et al. 2004, Guyot et al. 2004) that are distributed as head-to-tail tandem array.

1.3.3 The *Populus trichocarpa* genome

The genome of the poplar tree, *Populus trichocarpa*, is the first sequenced tree genome. It was chosen as the model forest species because of its small genome, fast growth, ease of transformation and the availability of numerous genetic tools (Tuskan et al. 2006). The draft genome sequence was published in Science in 2006 (Tuskan et al. 2006). A whole genome shotgun sequencing strategy was used to obtain 7.5 fold genomic coverage, and the genome sequence was assembled into 2447 major scaffolds. The genome size, estimated to be 485±10 Mb, is about 30% heterochromatin. Among the 45,555 predicted protein-coding genes, 89% were homologous to the proteins in nonredundant (NR) protein database from the National Center for Biotechnology Information (NCBI). Eight hundred seventeen putative tRNAs, 427 putative small nucleolar RNAs (snoRNAs) and 169 microRNA (miRNA) from 21 families also were identified in the Poplar genome. One thousand five hundred eighteen tandem duplicated arrays of two or more genes also were found in the genome. A whole genome duplication event was found by analyzing the Poplar genome. About 8000 pairs of paralogs survived from this event. Comparison between *P. trichocarpa* and *A. thaliana*

genomes showed that poplar has more protein-coding genes than *A. thaliana*, but that the relative frequency of protein domains in the two genomes is similar. Gene families involved in lignocellulosic wall biosynthesis, disease resistance, metabolite transport, and meristem development are overrepresented in *P. trichocarpa*.

1.3.4 The *Lotus japonicus* genome

Lotus japonicus was sequenced as the second model legume, in addition to *Medicago truncatula* because of its short life cycle, self-fertilization, a small genome size (472Mb), and a comparatively simple genome (2n=12) (Sato et al. 2008). About 315.1 Mb genome sequences, 67% of the genome, have been determined, and they cover 91.3% of the gene space (Sato et al. 2008). The sequenced lotus genome is composed of 30,799 protein-encoding genes, 638 tRNA genes, two complete units of 18S-5.8S-26S ribosomal RNA genes, 207 snoRNA gene, 53 miRNA, and 38% of transposable elements. About 52% of the protein-encoding genes are supported by ESTs with sequence identity of over 95% for a 50 base-long stretch. The structure of protein-encoding genes in *L. japonicus* is similar to that in *A. thaliana* except that both the average gene length (2917 vs. 1918 bp) and intron length (395 vs. 157bp) in lotus are longer than that in *A. thaliana*. The average gene length in *L. japonicus* and *A. thaliana* are estimated at 10.2 kb/gene and 4.5 kb/gene, respectively.

The previous studies between 149 Mbp of *M. truncatula* and 121 Mbp of *L. japonicus* revealed that they share at least 10 large-scale synteny blocks that often extend to the

length of the whole genome arms (Cannon et al. 2006). These large-scale synteny blocks constitute about 67% of the *M. truncatula* genome and about 64% of *L. japonicus*. The genome sequences in internal duplications either in *L. japonicus* (6.8%) or in *M. truncatula* (9.7%) are much less than in synteny blocks between the two genomes. Further more, the synteny between *M. truncatula* and *L. japonicus* tends to be extensive with numerous paired syntenic blocks observed. All the above duplication analyses indicate that there is no recent large-scale genome duplication either in *M. truncatula* or *L. japonicus* and that a whole genome duplication (WGD) preceded speciation of *L. japonicus* and *M. truncatula* (Cannon et al. 2006). The synonymous substitution analysis and the phylogenetic analysis also proved the above conclusion. Phylogenetic analyses suggest that the WGD event occurred within the Rosid I clade, after the split of the legume family and the Salicaceae (poplar).

1.3.5 The *Glycine max* genome sequence

Glycine max (soybean) belongs to the lineage Phaseoloids that split 50 million years ago (MYA) from the lineage Hologalegina containing *L. japonicus* and *M. truncatula* (Lavin et al. 2005). *G. max* was sequenced as the third model legume because of its economic importance, detailed and saturated genetic map, existing physical map, and medium-sized genome size (estimated as 1100 Mbp) (Jackson et al. 2006). The soybean genome, consisting of 20 chromosome pairs, is an ancient polyploidy that was thought to have undergone 2 to 3 times of genome duplication during the last 45 MY

(Shoemaker et al. 1996, Blanc and Wolfe 2004, Schlueter et al. 2004). The most recent duplication dates back to only 1-3 MYA and accounts for the low sequence drift and high sequence similarity between many duplicated blocks (Jackson et al. 2006). Thus, the present day *G. max* genome consists of highly conserved homeologous regions along with regions with high gene loss and rearrangement resulted from genomic reshuffling after multiple rounds of duplication (Jackson et al. 2006).

Even though the *G. max* genome is polyploidy, its genome structure still is organized like the *M. truncatula* genome, that is, euchromatin is located on the chromosome arms while heterochromatin is located on the centromeric and pericentromeric regions (Lin et al. 2005, Walling et al. 2006). Nearly 40-60% of the soybean genome is composed of repetitive elements based on DNA: DNA reassociation studies (Goldberg 1978, Gurley et al. 1979) and the gene space is estimated as about 24% of the genome (Mudge et al. 2004). The average gene density in soybean genome was estimated as 5.8-6.7 kb per gene (Mudge et al. 2005). Although macrosyntenic relationship between soybean and other legumes are difficult to detect because of chromosome rearrangement or gene loss after genome duplication, several studies showed microsynteny is frequently maintained over small chromosome regions (Choi et al. 2004, Yan et al. 2003, Mudge et al. 2005). Choi et al (Choi et al. 2004) identified 11 syntenic blocks between soybean and *M. truncatula* by mapping 60 homologous markers. In Yan et al. (Yan et al. 2003), 27 out of 50 soybean contigs were shown to have microsynteny with *M. truncatula* to some extent. The analysis in Mudge et al. revealed two large soybean regions showed high

synteny with two *M. truncatula* chromosomes (Mudge et al. 2005). The syntenic regions span 3 Mb and contain 500 predicted genes with 75% of these soybean genes collinear with *M. truncatula*.

1.4 DNA, gene, and genome

What is a gene? What is the relationship between a gene and DNA? What is a genome and what is chromosome? Why do we need to sequence a genome? To address these questions, we need to have the following basic knowledge.

A genome is the entire genetic material of an organism. Deoxyribonucleic acid (DNA) is the genetic material of most living things (Avery et al. 1944) except for some viruses that use ribonucleic acid (RNA) as genetic material. A genome exists in the form of chromosomes that consist of DNA. Some genomes are composed of a single chromosome while others are divided into multiple chromosomes. A gene is a segment of a chromosome that encodes RNA. In some cases this initial RNA transcript is translated into protein, while in other cases the RNA is a stable molecule. A gene is the fundamental unit of heredity. All of the genes encoded in a genome gives the genotype of the organism and their expression results in the specific characteristics or phenotype, of the organism.

1.4.1 DNA structure and central dogma

DNA, a polymer that consists of nucleotide units, consists of three components: a

pentose sugar, a phosphate, and a nitrogenous base. There are four nitrogenous bases that carry genetic information: adenine, cytosine, guanine, and thymine, abbreviated as A, C, G, and T, respectively. RNA also has A, C, and G, and instead of T, a uracil (U). A and G are purines and C, T, and U are pyrimidines. Another difference between DNA and RNA is that the pentose in DNA is 2-deoxyribose but the pentose in RNA is ribose. Ribose has an OH group at the 2' position while 2-deoxyribose does not. The nitrogenous base is attached to position 1' of the pentose ring through a glycosidic bond from the N1 of pyrimidines or the N9 of purines. The structure of DNA was first correctly proposed in 1953 by Watson and Crick (Watson and Crick, 1953), in a 2-page *Nature* paper, where they suggested that DNA has two helical polynucleotide chains that wind around the same axis. Each polynucleotide chain is built by the 3'-5' phosphodiester linkage that links the 3' position of one pentose ring to the 5' position of the next pentose ring by a phosphate group. Both chains are right-handed helices and their direction is opposite, i.e. one is 5' to 3', another is 3'to 5'. For each chain, the bases lie on the inside of the helix and the sugar-phosphate backbone is on the outside. The DNA helix makes a complete turn every 3.4 nm with 10 nucleotides per turn. According to Watson and Crick, the two DNA chains are joined by hydrogen bonding of the purine and pyrimidine bases with G hydrogen bound with C, and A hydrogen bound with T.

Chargaff demonstrated that although DNAs in different organisms have the same sugar-phosphate backbone, they differ in the amount of the 4 bases and likely the base order,

suggesting to him that the sequence of the bases carries genetic information (Chargaff 1950). The “central dogma”, the process of duplicating and expressing the genome, was first described by Francis Crick in 1958 (Crick, 1958) where he pointed out “Once information has passed into a protein it cannot get out again”. In 1970, Crick further clarified this idea by proposing there are three classes of information transfer: general transfer, special transfer, and unknown transfer (Crick, 1970). A general transfer occurs in most cells that include transfer from DNA to DNA, from DNA to RNA and from RNA to protein. A special transfer can only happen in special circumstances, including transfer from RNA to DNA, from RNA to RNA, and from DNA to protein. The transfer of biological information from protein to protein, protein to DNA, or protein to RNA have yet to be described and likely never happens in cells. The information transfer from DNA to DNA is called DNA replication. DNA is duplicated by a semiconservative replication (Meselson et al. 1958), as each of the polynucleotide parental strands of DNA acts as a template for the synthesis of a new daughter strand. The sequence of the daughter strand is determined by the parental strand as dictated by the base pairing rules (A pairs with T, G pairs with C). After replication, the parental DNA duplex forms two daughter duplexes that are identical to each other and contains one parental strand and a newly synthesized strand. The transfer of genetic information from DNA to RNA, called transcription, occurs in all cells. In retroviruses that have an RNA genome, the RNA can be reversely transcribed into DNA by reverse transcriptase enzymes or RNA-dependent DNA polymerase. The genome of retroviruses consists of

single-stranded RNA that is converted to a single-stranded DNA and subsequently to a double stranded DNA that is then inserted to the genome of a cell during infection cycle. Reverse transcription allows RNA to act as genetic information. Interestingly, some plant viruses have a double-stranded RNA genome that is not reverse transcribed to DNA but rather is replicated by an RNA dependent RNA polymerase (van Kammen 1985).

The transfer of information from RNA to protein is called translation. Translation allows for the expression of genetic information in the form of proteins. Translation is unidirectional, meaning that the information flow from RNA to protein is irreversible. The central dogma has been developed in great details, as more and more discoveries are made in the field of gene expression (shown in Fig1.2). Gene expression is regulated at the transcriptional, posttranscriptional, translational, and posttranslational levels. In eukaryotes, precursor messenger RNAs (pre-mRNAs), synthesized using DNA template undergo a process called splicing. During splicing, introns of pre-mRNA are excised and the exons are joined. In alternative splicing, exons from the pre-mRNA are reconnected in alternative ways to produce mRNA variants. These variants then are translated into isoform proteins (Lalli et al. 2003).

The process of chemical modification of a protein after translation is called posttranslational modification. In this process, biochemical functional groups such as the acetyl, methyl, phosphate, and glycosyl groups are attached to a protein or protein structure, and oxidation to form disulfide bridges.

Epigenetics refers to the study of a phenomenon wherein gene function is changed genetically but the sequence of the DNA is not changed (Adrian, 2007). The widely studied epigenetic processes include DNA methylation and histone modification.

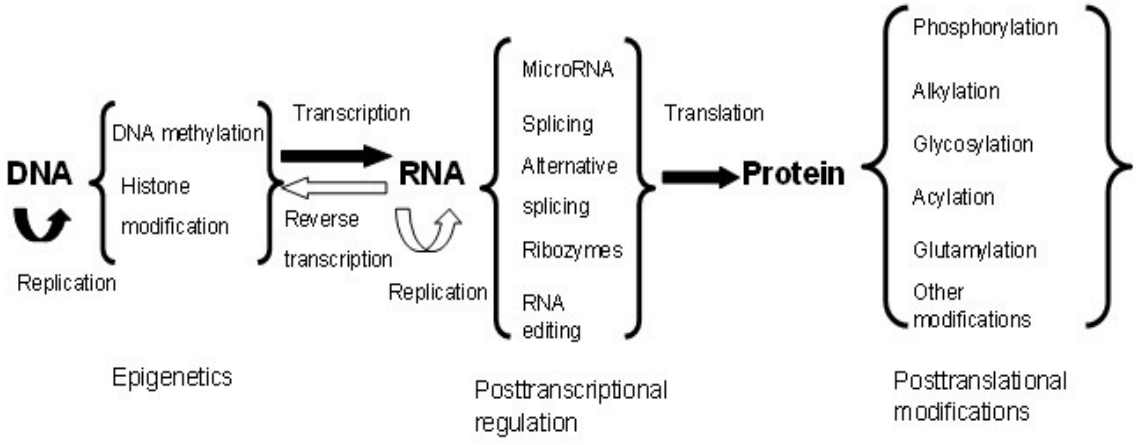


Figure1.2 The schematic diagram of updated Central dogma

1.4.2 Gene

In eukaryotes, the protein-coding gene consists of exons, introns, as well as 5' non-translated and 3' non-translated regions. Exons are DNA regions in a gene that are present in the mature mRNA while introns are DNA sequences between exons that are transcribed into pre-mRNA but will be spliced out when the pre-mRNA is processed into a mature mRNA. Generally speaking, exons constitute the open reading frame (ORF) that encodes a protein, but they also contain untranslated regions (UTR) that are located in 5'end and 3'end of a gene. Introns are removed from pre-mRNA by an

enzymatic complex called the spliceosome. The number and size of introns in different species and in genes of the same species are different. The introns of genes in higher organism, such as flowering plants, can be much longer than the nearby exons. Each gene has regulatory regions that can regulate gene expression. All genes are flanked by a regulatory region that includes a promoter that binds the RNA polymerase and numerous transcription factor binding sites necessary for transcription initiation. The promoter regions contain consensus sequences such as TATA box, CAAT box and GC box. TATA box is located in 25-35 base pairs upstream of the transcription start site and position RNA polymerase II to the transcription start site. Most house-keeping genes have GC boxes instead of a TATA box that is located within 100 base pairs upstream of the start site. Transcription in many eukaryotes can be up-regulated by regulatory elements, called enhancers that may be located upstream or downstream of a promoter, within an intron, or even downstream from the last exon of a gene.

RNA genes, also called non-coding RNA (ncRNA) genes in a genome, are transcribed to structural, catalytic or regulatory non-coding RNA transcripts that do not encode proteins (Eddy et al. 2001). Non-coding RNAs mainly include tRNA, rRNA, small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and microRNA (mRNA).

1.4.3 Genome

A genome is the entire genetic material in an organism, i.e. DNA (or RNA in some

viruses) that is composed of genes and non-coding DNA. As mentioned above, genes can be divided into protein-coding genes and stable RNA coding genes and therefore a gene is defined as a region of a genome that is transcribed into RNA. Some of the protein-coding genes that are transcribed into messenger RNA, or its precursor, are unique, in the genome, i.e., there is only one copy in a genome while other protein-coding genes have multiple copies, with the former called singletons and the latter typically called gene families. A gene family is a set of genes that evolved from a single ancestral gene and arose from gene duplication. A protein family is a group of homologous proteins encoded by a gene family. Gene duplication may result from an unequal crossover recombination (Fitch et al. 1991), a retrotransposition event, or duplication from a whole chromosome (Zhang et al. 2003). Gene duplication is considered a very important evolutionary driving force after the emergence of the common ancestral gene (Taylor 2004, Ohno 1970). Genome duplication is common in all living organisms, especially in plants. For example, the genome of *Medicago sativa,* a tetraploid, was duplicated at least once. Some genes have similar sequences to functional genes but they are not functional; these genes are called pseudogenes. Pseudogenes may arise by gene duplication, but their function was lost during evolution after duplication due to frameshift and/or nonsense mutations (Mighell et al. 2000). Pseudogenes also may result from retrotransposition-related process when the mature mRNA is reversely transcribed and inserted back into the genome (Vanin 1985). Therefore, these kinds of pseudogenes lack introns and upstream promoter sequences.

When a genome is sequenced, the focus is on finding protein-coding genes, stable coding RNA genes and their regulatory elements. However, repetitive DNA sequences cannot be ignored in a genome because they constitute a significant portion of the genome in higher eukaryotes. The amount of repetitive DNA is very high in plants. For example, in pea, over 95% of the DNA sequence is estimated to be repeats (Thompson et al. 1980). Therefore, the role of repetitive DNA is crucial in the determination of chromosome size and structure (Flavell 1986). There are two types of repetitive DNA sequences in genome: tandem repeats and interspersed repeats. Tandem repeats are a cluster of the same DNA sequence (2 or more nucleotides in length) located adjacent to each other. Tandemly arranged DNA sequences include different kinds of satellite DNA, the telomeric repeat, and the rDNA. They lie at specific positions of chromosomes, such as pericentromer, subtelomer or telomere (Kubis et al. 1998). Simple sequence repeats (SSRs) (Jacob et al. 1991) or microsatellites, one type of tandem repeats that are clusters of 2 to 6 nucleotides (usually di-, tri-, tetranucleotide repeats). They are repeated a few to hundreds of times in eukaryotic genomes. The repeat (TA)n is the largest group of SSR repeats in plants (Gianfranceschi 1998). SSRs are length polymorphic and are one of the most important molecular markers. The interspersed repeats are distributed throughout the genome and interspersed with other sequences. These types of DNA repeats mainly contain transposable DNA elements and their remnants (Kubis et al. 1998). Transposable DNA elements, the first of which the Ac-Ds control element was observed in maize (McClintock 1951), are divided into class I and

class II based on the method of transposition (Schmidt 1999, Flavell et al. 1994). Class II elements are called transposons since they move through DNA intermediates and include the Ac-Ds, En-Spm, Mu transposons, and miniature inverted repeat transposable elements (MITEs). Class I transposable elements contain retrotransposons and other retroelements. Retrotransposons are DNA sequences that move in a genome via RNA intermediates. They constitute a large proportion of a plant genome. There are two major types of retrotransposon. One is long terminal repeats (LTR), and the other is the non-LTR retrotransposon. LTR retrotransposons can be divided into two subgroups: Ty1-copia-like and Ty3-gypsy-like retroelements. They contain up to three open reading frames (ORFs) and are flanked by LTRs. Non-LTR retrotransposons include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LINEs do not have LTRs but have a poly(A) tail at their 3' termini. LINEs are several kilobases in length and consist of two ORFs that code for a gag protein, and endonuclease and reverse transcriptase, respectively. They have the ability for autonomous retrotransposition. LINEs are considered as a major component of plant genomes (Schwarz-sommer et al. 1987; Wright et al. 1996). SINEs are up to several hundred bases long and contain a tRNA-derived region, unrelated DNA sequence, and a LINE-related region. SINEs are dependent on the reverse transcriptase supplied by LINEs for retrotransposition (Luan et al. 1993). Most plant LINEs and SINEs are transcriptionally inactive.

1.5 The history of DNA sequencing methods

In 1977, two sequencing approaches were described by different groups (Maxam & Gilbert 1977; Sanger et al. 1977). The Maxam-Gilbert sequencing method utilizes different chemical reactions to cleave radioactively labeled double-stranded DNA at specific base positions. The Sanger, or chain termination method, utilizes dideoxynucleotides of the four nucleotides in the enzymatic DNA synthesis reaction to terminate DNA chain growth. The Sanger method is widely used today and was used during this dissertation research. Therefore, it will be discussed now.

1.5.1 The Sanger dideoxynucleotide DNA sequencing method

Sanger's chain termination sequencing method uses dideoxynucleotides (ddNTPs) to terminate the synthesis of DNA in the presence of a DNA template, the four deoxynucleotides (dNTPs), four ddNTPs, a primer, DNA polymerase, magnesium, and buffer. The primers anneal specifically to the DNA template and function as a starting site for the synthesis of a new DNA strand. Then, a corresponding nucleotide is attached to the end of the primer in a reaction catalyzed by DNA polymerase to form a 3', 5' phosphodiester bond between the last nucleotide of primer and the newly added nucleotide. According to base pairing rules, since an A pairs with a T and a G pairs with a C, if the template DNA has an A at a position, then a T will be esterified to the

corresponding position of the new DNA strand and vise versa. The new synthesized DNA continues extending until a terminator ddNTP is attached to the end of the new DNA. The absence of oxygen in 3' position of deoxyribose of ddNTP terminates the synthesis of the new DNA strand.

Radioisotopes were used originally in both the Maxam-Gilbert and Sanger methods to label and detect DNA fragments. Later, fluorescence dyes were introduced to replace the radioactivity and allow for automation of DNA sequencing. The primer-labeling method and the terminator-labeling method are utilized to fluorescently label the products of the DNA sequencing reaction. In the primer-labeling method, the 5' end of the primer is attached by fluorescent dyes. Four different dye-labeled primers were used initially. More recently, in the terminator-labeling method, the dyes are attached to the ddNTPs instead of to the primers. This approach only requires one reaction tube because each ddNTP has a different dye.

1.5.2 Massively parallel pyrosequencing, the GS20 and FLX systems

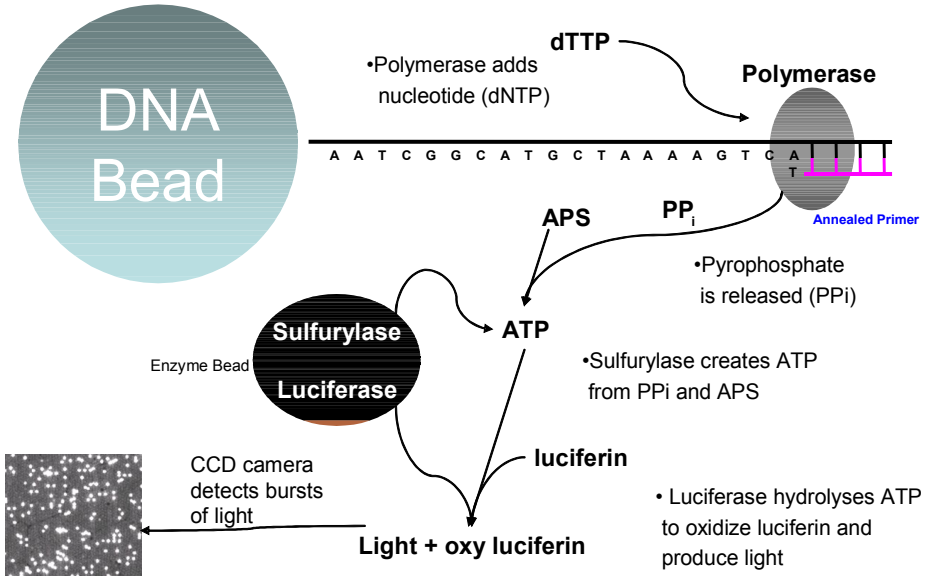


Figure 1.3 The process of pyrosequencing reaction

In 2005, Margulies M. et al (Margulies et al. 2005) described an ultra-high throughput automated DNA sequencing system based on massively parallel pyrosequencing that can generate over 20 M bases in a single 4.5-hour run. The software of this GS-20 system performs mapping or de novo assembly for genomes up to 50 M bases. The system is fast, cost-effective, simple, efficient, and convenient compared to the conventional Sanger technology. In this system, several million DNA templates are immobilized on each bead after sample preparation. When the sequencing reaction occurs (shown in Figure 1.3), nucleotides complementary to the template strand are

added into the growing DNA strand by the DNA polymerase enzyme. Every addition of a nucleotide releases one pyrophosphate (PPi) molecule. PPi and adenosine phosphosulfate (APS) are converted into ATP by sulfurylase. Luciferase enzyme hydrolyzes ATP and oxidizes luciferin to produce light and oxy-luciferin. The intensity of the light is proportional to the number of nucleotides added. This light is captured by a charge-coupled device (CCD) camera and then converted into a digital signal. The computer of the sequencer then combines the signal intensity with the positional information on the PicoTiterPlate device to determine the sequence of hundreds of thousands of individual reactions at the same time and generate millions of nucleotides per hour. The whole process for this sequencing method includes DNA library preparation and titration, emulsion-based clonal PCR (emPCR) amplification, sequencing by synthesis, and data analysis.

More recently, 454/Roche introduced the GS-FLX system that is more sensitive and powerful than GS20. It can generate reads up to 250 bp in length, compared to the 100 bp generated by GS20 system.

2. Materials and Methods

2.1 Sequencing strategies

In this present work, the mapped BAC-based shotgun sequencing strategies was used to sequence the euchromatic arms of chromosomes1, 4, 6, and 8 of *Medicago truncatula*. This procedure is illustrated in Fig 2.1. The construction of the BAC library (120kb HindIII library) and the mapping of the BAC clones into large contigs were done by the University of California-Davis group.

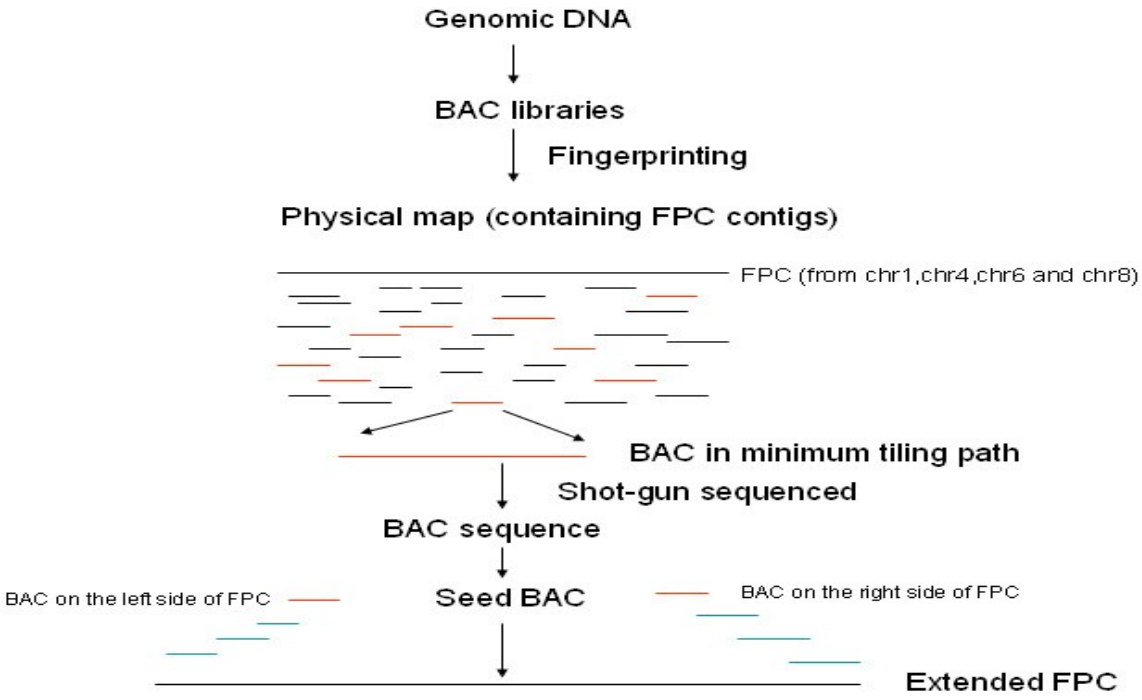


Figure 2.1 Idealized representation of the mapped BAC-by-BAC shotgun sequencing strategy

One thousand BAC seed clones were obtained from the UC Davis group and sequenced

in our lab. These clones contain known markers, or regions, of biological interest. They were fingerprinted to create a physical map for *M.trunctula*, and many of them have already been anchored to the genetic map by linkage analysis of simple sequence repeats (SSRs) and the use of EST. BAC ends that correspond to the seed BACs were located in BACs that could extend into the gaps between contigs. Since our lab is sequencing chromosome 1, 4, 6 and 8 of Medicago and the location of many seed BACs are already known, after sequencing the 1000 seed BACs, additional BACs from these 4 chromosomes were chosen for sequencing to extend contigs on these chromosomes. For most of the sequencing, we used the traditional Sanger sequencing technology, but more recently we have developed a combined Sanger-454 approach, which will be discussed later.

2.1.1 Sanger sequencing

2.1.1.1 Large scale BAC DNA isolation

The purpose of this procedure is to isolate BAC DNA from the *E. coli* host by using modified alkaline lysis procedure (Birnboim et al. 1979) followed by ethanol precipitation. First, *E. coli* cells containing BAC DNA were grown in LB media to get a large amount of BAC DNA. The cells were streaked onto the small LB Petri dish containing 15 μg/ml chlorophenicol. After incubating at 37 ^{o}C overnight, colony smear was picked using a bioloop and incubated into a falcon tube containing 3 ml of LB

media plus chlorophenicol. The tube then was incubated for 8-10 hours at 37°C in a shaker shaking at 250 rpm. After incubation, 3 ml of culture was transferred into a flask containing 200 ml media and then shaken for 8-10 hours. Cells then were collected by centrifugation at 5,000 rpm for 15 min in the RC5-B centrifuge at 4 °C. The supernatant was discarded into a container to be autoclaved later and the cell pellets were stored at -70 °C.

The following steps were performed to isolate BAC DNA from the *E.coli* cells. The cell pellets were thawed and resuspended in 8 ml of 10 mM EDTA (pH 8.0) using an Eppendorf pipette. 16 ml of alkaline lysis solution (1% SDS and 0.2N NaOH) was added very gently to lyse cell membrane and release the cell contents. The bottle was not shaken, so as to avoid shearing chromosomal DNA. The bottle was kept in an ice-water bath for 5 min and then 12 ml cold 3 M potassium acetate (KOAc) was added. The bottle was incubated in an ice-water bath for 30 min. The KOAc precipitates proteins, cell membrane debris, and chromosomal DNA and neutralizes the pH to renature single-stranded DNA to double-stranded (chromosomal DNA precipitates with proteins though it can be partially renatured, however, plasmid DNA can completely be renatured into double strand and remains in solution). After centrifuged at 10,000 rpm for 15 min to pellet SDS, proteins, chromosomal DNA, and cell debris, the supernatant was filtered through double-layered cheesecloth into an autoclaved bottle, and an equal volume of isopropanol was added to precipitate the BAC DNA. After mixing and then standing for 5 min at room temperature, the sample was centrifuged at 5000 rpm for 15

min and then supernatant was discarded. The DNA pellet was dissolved in 3.6 ml of 10:50 TE, transferred the solution into a 50 ml Sorval centrifuge tube and 1.8 ml of 7.5 M KOAc was added to remove residual chromosomal DNA and proteins and stored at -70 °C for at least half an hour.

After thawing at room temperature and then centrifuging at 10,000 rpm for 10 min, the supernatant was transferred into a 50 ml Corning tube and then appropriate amount of RNase A and T1 was added to digest RNA. The sample then was incubated in a water bath for 1 hour at 37 °C followed by the addition of 30 ml of 95% ethanol to precipitate DNA. After centrifugation at 3000 rpm for 25 min to collect the DNA pellet, the supernatant was discarded and the pellet was washed with 30 ml 70% ethanol and dried in the vacuum dryer or on the bench overnight.

2.1.1.2 DNA Fragmentation, size selection and insertion into pUC vector

After isolation, large-sized BAC DNA (100-150 kb) was sheared into small-sized DNA in the hydroshear (made by Gene Machines) (Oefner et al. 1996). Here the DNA pellets were dissolved in 250 µl ddH_2O. After electrophoresing 5-10 µl on an agarose gel to check the concentration of DNA (Studier 1973), the sample was centrifuged for 30 min at 12,000 rpm in the cold room and then the supernatant was transfered to a clean microcentrifuge tube. Next 100 µl of DNA solution was chilled to 0 °C and sheared at a speed code of 10 in the hydroshear for 20 cycles. The sheared DNA fragments were

precipitated with 2 volumes of 95% ethanol containing 0.12 M NaOAc and collected by centrifugation at 12,000 for 15 min in cold room. After washing the fragments with 70% ethanol, the sheared DNA was dried in vacuum dryer for at least 15 min.

After hydroshearing, DNA fragments have overhangs. Therefore, they needed to be end-repaired before they were ligated to blunt-ended vectors in the ligation step (Pan et al. 1994, Bankier et al. 1987). In the end-repair procedure, T4 DNA polynucleotide kinase and Klenow DNA polymerase were used to add phosphate group to 5' end and add nucleotides to pair with overhangs, respectively. Here the sheared DNA was dissolved in 27 µl autoclaved ddH_2O and 5 µl 10X kinase buffer, 5 µl 10 mM rATP, 7 µl 0.25 mM dNTPs, 1µl T4 DNA polynucleotide kinase and 2 µl Klenow were added. After incubation in a 37 ^oC water bath for 30 min, the reaction was halted by heating the tubes in a 70 ^oC water bath for 10 min to denature enzymes.

Since only 1-4 kb DNA fragments were needed, the DNA fragments were size selected on a low melting agarose gel after electrophoresis at 12 mA for 1-1.5 hours. Fragments in the 1-4 kb size range were excised with a sterile razor blade, placed into Eppendorf tube and frozen at -70 ^oC.

The DNA was extracted from the low melt gel by phenol (Sambrook et al. 1989). Here the gel containing 1-4 kb DNA was melted in a 70 ^oC water bath. An equal volume of TE-saturated phenol was added and mixed on a vortexer for 30 seconds. After centrifugation at 12,000 rpm for 5 min, the upper 90% of the aqueous layer containing DNA was transferred into a new tube, and then an equal volume of water-saturated ether

was added. After vortexing, it was centrifuged at 12,000 rpm for 3 min. The upper ether layer was discarded and the ether extraction was repeated. The tube then was placed in a vacuum dryer until the total volume was less than 700 µl, at which time the DNA was ethanol precipitated and ligated into SMA1-CIP treated pUC18 vector. The total volume of the ligation mix was 10µl, including the following reagents: X µl of DNA, 1 µl of 10X Ligase Buffer, 2 µl of pUC18 SMA1 vector, 1 µl of T4 DNA ligase 400 U/ul and Y µl of sterile ddH_2O. X plus Y should be 6. The tube with the ligation mix was then centrifuged at 1000 rpm for 1-2 seconds and could be stored at 4 ºC for 24-48 hours.

2.1.1.3 Subclone generation and isolation

The DNA was transformed into electro-competent *E. coli* cells by electroporation (Rakesh et al. 1996). Here 2.5µl of ligation mix was transferred into *E.coli* XL1blue-MRF competent cells, mixed, and then transferred into an electroporation cuvette. The cuvette was placed in the electroporation chamber, and 2.5 kV was applied. After centrifugation at 2500 rpm for 5 min, the supernatant was discarded and 25 µl of X-gal and 25 µl of IPTG were added to each tube to resuspend the transformed cell pellets. Transformed cultures were plated on LB plus ampicillin agar plates. The LB plates were dried in a hood, inverted, and then incubated at 37 ºC for 20 hours. After storing in the cold room to intensify the blue color, the Genetix colony picker was used to robotically pick the white, insert-containing colonies into 384-well microtiter plates containing TB

media and ampicillin. This was followed by shaking for 22 hours at 37 °C at 520 rpm with oxygen in a Genemachines HiGro incubator.

The DNA was automatically isolated from these cells after they were collected by centrifugation at 3000 rpm for 10 min, and stored at -80 °C for at least 30 min to 1hour. After thawing, 23 µl TE-RNase was added to each well using the Zymark station (Sciclone ACH500) and shaken at 1800 rpm for 8 min. Subsequently, 23 µl alkaline lysis solution was added and after shaking at 1800 rpm for 8 min, 23 µl 3 M NaOAc (pH4.5) was added and shaken at 1200rpm for 10 min. The plates were frozen at -80 °C overnight.

Next the plates were thawed and centrifuged at 3200 rpm for 45 min to clear the lysate. 40 µl of the supernatant then was transferred into a new 384-well plate using the Velocity11Vprep and 40 µl of 100% isopropanol was added using the Vprep. After standing at room temperature for 5 min and centrifuged at 3000 rpm for 30 min, the supernatant was discarded invertedly on paper towel. 50 µl of 70% ethanol then was added using the Vprep, and the plates were centrifuged at 3000 rpm for 10 min, decanted, and dried on bench or in a vacuum dryer. The DNA then was dissolved in 20 µl of ddH_2O and a portion was analyzed by agarose gel electrophoresis.

2.1.1.4 Subclone DNA sequencing

The Sanger chain-termination method (Sanger et al. 1977) was used to sequence the

DNA. Here 6 µl of DNA template was dispensed into a viper plate from 384-well microtiter plate using Hydra (Robins Scientific, Inc) or V-prep (Velocity 11. Inc.) robots. Then 2 µl of 1:16 BigDye or ET reaction mix (containing dNTPs, buffer, ddNTPs with dye attached, DNA polymerase, primers) was added to each well and the reaction mix was concentrated in the bottom of the tubes by centrifuging at 1000 rpm for 2-3 seconds. The plates were covered with a rubber plate sealer, placed into the thermocycler and incubated at 60 cycles consisting of a denaturation at 95 °C for 30 second followed by rapid thermal ramp to 50 °C, and an annealing at 50 °C for 20 second, rapid thermal ramp to 60 °C, and an extension at 60 °C for 4 min. Unreacted dNTPs, ddNTPs, primers were removed from the incubation mix by precipitating the resulted nested fragment set with 95% ethanol/0.12NaOAc followed by centrifuging 30 min at 3200 rpm. After washing the pellets with 70% ethanol and centrifuging 10 min at 3200 rpm, supernatant were decanted as before and the plates were covered with KimWipes and dried on bench after covering with KimWipes. The plates then were sealed and stored at -20 °C.

2.1.1.5 Sample loading and data analysis

Prior to loading the Applied Biosystem Inc (ABI) sequencers, 15 µl of 0.1 mM EDTA was added into each well using Hydra and then the plates were shaken for about 1.5 hours. The sample then was loaded onto an ABI 3730 fluorescence-based capillary

sequencer (Deschamps et al. 2003). The DNA sequence data was collected and then transferred to Sun computer workstations for automated base-calling with Phred (Ewing et al. 1998). The data was assembled into contigs with Phrap and viewed and analyzed with both Consed (Gordon D. et al. 1998) and Exgap (Hua A et al. 2003).

2.1.2 Massively parallel pyrosequencing on the 454 GS20 system

Massively parallel pyrosequencing on the 454 GS20 system includes 4 major steps: library preparation, emPCR, sequencing, and data analysis (Margulies et al. 2005). When this technology was first introduced in our lab, we made single-stranded libraries and purified them using Qiagen columns. However, more recently we have been making double-stranded libraries and purifying them using solid-phase reversible immobilization (SPRI) magnetic beads (DeAngelis et al. 1989). I will describe these two different approaches separately. The steps for emPCR and sequencing are almost the same for the two different libraries except for small details which I will discuss later.

2.1.2.1 Library preparation

2.1.2.1.1 Single-stranded template DNA (sstDNA) library

A single-strand template DNA (sstDNA) library was prepared using several steps including nebulization, end polishing, adaptor ligation, library immobilization, fill-in

reaction, library isolation, and assessment using RNA 6000 Pico assay.

One sample preparation is enough for each genome despite its size and no cloning and colony picking were needed during library preparation. To make the library, the genomic DNA was sheared to 300-500 bp-long fragments by nebulization (Bodenteich et al. 1994), followed by purification on MinElute PCR purification columns. The nebulized DNA was end polished (blunt-ended and phosphorylated) by adding T4 polynuleotide kinase, T4 DNA polymerase, dNTPs, ATP, BSA, and buffer. Adaptors A and B (44bp) then were attached to the ends of the end-polished DNA fragments under the catalysis of T4 ligase. 20 bp of the adaptor provides priming sequences for emPCR amplification, 20bp is sequencing primer and 4bp is base key that is used by the software for base calling. Adaptor B has a biotin tag at its 5'end that attaches the DNA library onto streptavidin-coated immobilization beads. There are 4 kinds of fragments with different adaptors. Only the fragments with B adaptors could attach to the immobilization beads. After ligation, nebulized, end polished, and adaptor-ligated DNA fragments were immobilized on immobilization beads followed by fill-in reaction. During fill-in reaction, DNA polymerase, dNTPs, and buffer were added to make the DNA fragments blunt ended. Next, the melt solution (NaOH) was added to denature the double-stranded DNA, and then the non-biotinylated strand was released. These non-biotinylated strands that have an A adaptor at one 5'end and a B adaptor at another 5' end, were purified and used as an sstDNA library. The sstDNA library was assessed for its concentration and average size using the RNA 6000 Pico assay and then titrated to

determine the optimal amount (1 copy DNA per bead) required for emPCR.

2.1.2.1.2 Double-stranded template DNA (dstDNA) library preparation

For the preparation of dstDNA, DNA is nebulized, end-polished, and ligated with adaptors, as that in sstDNA library preparation. After being ligated with adaptors, overhangs on DNA fragments are fixed by fill-in reaction directly instead of being immobilized onto immobilization beads and denatured by alkaline solution. Another difference is that we use SPRI beads to purify dstDNA instead of MinElute PCR purification column used in sstDNA library preparation. After dstDNA library is made, it is analyzed by the Caliper AMS90. The number of DNA molecules per µl was computed from the concentration (ng/µl) resulting from the caliper AMS-90 using the molecules calculator spreadsheet.

$$\text{Molecules/µl} = \frac{(\text{sample conc.; ng/µl}) \times (6.022 \times 10^{23}\ \text{mol./mole})}{(656.6 \times 10^{9}\ \text{gram/mole}) \times (\text{fragment length; nt})}$$

The dsDNA library stock then was diluted to the ratio of 1:4 and stored at -20°C for later use.

2.1.2.2 EmPCR (for both dsDNA and ssDNA libraries)

This step amplifies DNA using emulsion-based clonal DNA PCR. The purified and quantified sstDNA library was immobilized to the capture beads by hybridizing to the

complementary primers attached to the capture beads. For the dstDNA library, the above annealing step was omitted. The DNA concentration and bead number are optimized so that each bead contains 1 DNA segment or less on average. We use different equations to calculate DNA concentrations for dsDNA and sstDNA libraries. The capture beads attached by sstDNA library, or the beads and dstDNA mixture, were emulsified using the amplification mix including platinum HiFi Taq polymerase, pyrophosphotase, primer mix, $MgSO_4$, and buffer in a water-in-oil mixture. Each bead stayed within its own microreactor containing the complete amplification reagents. The PCR amplification takes place in these microreactors. The forward PCR primer was biotinylated for later use. All microreactor reactions occurred in parallel, resulting in bead-immobilized clonal amplified DNA fragments. After amplification, the microreactors were broken by adding isoproponal, followed by centrifugation. Next, streptavitin-coated magnetic enrichment beads were added to catch the positive capture beads containing biotinylated primers. After the removal of the uncaptured waste beads on a magnetic particle collector (MPC), melt solution (NaOH) was added to release the captured positive beads from enrichment beads and denature the double stranded DNA bound to the capture beads to single stranded DNA. Then the beads rich in single-stranded DNA were separated from the enrichment beads on MPC. Sequencing primers were annealed to the above DNA beads on thermocycler by running the annealing program. Then bead counting was performed for calculating the number of loading beads. The DNA beads then were mixed by vortexing. Five µl of beads were added into

2 ml of Coulter Counter cuvette. After swirling the cuvette to mix and placing it on the station of Coulter, the beads were measured and the number of beads was used to calculate the volume of beads required to obtain 300,000 beads per region that was to be loaded onto the PicoTiterPlate.

2.1.2.3 Sequencing

To load DNA beads onto the PicoTiterPlate for sequencing, the DNA beads first were combined with enzyme beads and packing beads and then centrifuged into wells of PicoTiterPlate. The diameter of each well is designed to allow only one DNA bead to be deposited into each well. Sulfurylase and luciferase required for sequencing were located on the enzyme beads. To generate the 200,000 high quality sequencing reads on average for each PicoTiterPlate, the sequencing reagents (containing buffers and nucleotides) were flowed across the wells of the plate and nucleotides were flowed in a fixed order: TACG. Nucleotides complementary to the template strand were added into the growing DNA strand by DNA polymerase enzyme. Every addition of a nucleotide releases one pyrophosphate (PPi) molecule. PPi and adenosine phosphosulfate (APS) were converted into ATP by sulfurylase. Luciferase enzyme hydrolyzed ATP and oxidized luciferin to produce light and oxy-luciferin. The light then was captured by the CCD camera. The intensity of the light was proportional to the number of nucleotides added. The sequential flow of the four nucleotides was repeated for 42 cycles, and

generated an average read per well of 100 bases for GS20 system (for FLX system, length was 250 bases per read). Twenty million bases (100bp x 200,000 reads) were generated in 5.5 hours. The FLX system, a more sensitive system, produced more than 100 million bases per 7.5-hour run with the average yield of 400,000 reads per run.

2.1.2.4 Data analysis

After the signal was captured by the CCD camera, it was processed. There the flowgrams and base-called sequences with corresponding quality scores were generated after image acquisition, image processing, and signal processing. A nucleotide sequence of each well was produced in the form of a flowgram. Every flowgram began from the 4 base key sequences that were utilized to identify and calibrate the wells. The length of homopolymer could be determined because the signal strength is proportional to the number of based added. Flowgrams were used as input information for training, assembling, or mapping application. Final read sequences were generated by a training application to improve the base calling accuracy. The assembly application used in *de novo* sequencing assembled the final reads into contigs and produced a consensus sequence of the whole DNA sample. The mapping application used in resequencing maps the final reads to the reference sequence and produced the consensus DNA sequence and the final output.

2.1.3 Paired end sequencing in GS20/FLX system

Although the GS20 system has many advantages over traditional sequencing methods, it does have some shortcomings. One of the shortcomings is that the contigs resulting from the GS20 are unoriented and unordered, which causes difficulty in closing gaps. To overcome this, a paired-end sequencing approach was developed that aids in ordering and orienting contigs (Javie 2006, Korbel et al. 2007) generated by sequencing shotgun genomic libraries. This method entails preparing a separate paired-end library from the same DNA sample used to produce shotgun DNA data. The paired-end library preparation differs from shotgun library preparation as it uses special adaptors, primers, and enzymes. The libraries then could be amplified using the same kit and protocol as for shotgun dsDNA emPCR and then sequenced using the GS20 sequencing kit, the GS PicoTiterPlate kit, and the GS20 or FLX instrument. The procedure of paired-end DNA library preparation was described briefly as follows. The DNA sample was cleaved into 2-3 kb fragments using the hydroshear. The fragments then were methylated to protect internal *EcoR*I sites from digestion by *EcoR*I enzyme and made blunt ended by DNA fill-in polymerase. Biotinylated Hairpin adaptors with non-methylated *EcoR* I and *Mme* I sites were attached onto both ends and exonuclease was added to digest any remaining DNA without ligated hairpin structures. *EcoR*I was added to cleave the hairpin structures and create sticky ends that then were ligated to circularize. The circular DNA then was sheared to about 500 bp fragments in the nebulizer. The 19 bp paired-end

adaptors were ligated onto either end to serve as priming sequences for both amplification and sequencing and as a 4 base key sequence utilized for base calling. Since there were gaps when adaptors were ligated with paired-end library DNA fragments because the ends of adaptors are not phosphorylated, a fill-in reaction using *Bst* DNA polymerase displaces the nicked strands and extends the strands to full length. The paired-end library then was amplified by PCR, quantitated on caliper and diluted to appropriate concentration for further emPCR.

2.1.4 Pooling strategies

To save money and labor, different pooling strategies were used during pyrosequencing (shown in Figure 2.2 A and B). For the first pooling strategy, we pooled BACs without adding tags. Taking 100 BACs as an example, 10 horizontal and 10 vertical pools were grown, nebulized, and 20 libraries were made as the steps in library preparation followed by emPCR and sequencing. In combination with the sequence data from 3730 sequencer and/or the paired end sequencing, each horizontal pool was compared with each vertical pool and the sequences for the intersectional BAC were obtained.

As to the second pooling strategy, each individual BAC was grown and nebulized. Library was made for each BAC as the steps in library preparation except that in adaptor ligation step, a multiple identifier (MID) adaptor containing a unique 10 nucleotide sequence instead of a common adaptor was ligated to DNA fragments. After library preparation, 12 DNA libraries with 12 different MIDs were pooled together and

quantitated followed by emPCR and sequencing. The sequences for each BAC with a unique MID tag were separated by the GS FLX analysis software that allows for automated grouping and analysis of MID-containing reads.

A

	VP1	VP2	VP3	VP4	VP5	VP6	VP7	VP8	VP9	VP10
HP1	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
HP2	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
HP3	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
HP4	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
HP5	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
HP6	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
HP7	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
HP8	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
HP9	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
HP10	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10

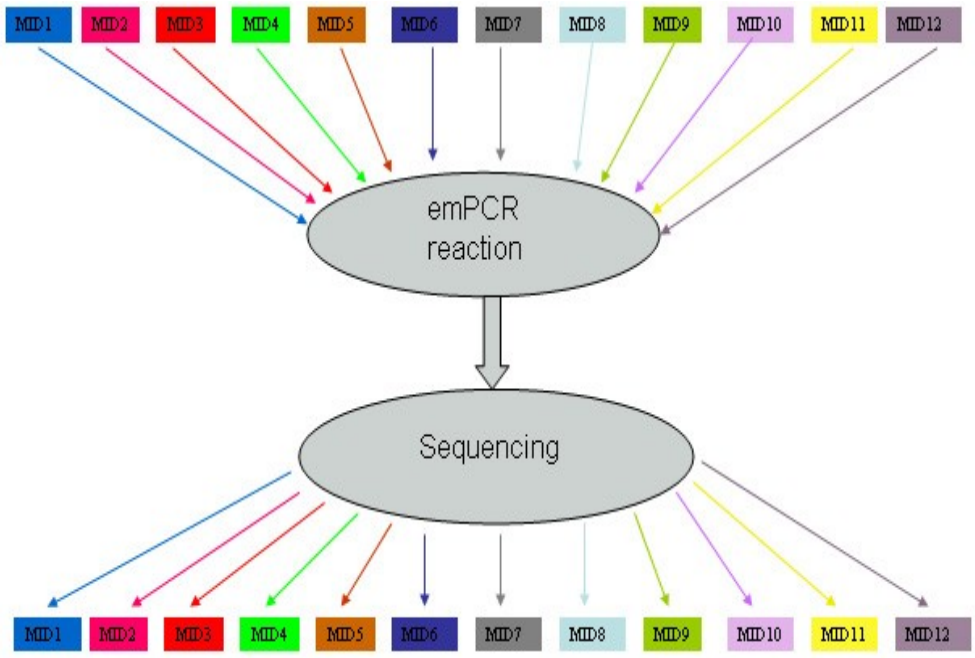


Figure 2.2 Pooling strategies in 454 technology: (A) BAC arrays in the first pooling strategy: VP stands for vertical pool and HP stand for horizontal pool; (B) The MID workflow in the second pooling strategy

2.2 Computational tools in DNA sequence analysis

After completely sequencing the Medicago genome, the genes were predicted using GENSCAN (Burge et al. 1997) and FGENESH (Salamov et al. 2000) after first masking out repeats using Repeatmasker (Smit and Green, 1999). Then, GeneSplicer (Pertea et al. 2001) or Splicepredictor was used to determine HnRNA processing splicing sites and tRNA scanSE (Lowe et al. 1997) was used to predict tRNA genes. The database of plant cis-acting regulatory DNA elements (PLACE) was used to recognize plant regulatory motifs conserved in several species (Higo et al. 1999). Programs such as RepeatMasker (Smit and Green, 1999), Signal Scan, and CpGFinder also were used to identify other

M.truncatula genomic features. To identify evolutionary conserved genomic features, tools such as Blast (Altschul et al. 1990), Crossmatch (Smith et al. 1981, Gordon et al. 2001), and clustal W (Larkin et al. 2007) were used to compare the *M. truncatula* genome data with that from *A. thaliana*, *Oryza sativa*, *Populus trichocarpa, Lotus japonicus*, and *Glycine max.* Gene Ontology (GO) was annotated automatically using Blast2GO (Götz et al. 2008) on the *M. truncatula* protein-encoding genes by searching InterPro domains.

GENSCAN, an *ab initio* gene prediction program that is based on a probabilistic model of composition/gene structure properties (Burge et al. 1997), captures the general and specific compositional properties of a eukaryotic gene: exon, intron, splice site, promoter and provides information that is independent and complementary to that provided by homology-based gene identification methods such as BLASTX. The substantial differences in gene density and structure (e.g. intron length) that exist between different C+G% compositional regions were considered in this program. GENSCAN predicts multiple genes in a sequence, where genes can be partial or complete, and genes occurring on both DNA strands. It also indicates the reliability of each predicted exon. FGENESH (Salamov et al. 2000) is the HMM-based gene prediction program with the algorithm similar to GENSCAN. The difference between these two programs is that GENSCAN includes known sequence features that include splice sites and start sites. A gene with identity or high homology to a protein was classified as ‘putative’ and a ‘–like protein’ based upon the protein name. A gene

without similarity to other proteins but with EST homology was termed as 'unknown' protein. A gene predicted by two prediction programs but without protein or EST homology was defined as a 'hypothetical' protein according to IRGSP standard (Leung et al. 2002). Genes predicted by one prediction program also were termed possible hypothetical proteins.

Genesplicer (Pertea et al. 2001), a program that can identify splice sites in eukaryotic mRNA is useful for *ab initio* gene prediction, after the exons and introns first are located. In most eukarotes, GT (GU in mRNA) and AG are located in the 5' boundary or donor site of introns and the 3' boundary or the acceptor site of introns, respectively. Besides these dinucleotides, other consensus sequences also are found near the donor or acceptor site, such as a pymimidine-rich region preceding the AG at the acceptor site, or a shorter sequence following the GT at the donor site. The spliceosome complex recognizes these consensus sequences and removes the introns from the hn-mRNA to generate the mature RNA. Although several programs, such as NetPlantGene, Netgene2, HSPL, NNSplice, GENIO, SpliceView and GeneSplicer also could be used to predict these splice sites, GeneSplicer, considered the best predictor in terms of its accuracy and its efficiency, was chosen as there was no limit for input sequence length, for example it could process 20Mb of chrII of *A.thaliana,* whereas other programs limit the submitted sequence length to a few kilobases. The method in GeneSplicer was derived from the maximal dependence decomposition (MDD) (Burge et al. 1997) and was improved with Markov models.

The program tRNAscan-SE (Lowe et al. 1997) combines three tRNA prediction approaches, i.e. tRNAscan1.3 (Fichant et al. 1991), Pavesi algorithm (Pavesi et al. 1994), and covariance model search program covels (Eddy et al. 1994) to obtain the speed, sensitivity and specificity. tRNAscan1.3 (Fichant et al. 1991) was the most widely used RNA detection program in which each tRNA candidate must have two intragenic promoters and can form base pairings in tRNA stem-loop structure. Pavesi (Pavesi et al. 1994) et al designed a different algorithm that only looks for linear sequence signals existing as eukaryotic RNA polymeraseIII promoters and terminators. Covariance models (Pavesi et al. 1994) use both primary consensus and secondary structure information to find putative tRNAs and thus tRNAscan-SE can identify 99-100% of transfer RNA genes in DNA sequence with a false positive rate of less than one per fifteen billion nucleotides. The first two programs were used to run input sequences to screen out the candidate tRNAs which then underwent the scrutiny of *covels.* Besides wild-type tRNA genes, tRNAscan-SE also could detect tRNA-derived SINEs and tRNA pseudogenes.

RepeatMasker, a program searching DNA sequence for repetitive elements and low complexity, was used to screen a query sequence before the query sequence was utilized to search against a database. After using RepeatMasker, a modified query sequence was generated in which all the identified repeats and low complexity sequences were masked by “N”s. A detailed table also was produced to annotate all identified repeats or low complexity sequence such as small RNA pseudogenes, LINEs, SINEs, LTR

elements. RepeatMasker was used in a database search to avoid producing misleading results, because a major proportion in plant genome comprises repetitive or low-complexity sequences.

Cross-match, a program designed for quickly comparing protein and nucleic acid sequences and for database search which used the slightly modified Smith-Waterman-Gotoh algorithm (Smith et al. 1981, Gotoh 1982), was used to compare sequences in RepeatMasker. Cross-match also was utilized for several other tasks. First, it was used to compare reads to vector sequences to screen out the vector. Second, it was use to find the overlap or repeats between two sequences. Third, it was used to search a profile against a database.

The Basic Local Alignment Search Tool (BLAST) (Altschul et al 1990), the most widely used program to compare sequences and search for their similarity, aligns sequences by measuring local similarity in the form of the maximal segment pair (MSP) score. BLAST is a robust sequence comparison tool, and it is one order of magnitude faster than other existing similar tools. Blast was used for DNA and protein sequence database searches, motif searches and gene identification searches. Since BLAST generated a bit score and an expect value (E value) for each alignment using statistics, the higher the score, the better the alignment and the lower the E value, the more significant the homology was. Different BLAST programs were used when different query sequences were compared with different databases. BlastP was used to compare an amino acid sequence with a protein database. BlastN was used to compare a

nucleotide sequence with a nucleotide database. BlastX was used to compare a nucleotide translated into all reading frames with a protein database. TblastN was used to compare a protein sequence with a nucleotide database translated into all reading frames. TblastX was used to compare a nucleotide sequence translated into six reading frames with a nucleotide database translated in all reading frames.

Gene Ontology (GO) has been developed by the GO Consortium to attempt to consistently describe gene products in different databases based on the knowledge that many genes functioning in the core biological processes are shared by all eukaryotes (The Gene Ontology Consortium 2000). GO consists of three independent structured controlled vocabularies (ontologies) that describe gene products according to their biological process, molecular function, and cellular components. The building blocks of GO are GO terms that are made of a unique ID with the form GO:nnnnnnn and a term name, e.g. GO:0000166 nucleotide binding. Go terms are constructed as nodes of a network and a child term can be related to several parent terms.

Blast2GO (B2G) is a tool designed to electronically assign GO terms to genomic sequences of non-model species based on similarity searches with statistic analysis (Conesa et al. 2005). Briefly, Blast is used in B2G to find homologs to query sequences. Mapping is then performed to obtain GO terms associated with the Blast hits. An annotation rule is applied to assign GO terms to the query sequence. Statistical analysis can be performed after GO annotation is available. Moreover, B2G provides different statistical charts describing the results obtained from blasting, mapping or annotation.

The alignment of multiple nucleotide or amino acid sequences is used to detect homology and evolutionary relationship among sequences and thus it is a very useful tool in molecular biological research. Most of the automatic alignments use 'progressive' method that first aligns the most related sequences and then gradually adds in the less related ones devised by Feng and Doolittle (Feng et al. 1987). Clustal W is a program that improves the sensitivity of progressive multiple alignment approach without losing the speed and efficiency. The basic multiple alignment algorithm contains the following three major steps. Firstly, all sequences are aligned pairwisely to calculate a distance matrix that shows the divergence of each pair of sequences. Secondly, a guide tree is generated based on the distance matrix of step 1 using the Neighbor-Joining method (Saitou et al. 1987). Thirdly, the sequences are progressively aligned based on the branch order in the guide tree.

3. Results and Discussion

3.1 Characteristics of the *Medicago truncatula* genome

3.1.1 Repetitive sequences and transposable elements

When the *M. truncatula* genomic sequences were searched against the *A. thaliana* repeat database using Repeatmasker, about 11% of the genome was identified as repetitive sequence, which is comparable with that in *A. thaliana* (~10%) (Arabidopsis Genome Initiative 2000) and in *G. max* (~15%) but less than that in *O. sativa* (~35%) (International Rice Genome Sequencing Project 2005), *L. japonicus* (34.3%) (Sato et al. 2008), and in *P. trichocarpa* (42%) (Tuskan et al. 2006). A total of 16,865 di-, tri-, and tetra-nucleotide simple sequence repeats (SSRs) were identified in Medicago genome with a frequency of occurrence estimated to be 1 SSR per 15.1kb when we consider the total length of the 8 pseudomolecules of the medicago genome is 255Mb. Di-, tri-, and tetra-nucleotide SSRs accounted for 56.6%, 30.6%, and 12.8% of the identified SSRs. The repeat (TA)n, the most abundant SSR in plants (Gianfranceschi 1998), accounted for 36% of the total SSRs with (TA)n, (TTA)n, and (TAAA)n each representing 63.62% of di-, 18.47% of tri-, and 23.27% of tetra-nucleotide repeat units, respectively in Medicago. A total of 8433 transposable elements were identified, among which the total copy number of class I TE subfamily is higher than that in *A. thaliana*, and *P. trichocarpa* but much lower than that in *L. japonicus, G. max*, and *O. sativa* and the

total copy number of class II is similar to that in *A. thaliana*, and *P. trichocarpa* but less than that in *L. japonicus, G. max*, and *O. sativa* (Table 3.1 and Table 3.2). Since, no SINEs could be identified in class I while searching against the *A. thaliana* repeat database, the SINE sequences from all other plants were retrieved and compared with the Medicago genomic sequence. As a result, 305 SINES were found.

Table 3.1 Transposon abundance on the *Medicago truncatula* chromosomes

Transposons	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
Class I								
LTR-copia	348	389	564	577	633	327	472	517
LTR-gypsy	171	274	449	301	348	299	322	308
LINEs	77	95	105	90	130	37	74	87
SINEs	36	55	39	40	44	15	45	31
Class II								
En_Spm	12	11	24	14	27	26	17	17
MuDR	34	37	50	53	45	16	40	49
Tcl-type	23	22	34	32	36	17	23	35
hobo-activator	19	10	25	37	37	21	20	33
Tourist/Harbing	8	9	15	17	18	12	11	11
unclassified	57	63	111	97	100	49	63	62

Table 3.2 The comparison of the transposon copy numbers in *M. truncatula* (Mt), *L. japonicus* (Lj), *G. max* (Gm), *P. trichocarpa* (Pt), *O. sativa* (Os) and *A. thaliana* (At)

	Mt	Lj	Gm	Pt	Os	At
Class I	7299	75343	~58200	~5000	61900	2109
Class II	1134	11786	~7000	~1000	163800	1385

3.1.2 Genes encoding non-coding, stable RNAs

3.1.2.1 tRNA genes:

The program tRNAscan_SE 1.21 identified 632 putative transfer tRNA genes in *Medicago truncatula* genome including 523 genes that decode 20 standard amino acids (as shown in Table 3.3), 100 pseudogenes, 1 selenocysteine tRNA gene, 4 possible suppressor tRNAs, and 4 unknown isotypes. For those tRNA genes decoding the 20 standard amino acids, the most abundant was $tRNA^{Leu}$ with 39 copies, while the least abundant was $tRNA^{Cys}$ with only 9 copies. Chromosome 5 contains the most tRNA genes, with 91 encoded, while both chromosome 6 and 7 encode the fewest, 48 tRNA genes each. Although some medicago tRNA genes are clustered, the majority of the genes are dispersed individually throughout the genome. The clustered tRNA genes contain 2 to 7 of the same or different genes, with the largest cluster encoding 7 $tRNA^{Ala}$ genes within a 27 kb region on chromosome 6, which most likely arose from recent tandem duplication since they are highly conserved with over 90% identity.

The total number of tRNA genes in *A. thaliana)* (629) and *L. japonicus* (638) are very similar to that in M. truncatula while the number is higher in *P. trichocarpa* (817) and in *O. sativa* (763), and the highest in *G. max* (1295). The genomic organization of the tRNA genes in the above organisms is similar to that in the Medciago genome in which most of the tRNA genes are dispersed individually through the genome except for some clusters.

It is well established that the number of tRNA gene copies determine tRNA abundance (Itoh et al. 2007) as shown by a plot of the frequency of each amino acid obtained from the entire rice or Arabidopsis protein set against the number of corresponding tRNAs as well as observed in *L. japonicus, P. trichocarpa,* and *G. max* (Figure 3.1). The same tendency also was observed in *C. elegans* (Duret 2000), suggesting that the use of gene dosage to regulate tRNA gene expression levels likely appeared during the early stages of eukaryotic evolution (Itoh et al. 2007). In *Medicago truncatula,* this tendency also exists when the copy number of tRNA genes is plotted against amino acid frequency (shown in Fig 3.1).

As seen from Table 3.3, each tRNA gene has a preferred isotype. Does the preferred isotype correlate with the relative synonymous codon usage (RSCU) that represents the ratio of the observed frequency of a codon over the frequency expected (Duret 2000). In medicago, we also can observe a similar, although scattered, correlation between the number of isoacceptors and RSCU as reported for rice and Arabidopsis (Itoh et al. 2007).

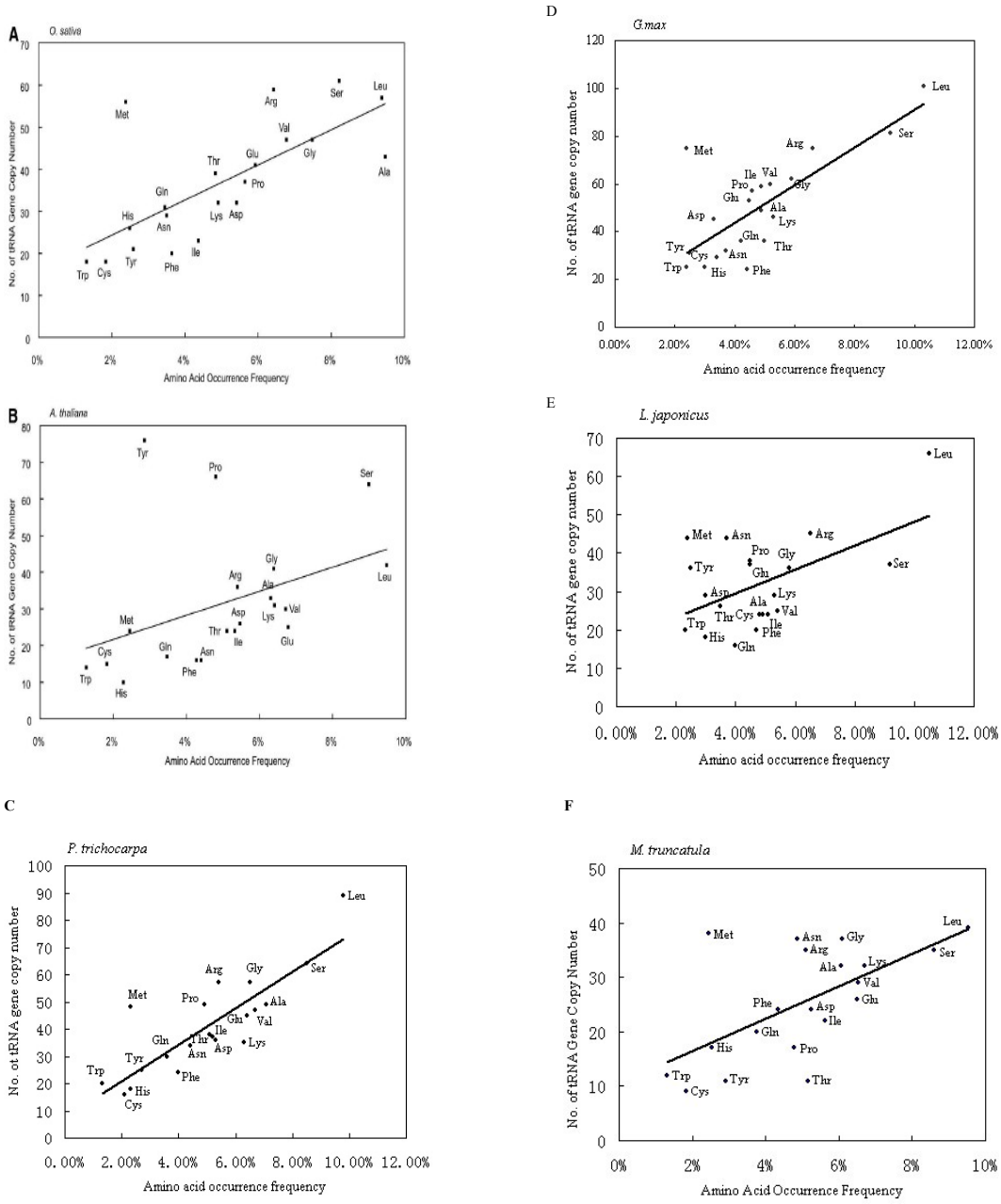


Figure 3.1 Correlation between the number of tRNA gene copies and occurrence frequency of amino acids in (A) *O. sativa*, (B) *A. thaliana*, (C) *P. trichocarpa*, (D) *G. max*, (E) *L. japonicus*, and (F) *M. truncatula* (A and B were taken from Itoh et al, 2007).

Table 3.3. Medicago truncatula isoacceptor tRNA gene copy number and the relative synonymous codon usage (RSCU)

Amino aci	Codon	Gene numbe	RSCU		Amino acid	Codon	Gene numbe	RSCU
Gly	GGU	2	**1.48**		Arg	AGA	9	**2.16**
	GGC	**18**	0.52			AGG	6	1.45
	GGA	14	1.39			CGU	**13**	0.9
	GGG	3	0.61			CGC	0	0.45
Val	GUU	**13**	**1.78**			CGA	4	0.62
	GUC	4	0.55			CGG	3	0.42
	GUA	5	0.67		Leu	CUU	8	1.56
	GUG	7	0.25			CUC	0	0.66
Lys	AAA	**16**	**1.04**			CUA	9	0.63
	AAG	16	0.96			CUG	3	0.57
Asn	AAU	8	**1.29**			UUA	6	0.96
	AAC	**29**	0.71			UUG	**13**	**1.62**
Gln	CAA	**15**	**1.27**		Ser	AGU	0	1.14
	CAG	5	0.73			AGC	**15**	0.63
His	CAU	0	**1.36**			UCU	9	**1.64**
	CAC	**17**	0.64			UCC	1	0.71
Glu	GAA	**16**	**1.17**			UCA	9	1.5
	GAG	10	0.83			UCG	1	0.36
Asp	GAU	0	**1.46**		Thr	ACU	**9**	**1.47**
	GAC	**24**	0.54			ACC	4	0.78
Tyr	UAU	0	**1.3**			ACA	8	1.43
	UAC	**11**	0.7			ACG	1	0.32
Cys	UGU	0	**1.26**		Pro	CCU	6	**1.61**
	UGC	**9**	0.74			CCC	0	0.49
Phe	UUU	0	**1.25**			CCA	**10**	1.48
	UUC	**24**	0.74			CCG	1	0.42
Ile	AUU	**17**	**1.5**		Ala	GCU	**16**	**1.77**
	AUC	0	0.72			GCC	0	0.58
	AUA	5	0.78			GCA	11	1.34
Met	AUG	38				GCG	5	0.31
Trp	UGG	8						

3.1.2.2 microRNA genes:

To identify the microRNA (miRNA) precursor candidates in Medicago genome, the 8 *Medicago truncatula* pseudochromosomes were compared with the *Arabidopsis thaliana* mature miRNA sequences (203 miRNA) downloaded from the miRNA registry database (http://microrna.sanger.ac.uk/sequences/) using crossmatch. As a result, 45 precursor candidates encoding 17 miRNA species were found with miR399 being the most abundant Medicago miRNA family with 5 members and 12 copies (shown in Table 3.4) and their putative target genes belong to ubiquitin-conjugating enzyme family. This number is comparable with that in *Lotus japonicus* where 53 miRNA were encoded on 6 Lotus chromosomes (Sato et al. 2008). The total number of miRNA genes in the *A. thaliana* (203), *O. sativa* (158), and *P. trichocarpa* (169) is much higher than that in *G. max* (120), *M. truncatula* (45) and *L. japonicus* (53) genomes, indicating that possibly the legume family has many legume-specific miRNAs that could not be identified here. Actually, 1312 miRNA legume-specific candidates were found in *L. japonica* (Sato et al. 2007). Recently, 8 novel microRNAs were identified from *M. truncatula* (Szittya et al. 2008) (Table 3.5). As seen from table3.6, 14 out of 17 miRNA gene families found in Medicago are conserved in *A. thaliana*, *O. sativa*, and *P. trichocarpa*, *G. max*, and *L. japonicus*, suggesting that a core set of miRNA gene families are conserved in angiosperms.

Plant microRNA Potential Target Finder (miRU) then was used to find the putative targets of the microRNAs by searching against the Dana Farber Cancer Institute (DFCI) Medicago gene index (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago), which previously was available from The Institute for Genomic Research (TIGR) as the Medicago Gene Index (MtGI). The putative targets and the putative proteins encoded by the target genes were shown in Table 3.4. It seems that the potential target genes of the counterparts of most of the conserved miRNAs in the six plants are homologous (Tuskan et al. 2006, Table S7, Sato et al. 2008, Table S6).

Table 3.4 List of miRNA precursor genes and the putative targets

Medicago Chromosome Number	Conserved Arabidopsis miRNA	Putative target	Annotation
1	Ath-miR167b	TC115313	Auxin response factor 8
1	Ath-miR166c	TC141583	Putative uncharacterized protein
1	Ath-miR408	TC116986	Basic blue copper protein
1	Ath-miR164b	TC128769	NAC domain protein NAC1
1	Ath-miR160b	TC135807	Auxin response factor 10
1	Ath-miR399c	TC115486	ubiquitin-conjugating enzyme family
2	Ath-miR399f	TC115486	ubiquitin-conjugating enzyme family
2	Ath-miR399d	TC115486	ubiquitin-conjugating enzyme family
2	Ath-miR169f	TC115415	Transcription factor
2	Ath-miR169b	TC115415	Transcription factor
2	Ath-miR169e	TC115415	Transcription factor
3	Ath-miR319b	TC135610	Putative uncharacterized protein
3	Ath-miR393a	TC115130	transport inhibitor response 1 (TIR1),
3	Ath-miR169j (2)*	TC117738	Transcription factor
4	Ath-miR390b	TC118766	Protein kinase
4	Ath-miR399f (2)*	TC115486	ubiquitin-conjugating enzyme family
4	Ath-miR171a	TC120850	Scarecrow-like protein
4	Ath-miR397a	TC112793	Putative uncharacterized protein
4	Ath-miR167d	TC115313	Auxin response factor 8
4	Ath-miR399a (5)*	TC115486	ubiquitin-conjugating enzyme family
4	Ath-miR166a (2)*	TC141583	Putative uncharacterized protein
4	Ath-miR169j	TC117738	Transcription factor
5	Ath-miR319a (2)*	TC135610	Putative uncharacterized protein
5	Ath-miR319c	TC135610	Putative uncharacterized protein
5	Ath-miR393b	TC115130	transport inhibitor response 1 (TIR1),
6	Ath-miR157b	TC131038	
6	Ath-miR399c	TC115486	ubiquitin-conjugating enzyme family
6	Ath-miR399b	TC115486	ubiquitin-conjugating enzyme family
7	Ath-miR164a (2)*	TC128769	NAC domain protein NAC1
8	ath-miR166c	TC141583	Putative uncharacterized protein
8	Ath-miR162b	TC123701	
8	Ath-miR171a	TC114268	Scarecrow-like protein
8	Ath-miR396b	TC116910	Cysteine protease precursor
8	Ath-miR396a	TC116910	Cysteine protease precursor
8	Ath-miR829.1	TC128533	Helix-loop-helix DNA-binding
8	Ath-miR171c	TC120850	Scarecrow-like protein

* The number in bracket indicated the copy number of the miRNA in one chromosome

Table 3.5 The novel microRNAs in *M.truncatula*

Novel microRNAs	Putative targets	Annotation
miR2086	TC125570	Ubiquitin carrier protein
miR2087	No target found	N/A
miR2088	TC126233	Peptidyl-prolyl cis-trans isomerase
miR2089	TC112724	TIR; Disease resistance protein
miR1507	TC128248	Disease resistance protein
miR1509	TC131818	Beta-glucan-binding protein
miR1510a	No target found	N/A
miR1510b	No target found	N/A

*Taken from (Szittya et al. 2008)

Table 3.6 Number of conserved microRNA families in 6 plant genomes

miRNA family	Arabidopsis	Rice	Poplar	Soybean	Lotus	Medicago
miR159/319	6	8	5	11	4	4
miR160	3	6	8	2	2	1
miR162	2	2	3	0	0	1
miR164	3	5	6	8	5	3
miR166	9	12	17	5	5	4
miR167	4	9	8	3	1	1
miR169	14	17	32	12	5	6
miR171	4	7	10	11	6	3
miR390	2	1	4	2	2	1
miR393	2	2	4	5	2	2
miR396	2	3	7	4	3	2
miR397	2	2	3	2	1	1
miR399	6	11	12	3	6	12
miR408	1	1	1	3	2	1

* The data for the first three plants are from (Tuskan et al, 2006, table S8), the data for lotus from (Sato et al, 2008, table S6)

3.1.2.3 rRNA genes:

Typically plant ribosomes contain four rRNAs, a 5S, 5.8S, 18S and 26S. The corresponding genes, 5S rRNA genes (5S rDNA) and 18-5.8-26 rDNA units tend to

cluster on the telomeric or centromeric regions. In medicago, a cluster of 5S rDNA is located near the pericentromeric region on chromosome 4 while another 5S rDNA cluster is located near the telomeric region on chromosome 5. A locus containing 10 copies 18S-5.8S-26S rDNA operons, 13 copies of lone 26S rDNA, and 12 copies of lone 18S rDNA spans 0.3 Mbp on chromosome 5 in the pericentromeric region. The similar rDNA distribution also was found in other plant genomes. In *O. sativa*, one 17S-5.8S-25S locus was found at the telomeric end of the short arm of chromosome 9 and a second 17S-5.8S-25S rDNA locus at the end of the short arm of chromosome 10 while a single 5S cluster is present on chromosome 11 in the vicinity of the centromere (International Rice Genome Sequencing Project 2005). In *A. thaliana*, one 5S cluster was found in the telomere of chromosome 3 and the other was in the telomere of chromosome 1 while one complete 18S-5.8S-26S rDNA was found in the centromere of chromosome 3 and the other the centromeric region of chromosome 1(Salanoubat et al. 2000, Theologis et al. 2000). In addition, two megabase-sized rDNA gene clusters are located at the tip of the short arms of chromosome 2 and 4, respectively (Mayer et al. 1999). In *P. trichocarpa*, FISH indicated one 5S repeat cluster on LGXVII and two 18-5.8-26 unit clusters, one of which is located on the telomere of LGXIV, other of which remains unlocated (Tuskan et al. 2006). Two copies of 18-5.8-26 unit and two or more copies of 5S rDNA are located together in *L. japonicus,* genome (Sato et al. 2008). In *G. max,*, three 5s rDNA clusters were identified: the first cluster containing 88 copies is located between 12.53 Mb and 12.57Mb on super contig 1 (near the centromere); the

second one containing 60 members is located in the beginning of the super contig 519; the third one containing 27 copies is clustered in the beginning of the super contig 872. As to the unit 18S-5.8S-26S in soybean, a large cluster containing 50 units was found between 11.41 Mb and 14.15 Mb on super contig 6 (near the centromere) and 46 small clusters including at least one unit were found in the beginning of their corresponding super contigs.

3.1.3. Characterization of the protein-coding genes

In total, 50,540 protein-encoding genes were identified within the latest updated assembly of the M. truncatula genome, with 23,175 genes (46%) having homology with sequences in the medicago EST database. The statistics for the 8 medicago chromosomes is shown below in Table 3.7. The percentage of GC content on each chromosome (~33%) is very similar to each other. Chromosome 5 has the highest gene number (8,802) and gene density (1 gene / 4.5 kb) while the chromosome 1 has the lowest gene density (1 gene / 5.5 kb). Chromosome 5 also encodes the highest number of short (<99 nucleotide) genes that may account for its high gene number and density and lowest average gene length (1,934 bp). As to exons, chromosome 1 has the highest average exon number per gene (3.64/gene) and the shortest average exon size (~240 bp/ exon) while chromosome 6 has the lowest exon number (2.98/gene) and the longest exon (~309 bp/exon). In contrast to the conservation of the average exon size, the average intron size varies on different chromosomes.

Table 3.7 *Medicago truncatula* genome summary statistics

Feature	Chr1	Chr2	Chr3	Chr4
Length	29,095,779	27,211,856	39,719,832	35,871,116
GC content (%)				
Overall	33.33	33.19	33.43	33.32
Coding	41.19	41.29	41.25	41.16
Noncoding	31.19	30.92	31.16	30.95
Number of genes	5316	5350	7881	7178
Gene density				
(kb/gene)	5.5	5	5	5
Average gene length	2149.57	2026.22	2236.87	2127.8
Average aa length	290.44	299.89	313.97	293.01
Exons				
Number of exons	19,325	18,564	27,002	24,918
Total length	4,647,819	4,829,217	7,446,900	6,331,251
Average per gene	3.64	3.47	3.43	3.47
Average size	240.51	260.14	275.79	254.08
Introns				
Number on introns	14,010	13,215	19,122	17,741
Total length	6,793,310	6,024,274	10,200,998	8,959,837
Average size	484.9	455.9	533.5	505
Number (%) expressed ge	2446 (46%)	2448 (46%)	3698 (47%)	3307 (46%)
	Chr5	**Chr6**	**Chr7**	**Chr8**
Length	39,917,350	19,087,130	30,489,780	33,703,334
GC content (%)				
Overall	32.9	33.64	33.17	33.15
Coding	40.93	40.9	41.08	41.18
Noncoding	30.77	31.33	31.01	30.92
Number of genes	8802	3621	6082	6310
Gene density				
(kb/gene)	4.5	5.3	5	5.3
Average gene length	1934.38	2035.23	2117.12	2216.29
Average aa length	296.75	305.55	295.08	295.42
Exons				
Number of exons	29,726	10,781	20,668	22,127
Total length	7,862,379	3,330,015	5,402,286	5,611,299
Average per gene	3.38	2.98	3.4	3.51
Average size	264.5	308.88	261.38	253.6
Introns				
Number on introns	20,926	7,161	14,587	15,818
Total length	9,184,958	4,046,704	7,488,639	8,389,338
Average size	438.9	565.1	513.4	530.4
Number (%) expressed ge	3998 (45%)	1640 (45%)	2731 (45%)	2907 (46%)

The *M. truncatula* genome was compared with the five other sequenced plant genomes, i.e. *A. thaliana* (Arabidopsis Genome Initiative 2000), *O. sativa* (International Rice Genome Sequencing Project 2005), *P. trichocarpa* (Tuskan et al. 2006), *L. japonicus* (Sato et al. 2008), and *G. max* and the results are shown in Figure 3.2.

The gene number in Medicago is less than that in soybean but more than that in the remaining four plant genomes (Figure 3.2A). The gene density of the Medicago is similar to that in *A. thaliana* and it is almost twice as much as that in *O. sativa*, *P. tricorcarpa,* and *L. japonicus* and three times as much as that in *G. max* (Figure 3.2B). The average gene length of the Medicago genome is the second shortest while the lotus has the longest gene size (Figure 3.2C). The average exon sizes of the six organisms have no much difference (range in 247-282 bp) while the average intron sizes have distinct difference (range in 148-498 bp) (Figure 3.2D&E). The *M. truncatula* genome has the lowest GC content (29.5%) while the *O. sativa* has the highest (43.6%) (Fig. 3.2F).The *M. truncatula* genome has the highest single-exon genes (around 40% and 55% of them are expressed) that account for its shortest protein length (299 aa) (Figure 3.2G&H). All the predicted proteins in medicago were compared with the proteins in the other five plant genomes and the results showed that 45.5%, 41.2%, 39.2%, 38.5%, and 36.6% of the medicago proteins are homologous to the proteins in *G. max, L. japonicus, A. thaliana*, *P. trichocarpa,* and *O. sativa*, respectively.

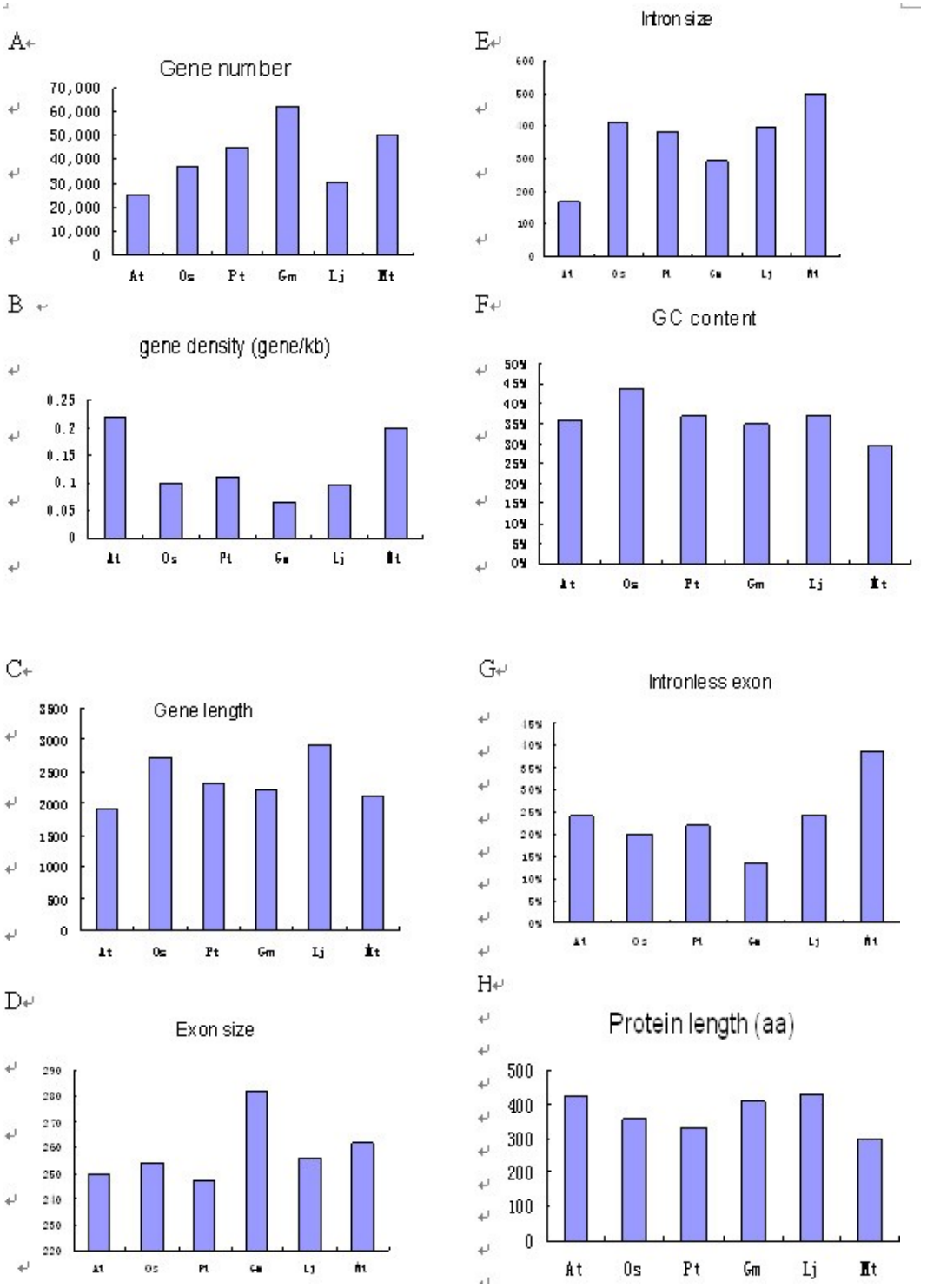
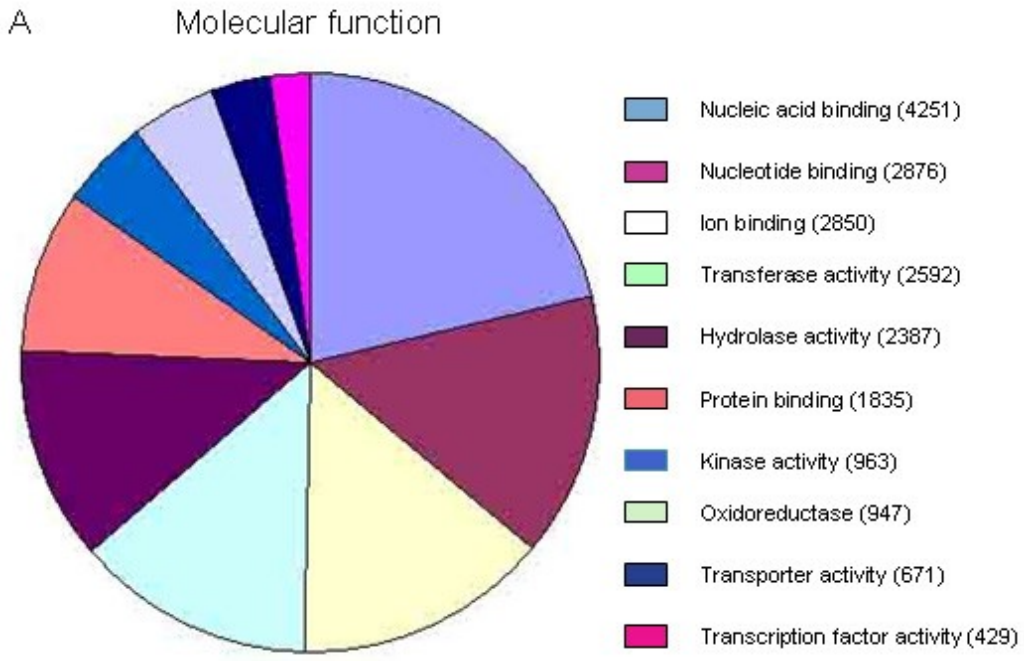


Figure 3.2 The comparison of the gene features in *A. thaliana* (At), *O. sativa* (Os), *P. trichocarpa* (Pt), *G. max* (Gm), *L. japonicus* (Lj), and *M. truncatula* (Mt)

It was possible to classify the predicted protein-encoding genes into functional groups according to GO (The Gene Ontology Consortium 2000) based on their InterPro domains. As seen in Figure 3.3, the nucleic acid binding domain is the largest domain in *M. truncatula* (4251) as well as in *A. thaliana* (3848) and *O.sativa* (10065) while the nucleotide binding domain is the largest in *L. japonicus* (2601), *G. max* (6573), and *P. trichocarpa* (5065). As to the category of biological process, the largest group is the domains functioning in protein metabolic process in *M. truncatula* and this is also the case in the other five plants. The second largest group is the domain functioning in DNA metabolism in *M. truncatula* as well as in rice while the transport domain is the second largest domain in the remaining four organisms.



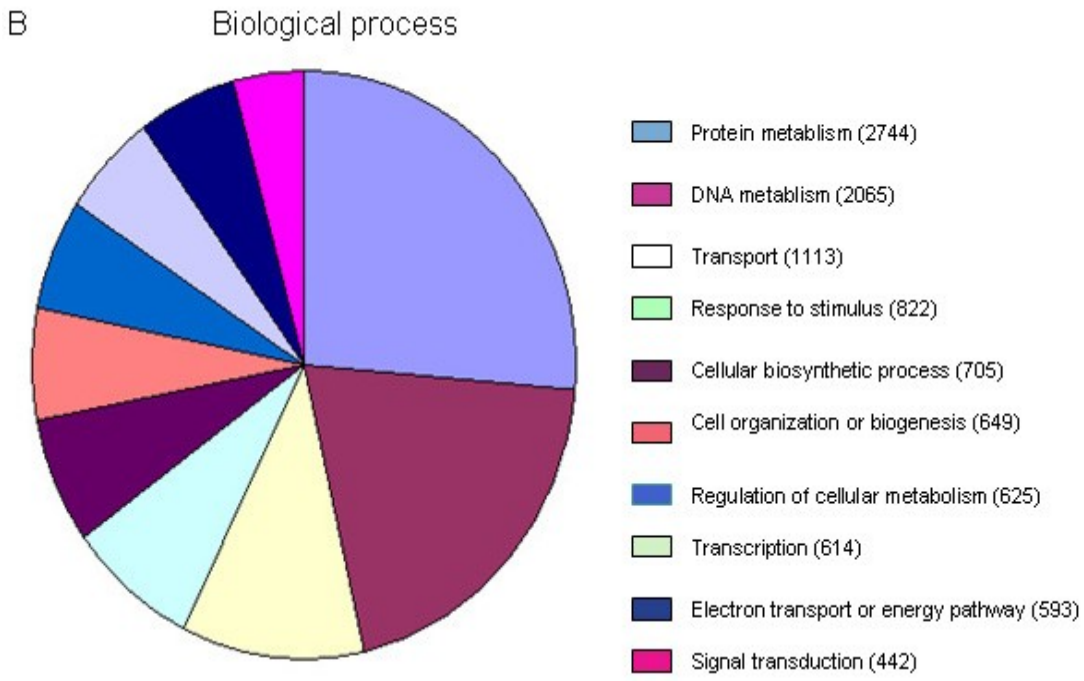


Figure 3.3 Gene Ontology (GO) category classifications. The results are shown for 10 representative classes of 'Molecular function' (A) and 'Biological process' (B). The predicted protein-encoding genes were automatically annotated by searching InterPro domains using program Blast2GO.

The comparison of the functional classifications among the six plant genomes based on GO annotation (Figure 3.4) showed that the Medicago genome contains comparable percentage of the main domains to that in other five plant genomes. From the 40 overrepresenting interpro domains in *M. truncatula* listed in Table 3.8 along with the corresponding domains in the other five genomes, we can see that most of the overrepresented domains in *M.truncatula* also were found overrepresented in the five other sequenced genomes with only a few exceptions. The most interesting domain among the exceptions is the late nodulin domain (IPR006810). This special domain was

only found in proteins of the galegoid group of legumes such as *M.truncatula*, *M. sativa*, and *Vicia faba*. *G. max* and *L. japonicus* don't belong to this clade of legumes. The late nodulin protein family is composed of several plant specific late nodulin sequences that are similar to the ENOD3 protein in *Pisium sativum* which is homologous to the nodule-specific cysteine-rich (NCR) protein is expressed in the late stages of root nodule formation and contains a signal peptide at the N-terminus and a cysteine-rich mature peptide at the C-terminus (Scheres et al. 1990).

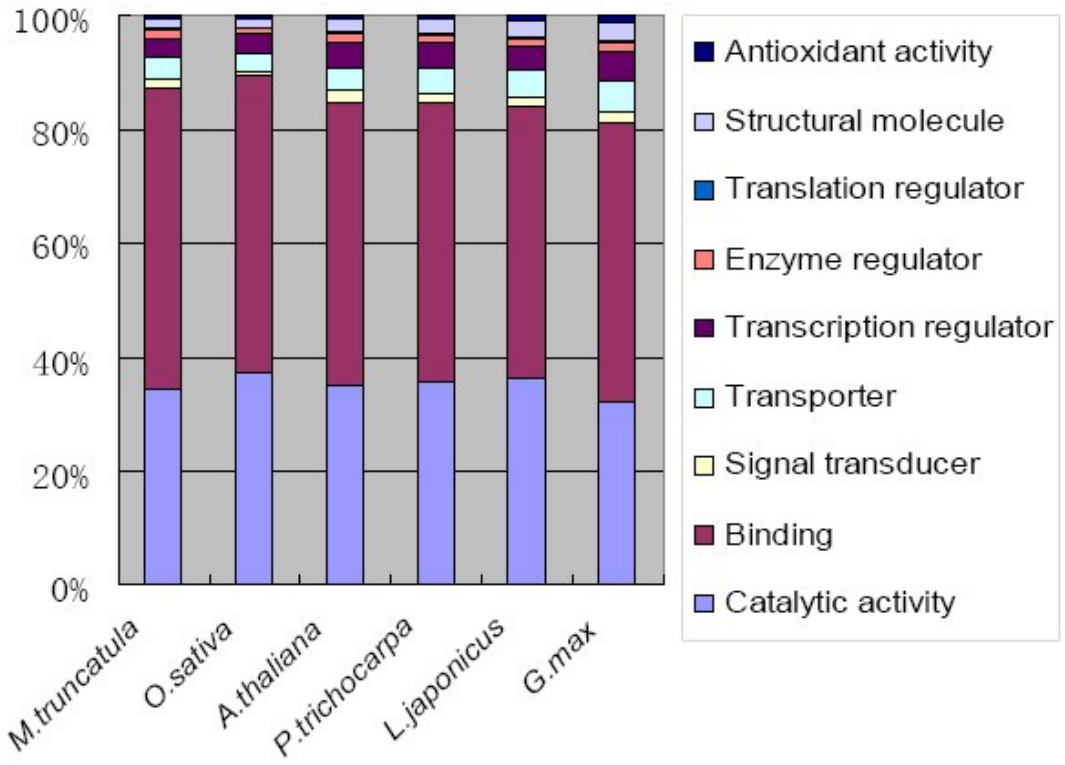


Figure 3.4 Percentage of functional domains in the six plant genomes based on InterProScan and Gene Ontology

Table 3.8 The comparison of the overrepresenting interpro domains in six plant genome

IPRID	Name	Mt	Lj	Gm	Pt	Os	At
IPR001878	Zinc finger, CCHC-type	1215	171	336	260	3280	514
IPR017441	Protein kinase ATP binding, conserved site	861	724	2229	1996	975	697
IPR001611	Leucine-rich repeat	837	400	1289	1286	1285	554
IPR008271	Serine/threonine protein kinase, active site	766	460	1552	1334	624	435
IPR001810	Cyclin-like F-box	727	397	279	357	928	903
IPR001969	Peptidase aspartic, active site	671	41	187	71	638	80
IPR002182	NB-ARC	482	122	373	556	594	160
IPR000767	Disease resistance protein	439	100	331	0	510	164
IPR013541	Protein of unknown function DUF1723	392	0	400	20	586	14
IPR002885	Pentatricopeptide repeat	366	433	921	633	626	481
IPR001128	Cytochrome P450	352	499	898	476	936	508
IPR001841	Zinc finger, RING-type	320	304	808	678	900	609
IPR013083	Zinc finger, RING/FYVE/PHD-type	313	313	799	692	846	518
IPR017451	F-box associated type 1	302	96	25	84	117	298
IPR013210	Leucine-rich repeat, N-terminal	255	182	487	522	742	304
IPR000209	Peptidase S8 and S53	248	109	224	143	317	113
IPR016040	NAD(P)-binding	236	329	801	619	715	121
IPR000157	Toll-Interleukin receptor	224	124	205	196	4	143
IPR012337	Polynucleotidyl transferase	217	120	347	154	4188	740
IPR012677	Nucleotide-binding, alpha-beta plait	181	237	545	370	568	256
IPR013242	Retroviral aspartyl protease	180	0	48	3	1110	45
IPR001650	DNA/RNA helicase, C-terminal	180	77	431	325	444	267
IPR000504	RNA recognition motif, RNP-1	180	218	540	363	556	267
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	172	126	303	231	272	116
IPR010285	Protein of unknown function DUF889	170	3	316	5	127	0
IPR005135	Endonuclease/exonuclease/phosphatase	155	29	192	92	256	182
IPR005225	Small GTP-binding protein	150	118	514	340	350	253
IPR003137	Protease-associated PA	147	84	271	250	179	118
IPR002198	Short-chain dehydrogenase/reductase SDR	147	136	379	181	350	196
IPR004330	FAR1	146	14	111	66	277	34
IPR008906	HAT dimerisation	140	28	29	84	270	76
IPR005123	2OG-Fe(II) oxygenase	135	128	290	189	186	139
IPR011992	EF-Hand	132	114	320	229	283	186
IPR001806	Ras GTPase	131	155	413	128	262	187
IPR000637	HMG-I and HMG-Y, DNA-binding	125	35	107	53	132	88
IPR013101	Leucine-rich repeat 2	120	88	41	39	125	174
IPR009007	Peptidase aspartic, catalytic	119	78	262	111	1594	246
IPR009810	Late nodulin	117	0	0	0	0	0
IPR004332	Transposase, MuDR, plant	112	0	21	22	812	332
IPR006670	Cyclin	108	54	169	120	160	101
IPR003676	Auxin responsive SAUR protein	106	55	170	125	73	84

3.2 In silico identification of nodule-specific Tentative Consensus sequences (TCs)

Expressed Sequence Tags (ESTs) are segments from either 5' or 3' end of a cDNA clone, usually 500-800 bp long, that are wildly used to identify expressed genes. Tentative Consensus sequences (TCs) are constructed by assembling ESTs into non-redundant transcripts based on the standard "overlap for at least 40 bases with at least 95% sequence identity" (Quackenbush et al. 2001). Since TCs represent more of the original mRNA, they are more useful than ESTs and they often represent a complete transcript. Moreover, the relative abundance of ESTs in a TC in the different libraries indicates the expression pattern of the gene and often is called an "electronic northern". The Dana Farber Cancer Institute database for medicago TCs (previously the TIGR MtGI) was constructed based on data from international *Medicago truncatula* EST sequencing and gene research projects (Quackenbush et al. 2001, Mergaert et al. 2003). When it was utilized to identify nodule specific genes, 340 TCs were identified that were expressed solely in root nodules (Federova et al. 2002). At that time, MtGI release 4.0 only contained 140,000 ESTs from 30 cDNA libraries. However, MtGI release 9.0 now contains 259,642 ESTs entries from 74 cDNA libraries and this new data could reveal new aspects about nodule-specific TCs, such as additional nodule-specific TCs, or TCs no longer nodule-specific. In addition, with the availability of about 75% *Medicago truncatula* genomic sequence, new insights about nodule-specific genes may

be revealed by study nodule-specific genes including their organization in *Medicago truncatula* genome, their evolution, and their regulatory regions.

Among the 74 cDNA libraries, 7 libraries now have been generated from nodules at different development stages (Table3.9). MtSN4, MtBB, R108Mt, and NOLLY were prepared from emerging or young nodules, and additional cDNAs were obtained from a nodulated root library, from effectively nitrogen-fixing nodules (GVN) and senescent nodules (GVSN). It should be noted that the MtSN4, MtBB and nodulated root libraries likely also contain plant genes expressed in root tissue because they were prepared from the mixture of nodules and adjacent roots.

Table3.9. The nodule libraries of *Medicago truncatula*

Names of nodule libraries	Total No. of ESTs	Total No. of TC	Description of the libraries	Library source
MtSN4	847	208	Nodules 4 days and 10 days after *Sinorhizobium meliloti* inoculation (pooled)	Centre National de la Recherche Scientifique (CNRS) (France)
MtBB	7785	2535	ESTs from emerging nodules and adjacent root segments of 21-day-old plant harvested 4 days after inoculation with *Sinorhizobium meliloti*	Genoscope and Centre National dela Recherche Scientifique institut National de la Recherche Agronomique (France)
R108Mt	447	309	ESTs from symbiotic, developing young nodule	Institut des Sciences Vegetales (France)
GVN	6446	2619	ESTs from one -month-old nitrogen-fixing root nodules	University of Minnesota
GVSN	2661	1491	ESTs from senescent nodules	University of Minnesota
Nodulated root	3185	2014	Mixture of roots and nodules	The Samuel Roberts Noble Foundation
NOLLY	3066	1696	ESTs from young nodules 4 to 8 days post infection with Sinorhizobium meliloti strain Sm41	Centre National de la Recherche Scientifique (CNRS) (France)
Total	24437	10872		

Table3.10. Nodule-specific TCs encoding known proteins

TC no.	Strong blast hit	Evalue	TC no.	Strong blast hit	Evalue
138811	Nodule-specific cysteine-rich peptide 10	E=10-35	118859	Carbonic anhydrase	E=10-80
126333	Nodule-specific cysteine-rich peptide 19	E=10-34	114678	a-type carbonic anhydrase (Lotus japonicus)	E=10-92
121233	Nodule-specific cysteine-rich peptide 19	E=10-20	118427	Calmodulin-like protein 6b	E=10-46
139055	Nodule-specific cysteine-rich peptide 53	E=10-32	134855	Calmodulin-like protein 1	E=10-98
124823	Nodule-specific cysteine-rich peptide 54	E=10-27	114830	Calmodulin-like protein 2	E=10-95
135570	Nodule-specific cysteine-rich peptide 68	E=10-24	114792	Calmodulin-like protein 4	E=10-61
136099	Nodule-specific cysteine-rich peptide 74	E=10-27	125585	Calmodulin-like protein 5	E=10-76
116268	Nodule-specific cysteine-rich peptide 76	E=10-31	114830	Calmodulin-like protein 3	E=10-98
118128	Nodule-specific cysteine-rich peptide 324	E=10-28	119502	Putative cysteine proteinase	0
130739	Nodule-specific cysteine-rich peptide 94	E=10-32	124698	Putative cysteine protease	E=10-106
126448	Nodule-specific cysteine-rich peptide 103	E=10-36		(Trifolium pratense)	
132307	Nodule-specific cysteine-rich peptide 111	E=10-16	117139	Cysteine proteinase(Lotus japonicus)	E=10-133
130698	Nodule-specific cysteine-rich peptide 144	E=10-25	113318	Putative cysteine proteinase	0
134562	Nodule-specific cysteine-rich peptide 147	E=10-31	134079	Lipoxygenase	E=10-149
128830	Nodule-specific cysteine-rich peptide 159	E=10-33	113617	Lectin-related polypeptide	E=10-64
131174	Nodule-specific cysteine-rich peptide 159	E=10-15		(Robinia pseudoacacia)legume	
131588	Nodule-specific cysteine-rich peptide 201	E=10-32	115344	LCB3-ROBPS putative bark agglutinin	E=10-19
136689	Nodule-specific cysteine-rich peptide 217	E=10-25		LECRPA3 precursor	
135979	Nodule-specific cysteine-rich peptide 265	E=10-33	124863	Aspartyl protease family protein	E=10-85
128856	Nodule-specific cysteine-rich peptide 301	E=10-20		(Arabidopsis)	
134771	Nodule-specific cysteine-rich peptide 310	E=10-16	118074	Albumin1(Phaseolus vulgaris)	E=10-32
120047	Leghemoglobin (Medicago sativa)	E=10-52	124996	sst1 protein(Lotus japonicus)	0
139434	Leghemoglobin 1	E=10-65	131886	Hypothetical protein (Vitis vinifera)	E=10-31
131798	Leghemoglobin 2	E=10-68	132384	Lupeol synthase (Lotus japonicus)	0
127422	Leghemoglobin (Medicago sativa)	E=10-74	118824	Thioredoxin M-type,chloroplast precursor	E=10-22
139029	Leghemoglobin (Medicago sativa)	E=10-67		(Brassica nupus)	
124562	Leghemoglobin (Medicago sativa)	E=10-75	120929	Putative non-LTR retroelement	E=10-126
131669	Leghemoglobin (Medicago sativa)	E=10-66		reverse transcriptase	
120047	Leghemoglobin 29 (Vacia faba)	E=10-58	112566	Nodule-specific IRE-like protein	0
119713	Leghemoglobin 29 (Vicia faba)	E=10-55	122801	Lectin-related polypeptide (Robinia pseudoacacia)	E=10-25
133101	leghemoglobin	E=10-14	124612	B12D-like protein (Phaseolus Valguris)	E=10-33
126384	Nodule-specific glycine-rich peptide 2D	E=10-32	116072	REMO_SOLTU Remorin (Solanum tuberosum)	E=10-40
127188	Nodule-specific glycine-rich peptide 3B	E=10-65	123589	Putative repetitive prolin-rich protein	E=10-22
119756	Nodule-specific glycine-rich protein 3A	E=10-93	118528	Zinc finger (C3HC4-type RING finger)	E=10-29
132746	MtN1	E=10-29		family protein	
113105	MtN6	0	136882	Unnamed protein product (Medicago sativa)	E=10-147
114236	MtN9	E=10-169	1134070	Unnamed protein product (Vitis vinifera)	E=10-25
138204	MtN11	E=10-34	139724	Unnamed protein product (Vitis vinifera)	E=10-66
134290	MtN15	E=10-42	130876	Unnamed protein product (Vitis vinifera)	E=10-12
131451	MtN16	E=10-33	124823	Unknown protein (Arabidopsis)	E=10-36
121205	MtN22	E=10-84	134887	ORF2 (Glycine max)	E=10-24
115962	MtN22	E=10-96	137023	ORF2(Glycine max)	E=10-39
115929	MtN22	E=10-107	126121	LATE BLOOMER (Pisum sativum)	E=10-17
114025	MtN24	E=10-115	117825	Hexose transporter(Solanum lycopersicum)	0
131314	MtN25	E=10-16	114041	Putative purine permease (Oriza sativa)	E=10-86
129897	MtN29	E=10-35	119349	Putative thioredoxin m2 [Pisum sativum]	E=10-21
113188	Enod8.1	0	116878	CAF1 family ribonuclease	E=10-14
127504	Early nodulin 12 precursor (N-12)	E=10-24	117384	Peroxidase precursor	E=10-92
114187	Early nodulin ENOD18 [Vicia faba]	E=10-74	133034	Embryo-specific 3	E=10-96
114239	ENOD20	E=10-73	127962	Low affinity sulphate transporter	E=10-60
113614	Nod25	E=10-139	121687	Homeodomain-like	E=10-33
130836	Late nodulin	E=10-29	121445	Basic blue protein (Medicago sativa)	E=10-61
138028	Carbonic anhydrase	E=10-130	137391	N8 protein	E=10-32

191 nodule-specific TCs were obtained from the in silico comparison of gene expression of all the medicago cDNA libraries using the "EST expression" web site under the DFCI Medicago gene index set to a likelihood statistic, R, of greater than 9, that indicates the expression variation was significant based on Stekel et al's study (Stekel et al. 2000). To further identify the function of these TCs, the program blastX was used to search these TCs against non- redundant databases (nr) in NCBI with the E value less than 10-10. One hundred TCs were found similar to known GenBank sequences (shown in table 3.10). Ninety one TCs have no homology in GenBank.

3.2.1 Characterization of nodule-specific TCs

3.2.1.1 Nodule-specific cysteine-rich peptide

Among 100 TCs with strong similarity to known proteins, 21 were identified either as nodule-specific cysteine-rich peptides (NCRs) or cystein cluster proteins (CCPs) (see table 3.10). Interestingly, 29 more TCs similar to the NCR gene family were identified by aligning the remaining 91TCs with no homology in GenBank using ClustalW2. In total, 50 NCR genes were identified in my study (shown in Fig3.4). Previously, 114 CCPs and more than 311 NCRs were identified by two research groups, respectively (Fedorova et al. 2002, Mergaert et al. 2003). The number of NCRs identified in my research was much less because only TCs with R greater than 9 in the *in silico* gene expression analysis were chosen and considered as nodule-specific since the study (Stekel et al. 2000) showed when R>9, the identified genes in the computational gene

expression analysis represent genuine variation, and are not false positive results. Forty six out of 50 NCRs have more than 4 ESTs, among which, TC129888 contains the highest number of ESTs (138ESTs). TCs must contain more than 4 ESTs to be considered as a true differential expression when analyzing expression data *in silico* (Audic et al. 1997). Fedorova and the associates (Fedorova et al. 2002) also found that TCs identified as nodule specific and consisting of six of more ESTs could usually be confirmed by physical measurements of transcript abundance on microarray or northern blots. They also mentioned that only 40 of the 114 TCs identified as CCP transcripts were composed of more than 5 ESTs.

The NCR genes were expressed at different developmental stages, 27 in young, mature and senescent libraries, 7 in mature and senescent libraries, 6 in young and mature libraries, and 1 in young and senescent libraries. Two TCs and 7 TCs were found solely expressed in young and mature libraries, respectively. Most of the NCR genes were mainly expressed at the mature stage.

The NCR protein family that contains a late nodulin domain also consists of short peptides (60-90 amino acids) with characteristic of a conserved signal peptide, and a conserved cysteine motif (Fodorova et al. 2002, Mergaert et al. 2003). Except for the conserved signal peptide and the conserved cysteine motif, the remainder of an NCR sequence shows extensive divergence. All NCR genes encode small mRNAs of about 400 to 700 nucleotides long and encode for polypeptides of 60 to 70 amino acids. The highly conserved N-terminal region was composed of 20-29 hydrophobic amino acids

and was predicted as a signal peptide by SignalP (Mergaert et al. 2003). In contrast to the conserved signal peptide, the remainder of the polypeptides was highly divergent except for conserved Cys with constant number of amino acids between them (shown in Fig3.5). C1 and C2 were spaced by 5 amino acids, while C3 and C4 were spaced by 4 amino acids. Moreover, hydrophobic residues, located one amino acid N terminal to C1, an Asp and a Pro adjacent to C2, a basic amino acid (Arg or Lys) preceding and a hydrophobic amino acid after C3, the second Asp between C2 and C3, one or several Pro between C2 and C3, and one Gly between C3 and C4 were relatively well conserved. The alignment of the deduced amino acid from TC126333 could not show its conserved characteristics since it was 40-amino-acid longer than the other deduced NCR amino acids, although they did have a conserved signal peptide and a conserved cysteine motif. The deduced amino acid from TC121233 was 100% identical to partial of TC126333 and it was lack of the conserved signal peptide and the conserved C1 and C2 while the signal peptide in 5 of the 50NCR was not present.

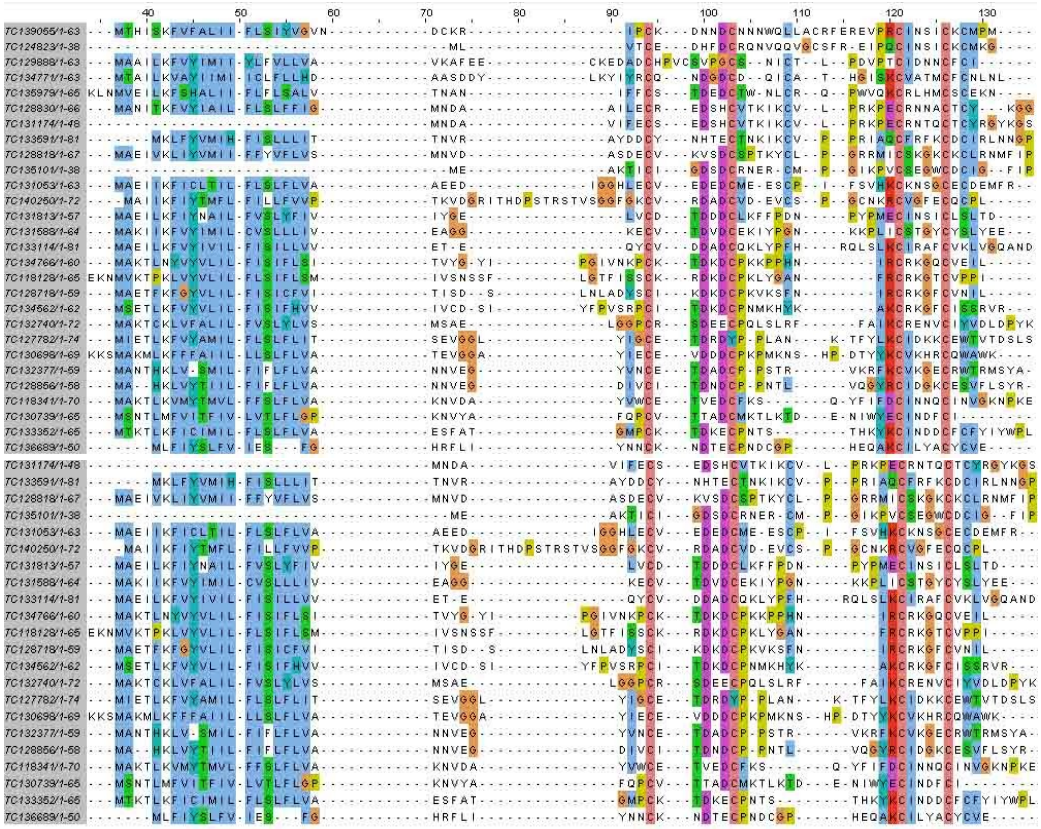


Figure 3.5 Jalview of ClustalW2 results for the deduced amino acids from 50 TCs similar to the NCR gene family. Conserved signal peptide and cysteines were marked.

3.2.1.1.1 The genomic organization of NCR genes

When the NCR genes were searched against each Medicago pseudochromosome using Chromosome Visualization Tool (CViT) to find the location of NCR genes, the results, as shown in Figure 3.6 reveal that although most of the NCR genes were clustered, others were dispersed throughout the genome. The genes in clusters were homologous to one or two NCR genes and likely evolved through tandem gene duplication, while genes dispersed on chromosomes most likely arose from segmental duplication.

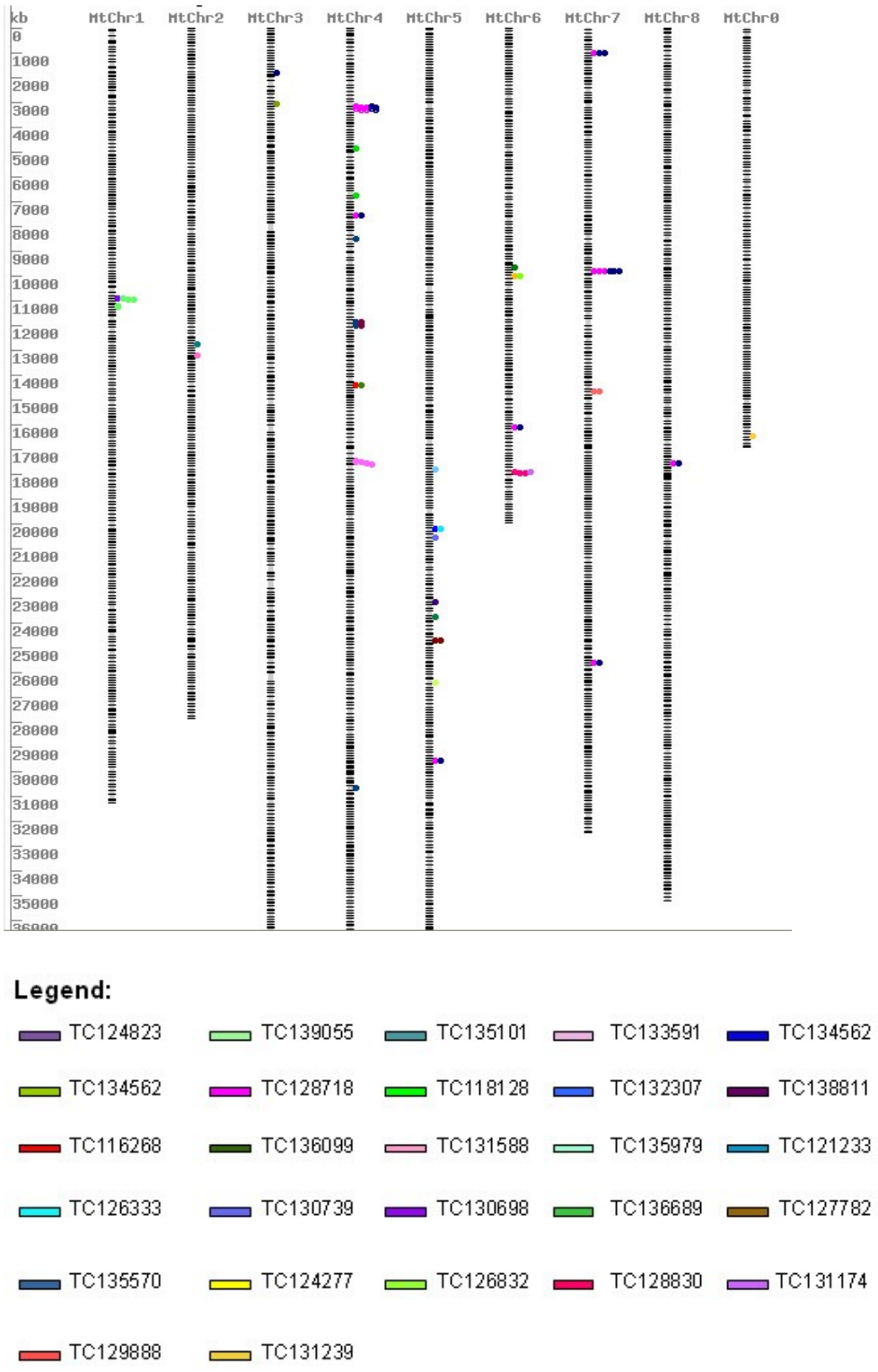


Figure 3.6 The positions of NCR genes on different chromosomes shown by CViT (http://www.medicago.org/genome/cvit_blast.php)

3.2.1.1.2. Evolution of NCR genes in *Medicago truncatula*

To examine the evolutionary pressures on NCR genes, the rates of non-synonymous (Ka) substitutions and synonymous substitutions (Ks) were determined for these genes. A synonymous substitution means that the substitution of one base for another in an exon of a protein-coding gene doesn't change the corresponding amino acid sequence while a non-synonymous substitution means a substitution in coding sequence change the amino acid sequence. A ratio of Ka/Ks greater than 1 indicates that a gene is under positive or diversifying selection that preserves the non-synonymous substitutions and thus may cause a protein has a new function or a protein binds a new substrate. When the ratio is less than 1, a gene is under purifying selection that preserves the synonymous substitutions and thus maintains the function of the gene product. When the ratio is 1, a gene is under neutral selection. A Ka/Ks tree based on the calculation of Ka and Ks is shown in Figure 3.7. From the tree, we can see that several NCR genes underwent purifying selection and others underwent positive selection as the Ka/Ks ratios of the purifying selection and positive selection ranged from 0.30 to 0.95 and from 1.1 to 3.4, respectively.

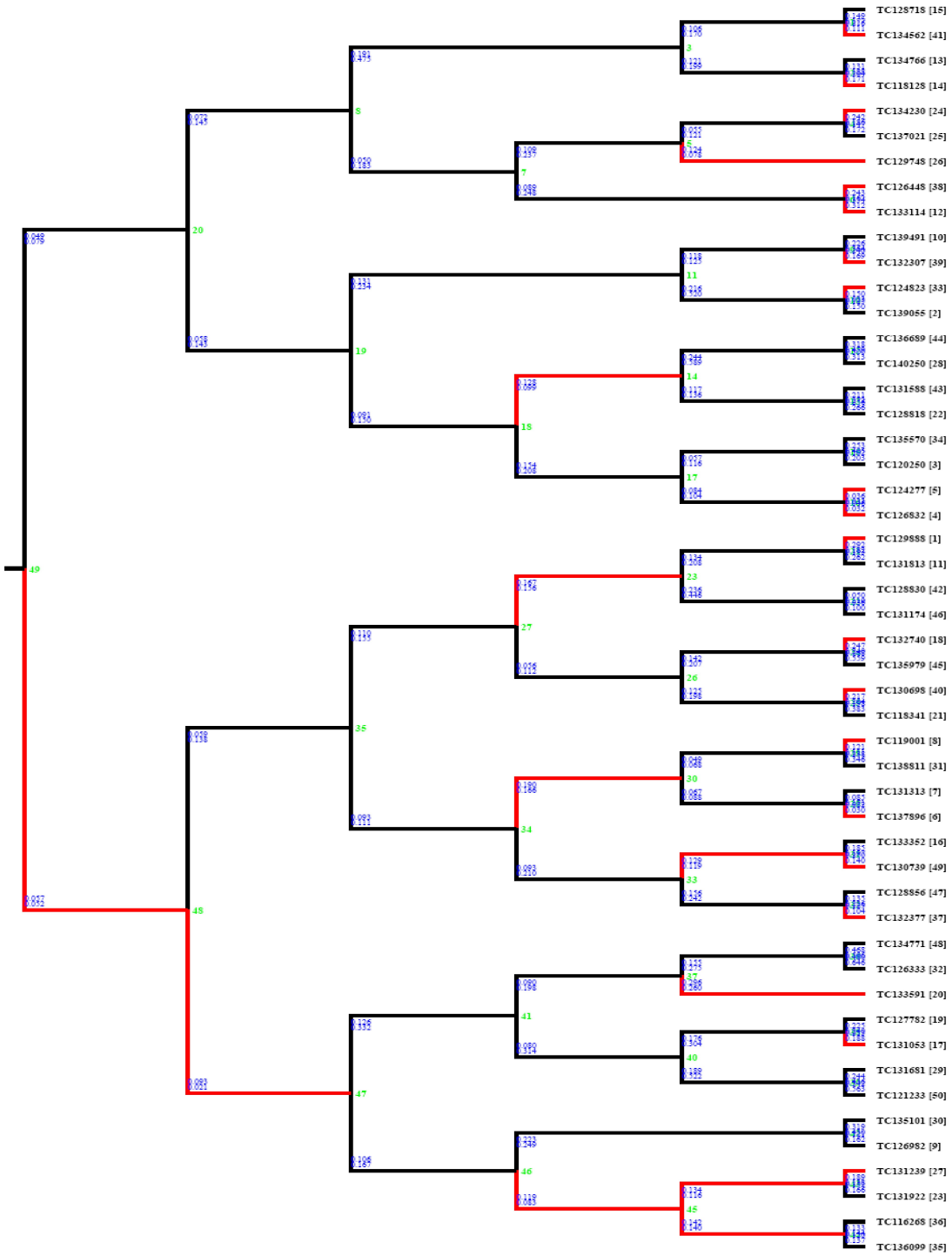


Figure 3.7 Ka/Ks tree indicating the positive (red) and purifying selection (black)

3.2.1.1.3 The gene features of NCR genes

All NCR TCs were searched against non-redundant nucleotide database in GenBank using BlastN to find the BACs containing these genes and the location of these genes on the BACs. Only those TCs with at least 99% identity to Medicago genomic sequence were chosen for the further analysis. By analyzing the blast result, NCR genes were found to contain only one intron and two exons. The names of TCs and the corresponding BAC clone, the position, the chromosome name and the size of gene and intron were listed in Table3.11. As can be seen in table 3.11, the size of most of NCR genes were in the range of 400 to 820 bp except for 3 genes that were somewhat longer (up to 1365 bp). The intron size was around 100 bp except for all but two longer introns. The 5' splicing donor site (GT) and 3' splicing acceptor site (AG) were found in 15 out of 18 introns. The remaining 3 introns without normal donor and acceptor sites may result from the TC assembly errors.

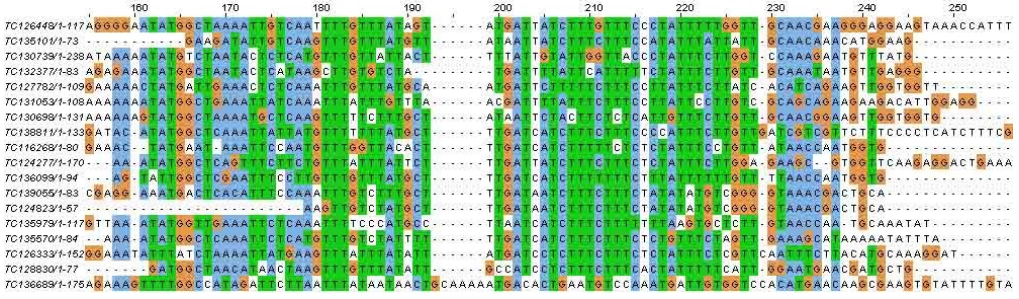
Since most of NCR genes were mainly expressed in the mature libraries, it is possible that all the gene expression were under the control of some common regulatory elements. To identify the common regulatory elements, 1000 bp upstream of the 18 NCR genes were extracted and compared with the previously published cis-acting regulatory elements in the database of plant cis-acting regulatory DNA elements (PLACE). The result was shown in Table 3.12. The position of the base just prior to the first base of NCR TC was designated as -1. To compare the relative position of the

common motifs, the position of putative TATA box in the promoter region also was listed, and both strands were searched for the common regulatory elements using PLACE.

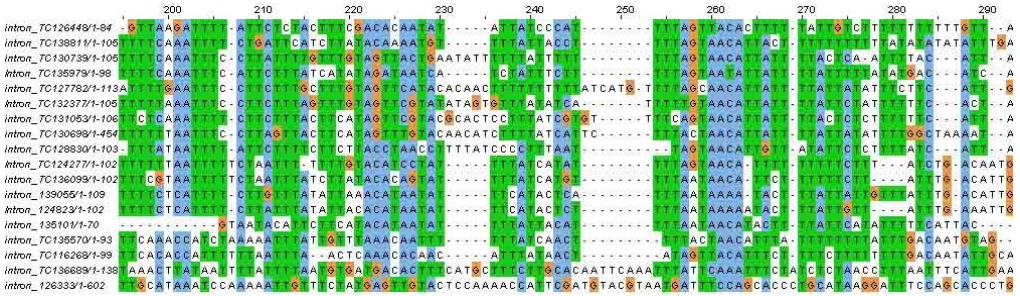
Table 3.11. The location and gene feature of NCR genes

TC name	Clone name	Position	Chr. No.	Gene size	Intron size
127782	CU302328	13437-12722	5	715	113
131053	CU459032	42728-42156	5	573	106
135979	CR932957	22813-23628	5	816	98
116268	AC152176	113512-114876	4	1365	99
130698	CR932039	118414-119417	5	1004	454
135101	AC161400	21723-21286	2	438	70
135570	AC146864	82310-81849	6	462	93
136099	AC152176	120032-119576	4	457	102
126333	CR962123	114844-113648	5	1197	602
124823	AC149493	70598-71336	1	739	102
139055	AC149493	83337-84009	1	673	109
132377	AC138527	99893-99347	3	547	105
124277	AC148657	21990-22736	5	747	102
136689	CR954192	74254-73647	5	608	138
138811	AC202469	12230-11773	4	458	105
128830	AC150703	92135-91518	6	618	102
126448	CT030243	5957-5144	3	814	83
130739	CT963114	128359-128988	5	630	104

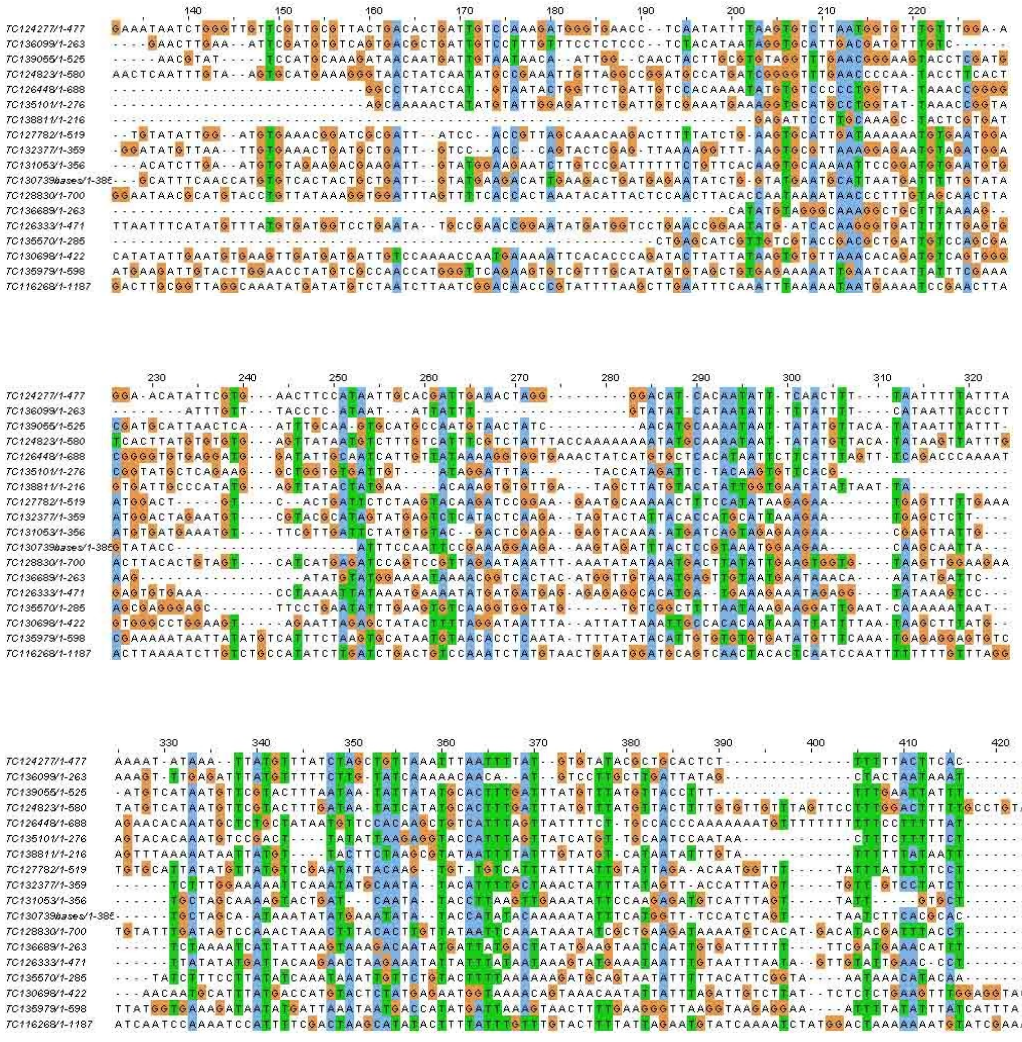
A



B



C



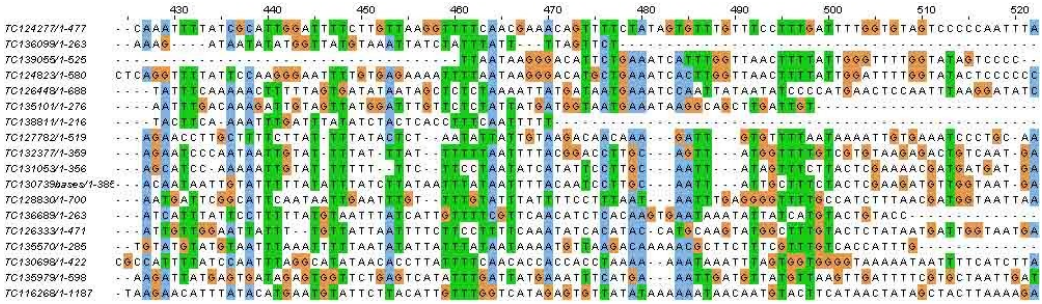


Figure 3.8 The Jalview showed the conservation among the first exons (A), introns (B) and the second exons (C)

Table 3.12 Common motifs found in NCR genes using PLACE

TC name	AAAGAT motif	CTCTT motif	TATA box	libraries
127782	-28	-130*/-86/-78*	-388*	Mature
131053	-26/-255*	-533*	-58*	Young
138811	-109*	-320/-26*	-68	Young
136099	-203/-248	-395*/-149*	-72*	Mature
130739		-239*/-13*	-63	Mature
126448	-469*	-311/-106*	-43	Mature
128830	-420*	-290/-401	-50	Mature
136689	-296*	-253*/-197*	-91*	Mature
124277	-211	-544*	-65	Mature
132377		-351*/-144*/-92*	-79*	Mature
139055	-411	-359	-119*	Mature
124823	-218/-88/-111*	-356*	-541	Mature
135570		-996	-309	Mature
135101	-378/-399	-740	-52	Mature
130698	-105	-74*/439*	-91*	Mature
116268	-184/-266/-186*	-5/-693	-31	Mature
135979	-144	-289	-94	Mature

* indicated the motif from the reverse strand; different locations for the same motif were separated by slash /.

Two short putative nodule-specific consensus sequence motifs, 5’-CTCCT and 5’-

AAAGAT, first described in soybean nodulin 20, 22, 23, and 44 genes (Sandal et al. 1987), were found present in the 5' upstream region of most of the NCR genes. The sequence AAAGAT was located at approximately -95 in these genes, -130 in the Leghemoglobin gene (Lb) genes (Sandal et al. 1987) and -193 in nodulin 24 (Verma et al. 1978). The CTCCT sequence was located about at position -130 in nodulin 20, 22, 23 and 44, while in the Lb genes it was at approximately positions -120 and -80 in the inverted form. In nodulin 24, this motif was observed at position -153 and -77 in the inverted form. The CTCCT motif located at -325 in soybean nodulin 23 promoter region was reported to be important for high-level organ specific expression (Stougaard et al. 1990). Mutation analysis also was confirmed that the CTCCT and AAAGAT sequences were positive specific elements in the N23 promoter (Jørgensen et al. 1991). It also was pointed out that these two motifs could not direct nodule-specific transcription without the presence of distal positive elements such as PE-AB (Jørgensen et al. 1991).

As seen in table 3.12, the reference sequence TATA box was found in the right position range in most of the sequences except for 3 (TC127782, TC124823 and TC135570) that may not contain this conserved box. The AAGAT sequence occurred in 15 out of 18 NCR gene regulatory regions and the CTCCT sequence occurred in every gene except TC135101 and TC135570, where it was located further away from -1 position. Therefore, we can conclude that these two motifs are highly conserved in NCR regulatory regions and they mostly are located at the proximal promoter regions or even in the promoter regions and speculate that they likely act as positive nodule-specific

element as they did for soybean nodulin genes.

Other nodule-specific motifs also may exist in the NCR regulatory region besides AAAGAT and CTCCT. Since only known cis-acting motifs can be found using PLACE, the WordSpy program was used to see if there is new nodule-specific motifs. The top five overrepresented motifs were TCCTT, TTGAA, TGTTG, TTTGTT and TTTCAT. Although not exprerimentaly confirmed, these motifs also likely play a role in nodule specific expression.

In eukaryotes, positive cis-acting elements or enhancer may locate not only in the upstream of a gene, but also in downstream of a gene and in the introns. The intron sequence of NCR genes are highly conserved (Figure 3.8), indicating that the introns in the NCR genes likely act as an enhancers and determine nodule-specific NCR gene expression in combination with the conserved upstream cis-acting motifs.

3.2.1.1.4 The possible role of NCR genes in nodulation

The identification of 50 NCR genes indicates the existence of a large nodule-specific gene family in *Medicago truncatula*. Although these genes have diverged, they still can be grouped because of the presence of the conserved signal peptide, their small size, the conserved Cys motifs, and their nodule specificity, observations indicating that NCR genes most likely are functional related. What is the biological role they might play in nodulation with such high sequence divergence? To address this question I investigated

several large multigene families in plants. For example, a large defensin-like gene family found in *Arabidopsis thaliana* is required to protect plants from wide-spectrum pathogens (Silverstein et al. 2005), plant receptor-like genes that are involved in different signaling processes (Shiu et al. 2001), resistance genes that recognize elicitors and protect plants from pathogen invasion (Bergelson et al. 2001), and pollen determinants for self-incompatibility (SCR) and pollen coat proteins (PCPs) (Schopfer et al. 1999, Vanoosthuyse et al. 2001). The underlying common property of these gene families is that they all are involved in recognition events, and thus it is possible that NCR genes also might be involved in recognition events as well (Mergaert et al. 2003) because NCR polypeptides are similar in structure to several known proteins, that include defensin and γ-thionin antimicrobial peptides (Broekaert et al. 1995, Zasloff 2002), SCR proteins (Schopfer et al. 1999), and scorpion neurotoxins (Bontems et al. 1991). The comparison between NCR TCs with the Medicago genomic sequence showed that the first exon corresponds approximately to the signal peptide and that the second exon corresponds to the mature NCR peptide, that is similar to the organization in the plant defensin, SCR, and in the scorpion toxin genes (Froy et al 1998, Vanoosthuyse et al. 2001). It is possible that these gene families arose from the same ancestor but they were highly diverged after duplication.

NCR genes are expressed in young, old, and mainly in mature nodule libraries. The abundance of NCR transcripts suggests large amount of encoded proteins are required in nodules, indicating there might be a gene dosage effect. The high expression level

shows that this gene family might play a very important role in nodule development. They also might act as antimicrobial defensins to avoid infections by other soil microorganisms during nodule formation or alternatively they act as signal molecules assuring communication between plant cells or between plant cells and rhizobial bacteria (Mergaert et al. 2003). Two nodule-specific NCRs were found to have antimicrobial activity against *P. syringae* and *Clavibacter michiganensis* but not against the growth of *S. meliloti* (Samac et al. 2007). That several of the NCR genes are under positive selection suggests that maintaining the nonsynonymous mutation may be needed to adapt to a fluid environment including different and changing microorganisms to more efficiently recognize and destroy these potential pathogens. As discussed above, the plant defensins are similar to NCR genes with a conserved signal peptide, a highly diverged mature peptide and a conserved cysteine motif. The study of Silverstein et al (Silverstein et al. 2005) showed that more than 300 defensin-like genes were found in Arabidopsis and their genomic organization and Ka /Ks pattern are similar to NCR genes. It is possible that the NCR genes originated from plant defensins but they highly diverged after duplication. To test this hypothesis, a phylogenetic tree was drawn using all the identified NCR genes (TC numbers), two near identical cysteine-rich defensin-like genes (Os1, Os2) in rice, the available cysteine-rich defensin-like genes in Arabidopsis from NCBI (gi numbers), and several cysteine-rich genes expressed as defensins in *M truncatula* (TC120449, TC124221, TC128939, TC138046, and TC132207) (Figure 3.9). Although there is a high divergence in the

mature NCR peptides, we still can see that the NCR genes belong to a large clade while the majority of the defensins in Arabidopsis, *O. sativa*, and *M. truncatula* belong to another clade that is more ancient than the NCR clade. Three Arabidopsis defensins are close to but are more ancient than the NCR genes. We can deduce from this phylogenetic tree that after the defensin gene duplication, some genes still remained defensins such as the five medicago defensins in the tree, while the other genes mutated such that they now seem to function in symbiosis.

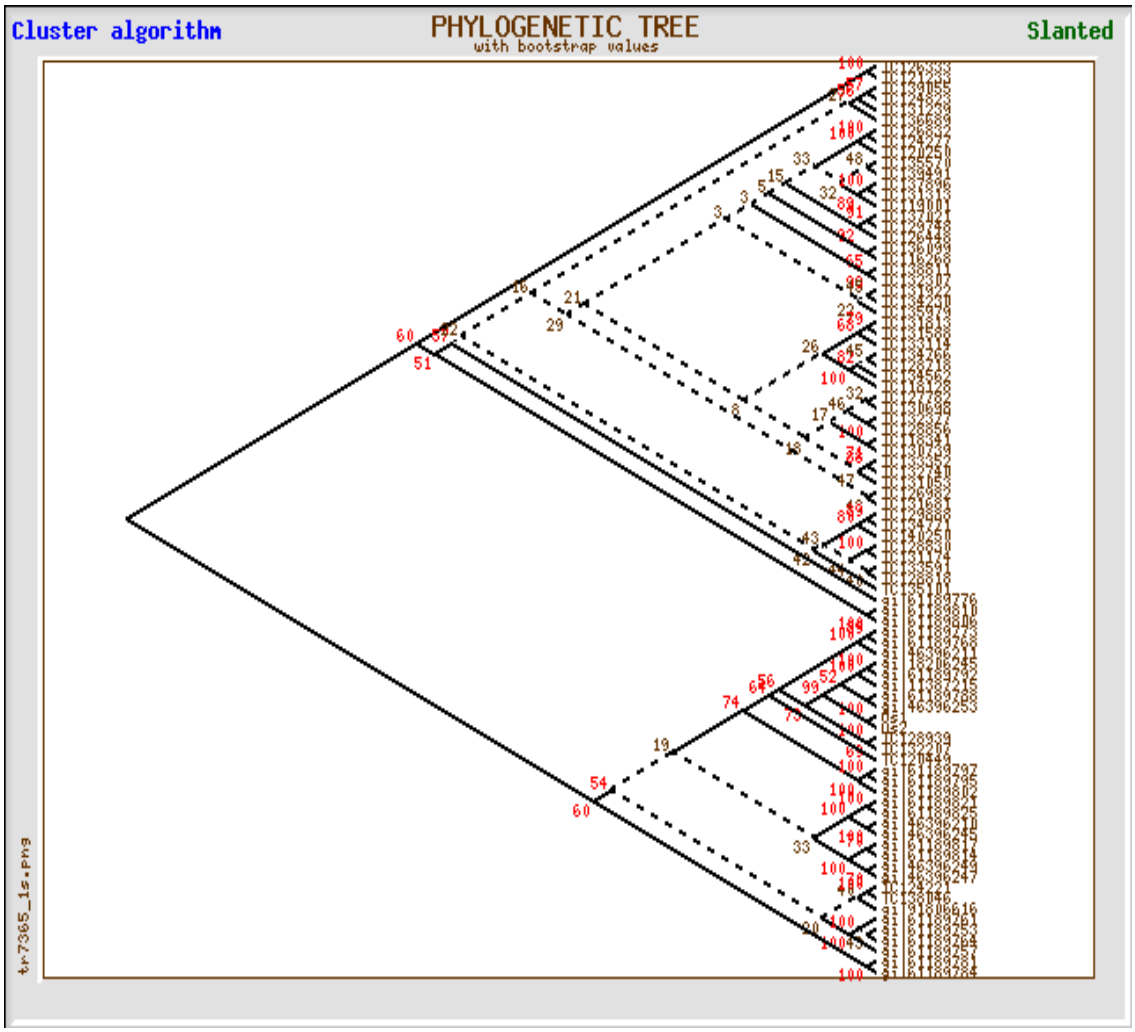


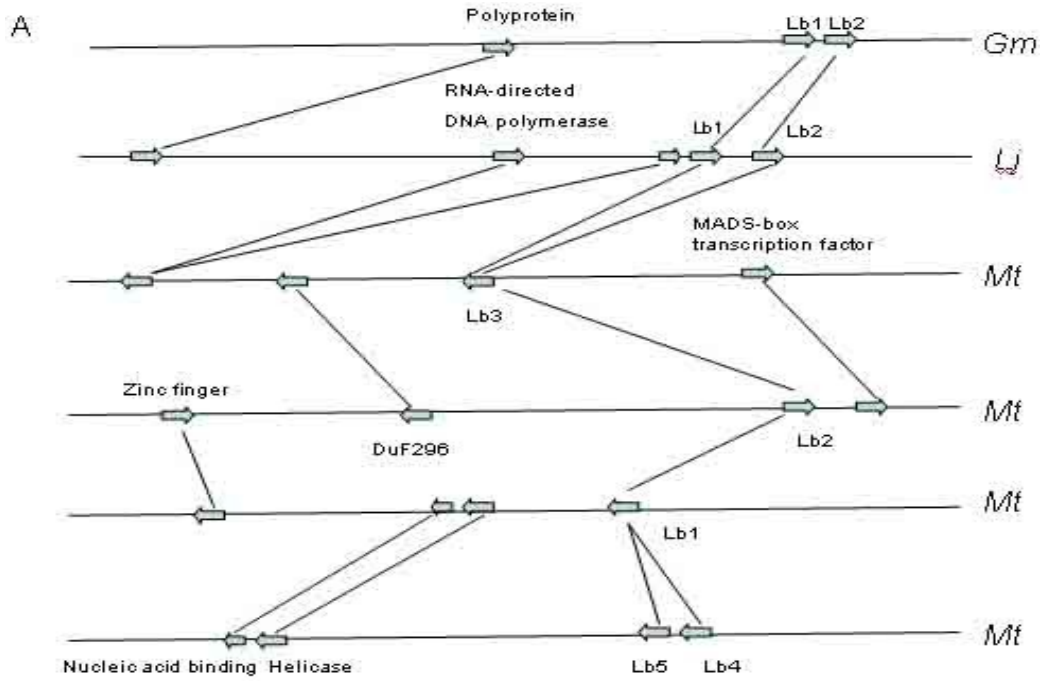
Figure 3.9 The phylogenetic tree of plant defensins and NCR genes

It was found that only the members in galegoid group of legumes (containing *M. truncatula, Pisum sativa, Vicia faba, Trifolium repens*) have NCR genes (Mergaert et al. 2003). Neither *L. japonicus* nor *G. max* has this gene family. Therefore, the NCR family could be specific to the galegoid group in legumes and other legumes could have different ways to deal with the functions performed by the NCR family.

3.2.1.2 Leghemoglobin

Ten leghemoglobin (Lb) TCs, one of the most abundant expressed nodule-specific genes, were observed. Each Lb-encoding TC was searched against the available *M. truncatula* genomic sequence using BlastN, and 5 Lb genes were found, 4 of which (Lb2, Lb3, Lb4 and Lb5) were on chromosome 5 and one (Lb1) was on chromosome 1. The identity ranged from 85% to 92% for the Lb cDNA sequence level and from 75% to 93% at the protein level. Lb4 and Lb5 may have resulted from tandem duplication since they were only 7,169 bp apart. Other Lb genes may have evolved thought segmental duplication as microsynteny can be identified in their neighbor regions (shown in Figure 3.9). To compare leghemoglobin genes in other legumes, the Lb genes in *Glycine max* (Gm) and *Lotus japonicus* (Lj) were identified by searching against the corresponding genome sequence using the corresponding leghemoglobin TCs. As a result, 3 Lb genes from *L. japonicus* and 4 Lb genes from *G. max* were found. Two lotus Lb genes were clustered and the 4 soybean Lb genes were found in two clusters. To investigate how

Lb genes might have evolved among these legumes, the DNA regions of 100 kb upstream and 100 kb downstream of each gene (including the gene) were extracted, masked repeats by Repeatmasker, and annotated using FGENESH. Then, genes in each region of one Lb gene were compared with those of another Lb gene using blastP. The genes with the highest homology by a reciprocal blast were chosen to draw the resulting figure (Figure 3.10a and b). Several genes in the neighborhood of Lb3 genes were found very similar to the genes near Lb1 and Lb2 genes in *Lotus japonicus* (Figure3.10 b). Lb3 and Lb4 in *Glycine max* were not shown in Figure 3.9a because in current Gm genome sequence, only sequences of the two genes were available. Lb4 in *Glycine max* was a pseudogene probably since it was truncated.



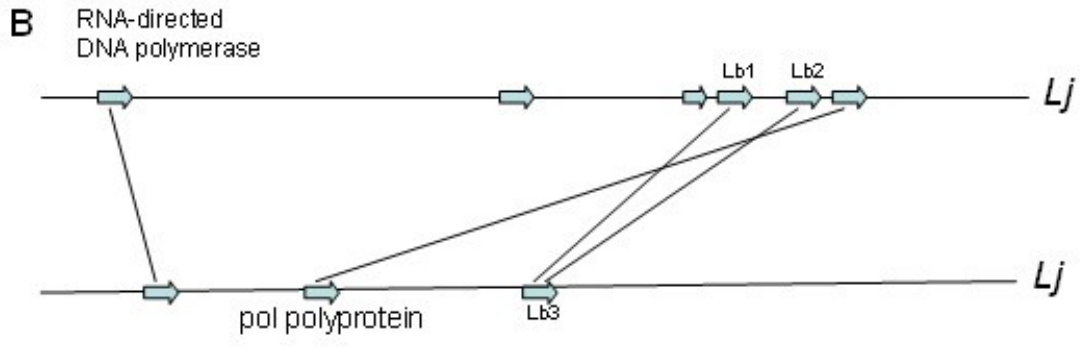


Figure 3.10 Microsynteny of leghemoglobin genes: (A) showed the microsyntenic relationship among *Glycine max*, *Lotus japonicus* and *Medicago truncatula*; (B) showed microsynteny in *Lotus japonicus*.

From Figure 3.10, we can observe the Lb segmental gene duplication in *Medicago truncatula*, although the genes in the neighborhood of the medicago Lb genes were not highly conserved, likely because of gene loss or rearrangement after segmental gene duplication. The genomic regions of *M. truncatula* and *L. japonicus* considered here probably share a common origin and the Lotus and soybean genomic sequences probably also share a common origin since somewhat microsynteny can be detected. To further compare the evolution of leghemoglobin genes, the Ka/Ks ratio was calculated between the genes and the result Ka/Ks tree was shown in Figure 3.11. All the Ka/Ks ratios were less than 1, indicating that Lb gene was under purifying selection. Leghemoglobin proteins predominantly have been found in legume nodules and Lb proteins function to help oxygen transport (Arredondo-Peter et al. 1998). Nonsymbiotic hemoglobins, another type of plant hemoglobins, are considered an ancestor of the Lb

proteins and have a high affinity for oxygen (Arredondo-Peter et al. 1998). To show the evolutionary relationship between these two types of plant hemoglobins, the Ka/Ks ratios were compared among the hemoglobins in *A. thaliana, O. sativa* and *M. truncatula* and the Lbs in *M. truncatula, L. japonicus* and *G. max*. The results shown in Figure 3.11 suggested that plant hemoglobins mostly underwent purifying selection although some positive selection was detected. The Lb genes in *G. max* are more closely related to the nonsymbiotic hemoglobins.

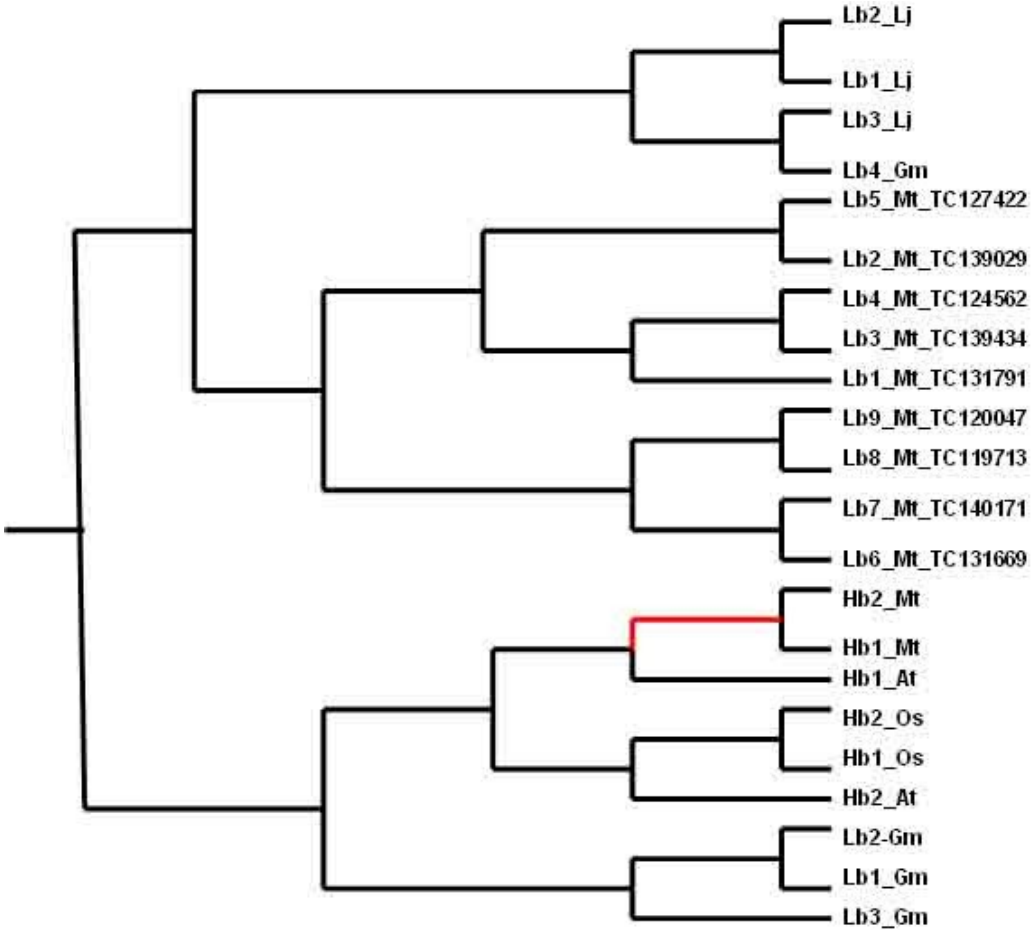


Figure 3.11 Ka/Ks tree of plant hemoglobins

To investigate the conservation of the cis-acting elements in Lb genes, 1000 bp regulatory regions upstream of Lb genes in *M. truncatula, L. japonicus* and *G. max* were searched against PLACE. The results, shown in Table 3.13, reveal that the putative nodule-specific consensus CTCCT motif could be found in all the regulatory regions of the Lb genes but the AAAGAT motif was less conserved as it could not be observed in two of the *L. japonicus* Lb genes and one of the *M. truncatula,* Lb genes. The lower conservation of the AAAGAT sequence might suggest it is less important than the CTCTT motif in the regulation of nodule-specific Lb gene expression. Another motif, AGATT, found in the promoter of *O. sativa* non-symbiotic haemoglobin-2 (NSHB) gene (Ross et al. 2004), also was detected upstream of all Lb genes. When the position of TATA box immediately upstream of Lb gene was used as a reference, the position of the CTCTT motif for the *M. truncatula* Lb gene was at approximately -50, and approximately -100 in Lotus, but less conserved in soybean as it was within 100 bp upstream of the gene. The positions of AAAGAT and AGATT also were much less conserved.

Table 3.13 Conserved motifs found in Lb genes from different organisms
(* indicates a motif that is located on the complementary strand)

Leghemoglobin	**Motifs**			
Gene	**AAAGAT**	**CTCTT**	**AGATT**	**TATA box**
Lb1_Gm	-92	-42*	-482	-112
Lb2_Gm	-86	-74/-39*	-485	-106
Lb3_Gm	-585	-91*/-312*	-420*	-32
Lb1_Lj	-302	-106	-338	-42
Lb2_Lj		-106/-194*	-346	-182
Lb3_Lj		-96	-172*	-149
Lb1_Mt		-52/-129	-57/-156/205	-30
Lb2_Mt	-162	-73/-150	-160/-196/-220/-249	-53
Lb3_Mt	-296	-53/-128	-360*	-32
Lb4_Mt	-141	-53	-91/-156/-257	-31
Lb5_Mt	-159	-71/-147	-110/-157	-50

3.2.1.3 Nodule-specific glycine-rich proteins

Glycine-rich proteins (GRP) have diverse structures in plants, although they all contain quasi-repetitive glycine-rich domains, for example GGGX, GGXXXGG or GXGX (Sachetto-Martins et al. 2000). Different GRPs are involved in different physiological processes since they have diverse expression patterns and subcellular localizations. Five GRP genes (Vfnod-GRP1-5) isolated from *Vicia faba* and 4 GRP genes from *Medicago sativa* (alfalfa) showed nodule-specific expression (Schröder et al. 1997, Kevei et al. 2002). Nodule-specific GRPs, as the NCR family, only were found in the galegoid group of legumes. In my study, 3 TCs (TC126384, TC127188 and TC119756) encoding nodule-specific glycine-rich proteins were identified. TC126384 (designated as GRP1), TC127188 (GRP2) and TC119756 (GRP3) mainly were expressed in the young nodule library, MtSN4, the senescent library, GVSN and mature library, GVN. All of the nodule-specific GRPs contain a putative hydrophobic N-terminal signal peptide predicted by SignalP 3.0 (Bendtsen et al. 2004) and a glycine-rich C-terminal. GRP1 has little homology with GRP2 and GRP3 except that they are rich in glycines with GRP2 and GRP3 about 85% identical at protein level and about 60% identical at the coding nucleotide level.

When these three GRPs were searched against the Medicago genomic sequence using BlastN to find their locations on chromosomes, all were found located on chromosome 2. The GRP2 and GRP3 genes are only 3.8 kb apart, clustered on the reverse strand

while GRP1 gene is about 5.2 Mbp away from the GRP2 and GRP3 genes. Both the GRP1 and GRP 2 genes contain 2 exons but the GRP3 gene contains 4 exons.

To further investigate the evolutionary relationship between GRP2 and GRP3 cluster, ClustalW was used to compare GRP2 and GRP3. The result showed that the major difference between these two GRPs was that there was an 80 amino acid deletion in GRP2 that makes it much shorter than GRP3 (Figure 3.12). Therefore, these two genes likely resulted from tandem duplication followed by either a gene deletion or an insertion. The alignment of GRP1, GRP2 and GRP3, in Figure 3.13, shows that the N-terminal signal sequences, the sequences near the signal sequences and glycines in the C-terminal sequences are somewhat conserved but other sequences in GRP1 diverge greatly from GRP2 and GRP3. These results suggest that GRP1 and GRP2-GRP3 genes might have the same ancestor but that GRP1 diverged greatly from GRP2-GRP3 after gene duplication.

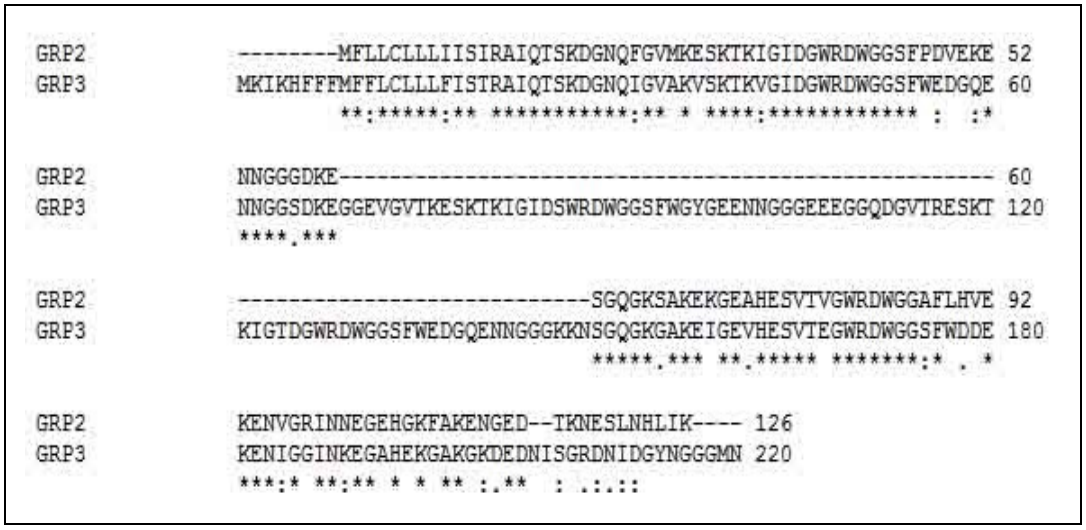


Figure 3.12 Clustal W result showing the conservation and divergence of GRP2 and GRP3

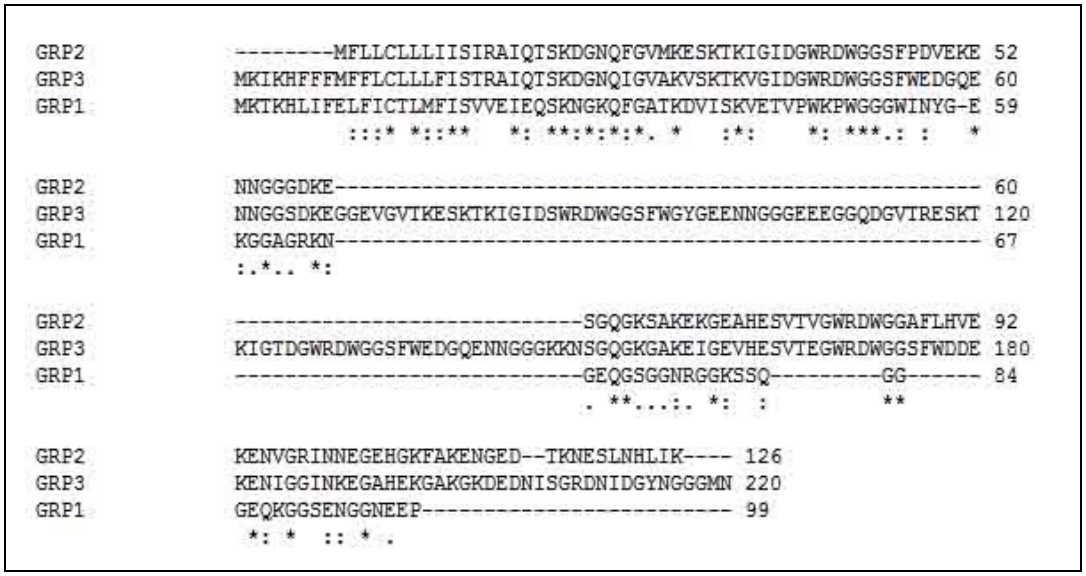


Figure 3.13 Clustal W result to show the alignment among GRP1, GRP2 and GRP3

The 1 kb upstream sequences also were compared and a conserved 200 bp region (Figure 3.14) was found. The conservation of the upstream sequence again suggested

GRP1, GRP2 and GRP3 might share a common ancestor. Nodule-specific regulatory motifs also were observed within 500 bp upstream of the genes (Table 3.14). However, both AAAGAT and CTCTT only could be found in GRP1, CTCTT only was found in GRP2, and AAAGAT only was found in GRP3.

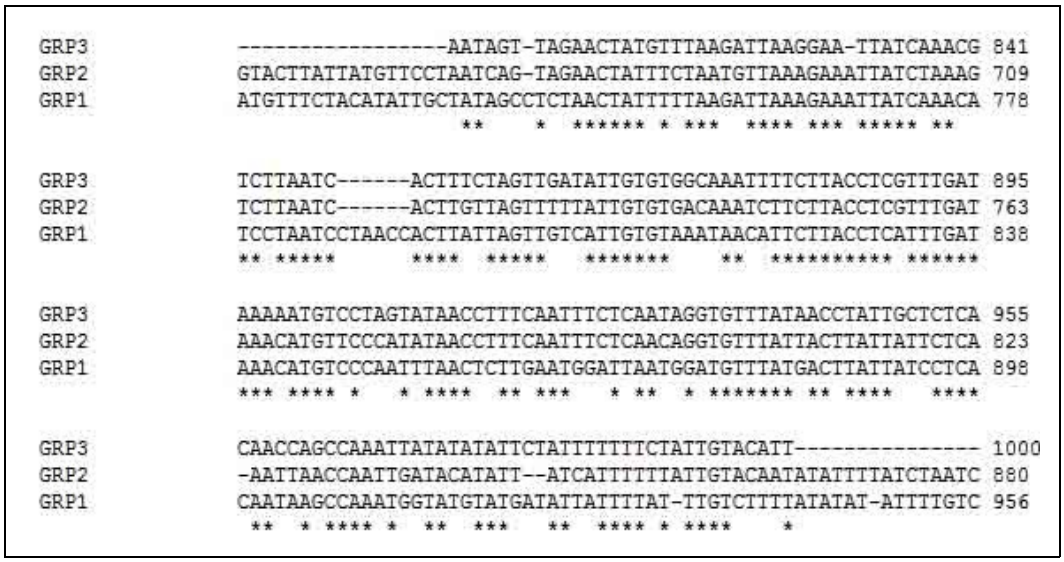


Figure 3.14 The conservation among the upstream sequences of GRP1, GRP2 and GRP3

Table 3.14 Nodule-specific regulatory motifs in GRP genes

	Motifs			Nodule
Name	**AAAGAT**	**CTCTT**	**TATA box**	**library**
GRP1	-447	-144	-71	Young
GRP2		-494	-105	Old
GRP3	-209*		-33*	Mature

To understand the evolution of nodule-specific GRP in different legumes, the GRPs in medicago were compared with GRPs in *Vacia faba* and in *Medicago Sativa*. The

resulting phylogenetic tree (Figure 3.15) revealed that GRP2 and GRP3 belong to one clade and while GRP1 belongs to another.

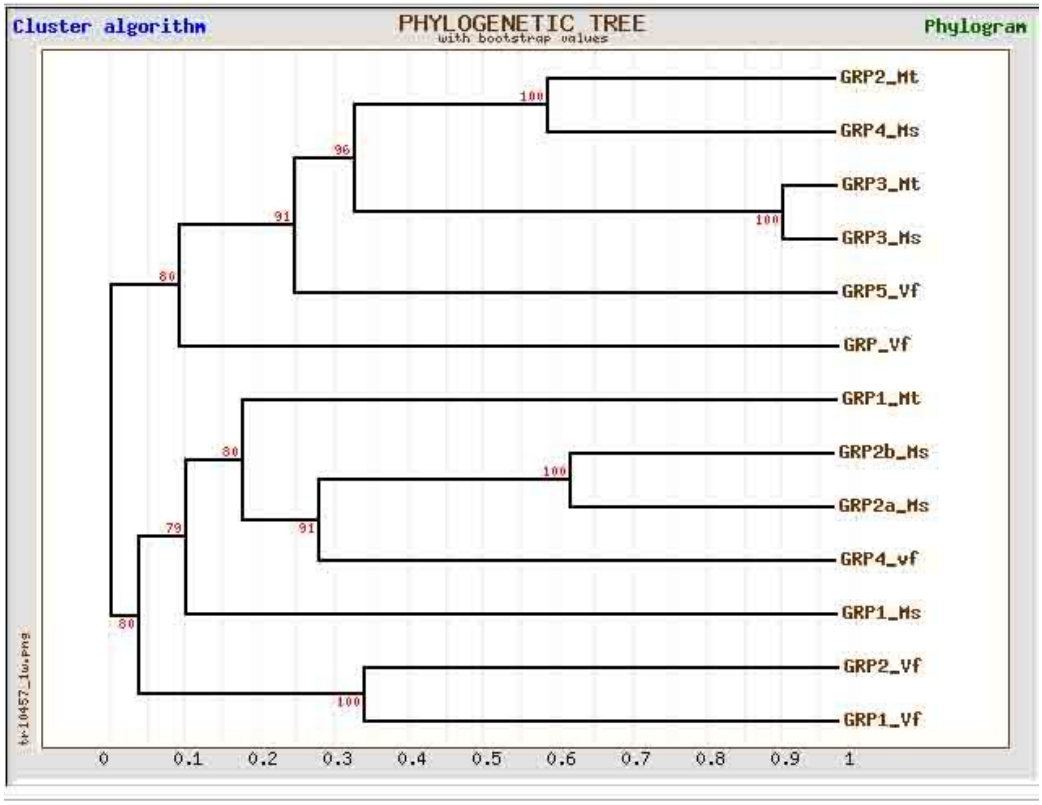


Figure 3.15 The phylogenetic tree of GRP genes in *Medicago truncatula*, *Vacia faba*, and *Medicago sativa*

To examine the evolutionary pressures on GRP genes, the rates of nonsynonymous (Ka) and synonymous (Ks) were determined among these genes. The resulting Ka/Ks tree (Figure 3.16) shows that positive selection plays a very important role in GRP gene evolution. The ratio for purifying selection ranged from 0.35 to 0.94 while the ratio for the positive selection ranged from 1.22 to 2.21(Table 3.15).

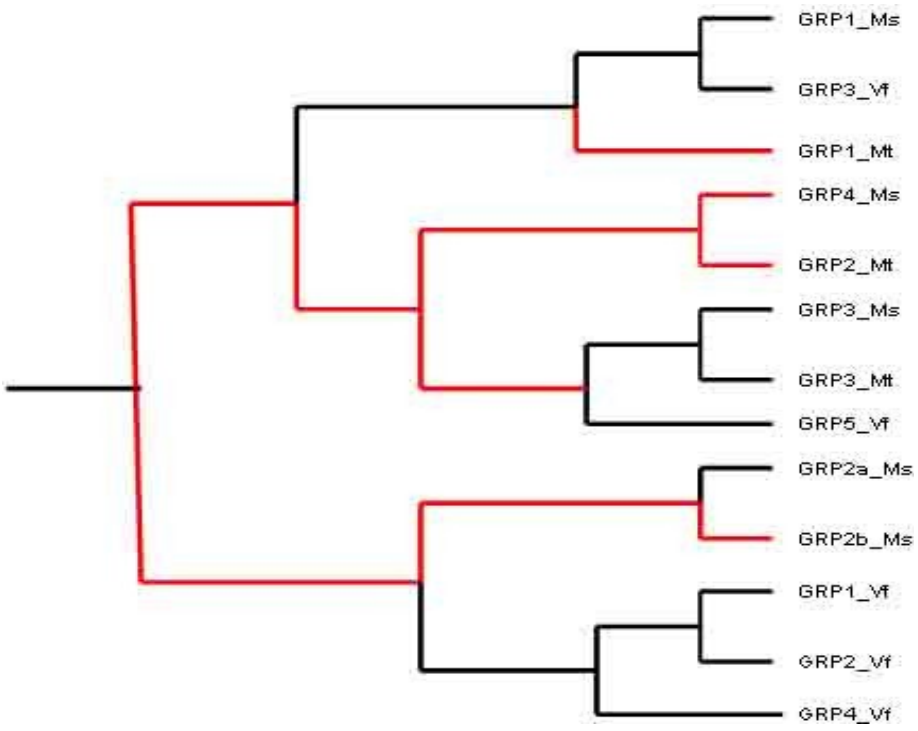


Figure 3.16 The Ka/Ks tree of GRP gene (the red line indicates positive selection and the black line indicates negative selection)

Table 3.15. Ka/Ks value for each node in the tree

Node#	Ka/Ks branch 1	Ka branch1	Ks branch1	Ka/Ks branch 2	Ka branch2	Ks branch 2
1	0.7801	0.3677	0,4713	0.9367	0.2498	0.2667
2	0.9220	0.1322	0.1434	1.2527	0.3024	0.2414
3	1.2803	0.0186	0.0145	1.3492	0.0261	0.0194
4	0.5339	0.0219	0.0410	0.9983	0.0279	0.0280
5	0.7697	0.0871	0.1131	0.6763	0.1848	0.2733
6	1.3001	0.1250	0.0961	1.0088	0.0837	0.0830
7	0.3478	0.0741	0.2129	1.00665	0.1850	0.1735
8	0.6797	0.0619	0.0911	2.2166	0.0446	0.0201
9	0.8230	0.1494	0.1816	0.5659	0.1736	0.3067
10	0.6755	0.3519	0.5210	0.9407	0.2643	0.2809
11	1.9239	0.1490	0.0775	0.5603	0.1024	0.1828
12	1.4451	0.1201	0.0831	1.2224	0.1236	0.1011

3.2.1.4 Nodulins

TC132746 encodes the *M. truncatula* nodulin 1gene, MtN1, which is similar to a defense protein in *Pisum sativum* (common pea) (Gamas et al. 1998). This 73 amino acid protein contains an N-terminal hydrophobic signal peptide and is rich in cysteine. The deduced amino acids from TC96169 are 100% identical to MtN6, a gene that was considered as a marker of a pathway involved in preparation to infection (Mathis et al. 1999). The early nodulin 12 precursor (N-12) encoded by TC 127504 has been postulated to play a role in the pre-infection processes (Bauer et al. 1997) while TC114187 encodes a protein similar to ENOD18 in *Vicia faba* that belongs to a novel ATP-binding family in plants (Becker et al. 2001). TC114025, TC131314, TC129897 and TC138204 encode MtN24, MtN25, MtN29, and MtN11, respectively. Although the functions of these three nodulins are not known, they are expressed in the early stages of nodulation. None of the TCs mentioned above are present in the most recent *M. truncatula* genome assembly but other nodulation genes, such as TC134290, whose corresponding gene is located on chromosome 8, encodes the single copy of MtN15. The deduced amino acid sequence from TC131451 was similar to MtN16 and a portion of this gene is located on an incomplete BAC sequence from chromosome 5. TC115962, TC115929 and TC121205 encode the same nodulin MtN22 from a single MtN22 gene located on chromosome 3 that is mainly expressed in mature library GVN.

Nod25, encoded by TC113614, is located about 3.8 kb from MtN22 on chromosome 3. Interestingly, 6 copies of nodule-specific Calmodulin-like were found clustered with Nod25 and MtN22 within a 59-kb region. TC113188 was identified as the transcript of ENOD8.1 that encodes a nodule-specific esterase locating in symbiososme membrane or symbiosome space around the bacteriods in the infected nodule cells (Coque et al, 2008). ENOD8.1 was highly expressed in nodules since TC113188 is composed of 145 ESTs in which 60% came from young nodule library MtBB. A cluster of 6 Enod8 genes was found within 20 kb on chromosome 1 (Figure 3.17). Interestingly, only ENOD8.1 was exclusively expressed in nodules at high level while ENOD8.2 gene (corresponding TC113207) was expressed in developing root, ENOD8.3 (TC113511) was expressed in aphid- infected shoots, and ENOD8.6 (TC122750) was expressed in pod walls (GPOD library) and pods with seeds (MTPOSE library) at low level. Their coding sequence identity ranged from 79% to 91% and their amino acid identity ranged from 53% to 87%. ENOD8.1, ENOD8.5 and ENOD8.6 genes contain 5 exons, ENOD8.2 and ENOD8.3 have 6 exons and ENOD8.4 only has one exon that is most similar to the first exon and 40 bp of first intron of ENOD8.5 (Figure 3.18), suggesting that ENOD8.4 originated from ENOD8.5 by unequal recombination crossover.

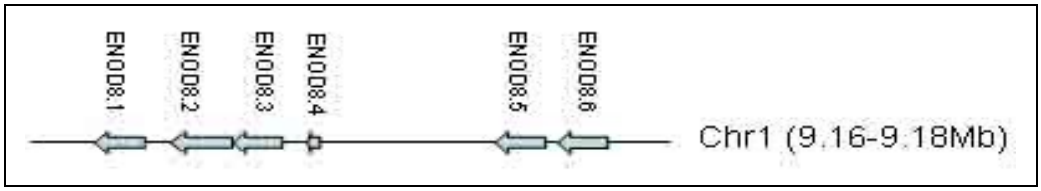


Figure 3.17 ENOD8 gene cluster on chromosome 1

```
ENOD8.5.exon1+intr      ATGGCACCAATTGAG------------ATGTCACTTGTGTCTTTGATTGTGCTTATTTTA 48
ENOD8.4_Mt              ATGGCACAAATAGAGTTTCTTAGGCACATGTCACTTGTGTCATTGATTGTGCTTATTTTA 60
                        ******* *** ***            ************** ******************

ENOD8.5.exon1+intr      TGTACTAGTGCCCCAATATCTGCCACAAATTTTTTTGATTGTCCTGCCATATTTAACTTT 108
ENOD8.4_Mt              TGTTCTACACCCCCCATATTTGCCACAAAGAATTGTGATTTTCCTGCCATATTTAACTTT 120
                        *** ***   **** **** *********   ** ***** ******************

ENOD8.5.exon1+intr      GGTGCTTCAAACGCTGATACTGGTGGTTTGGCTGCTGCCTTTCAAGCACTTCAGTTGCCA 168
ENOD8.4_Mt              GGTGCTTCAAACTCTGATACTGGTGGTTTAGCTGCTGCCTTTCAAGCACTGCCGTTGCCA 180
                        ************ ********************* ******************* * *******

ENOD8.5.exon1+intr      AATGGAGAGACTTTTTTCAATAGATCAACTGGAAGATTTTCTGATGGACGAATCATTATT 228
ENOD8.4_Mt              AATGGAGAGACTTTTTTCAATAGATCAACTGGAAGATTTTCTGATGGTCGAATCATTCTT 240
                        ********************************************** ********* **

                                          ↓
ENOD8.5.exon1+intr      GATTTCATAGGTAATCCT-TTCTTTCTCTATTTCATTTTCCATATTGTAGATAACTACAG 287
ENOD8.4_Mt              GATTTCATAGGAAATATTATTATTTTTCTTTCTTTTCAAGTTCATTTTCTATGA------ 294
                        *********** ***  * ** *** *** * *  *       *** *  ** *

ENOD8.5.exon1+intr      TAAGAAATATTGCACGGTCGGTGTAAATAAGTTCTACACATACATCTAATA 338
ENOD8.4_Mt              ------------------------------------------------------
```

Figure 3.18 Clustal W alignment on ENOD8.4 with the first exon and partial first intron of ENOD8.5 (The arrow shows where the first exon ends)

To investigate evolutional history of ENOD8 gene, a phylogenetic tree was constructed from the amino acid sequence of ENOD8 from *A. thaliana*, *O. sativa*, *L. japonicus*, *G. max*, *P. trichocarpa, and M. truncatula* (shown in Figure 3.19). The tree indicates that 4 ENOD8 genes in *M. truncatula* likely arose from a single ENOD8 progenitor. The step-like pattern observed in the *M. truncatula* ENOD8 clade suggested these genes evolved by tandem duplication followed by divergence. ENOD8.4 is much shorter and not so

comparable with the rest of ENOD8 genes. It seemed like ENOD8.6 is more ancient than ENOD8.1, ENOD8.3, and ENOD8.5, suggesting that ENOD8.6 may be the ancestor of the 3 ENOD8 genes. These four genes may share a common ancestor with Enod8-like genes in Lotus and in soybean since they are in the same clade. Enod8.2 is close to Enod8-like gene in Populus. Since ENOD8-like genes can be detected in rice and in Arabidopsis, ENOD8 genes may come from the same ancestor preceding the separation of dicots and monocots.

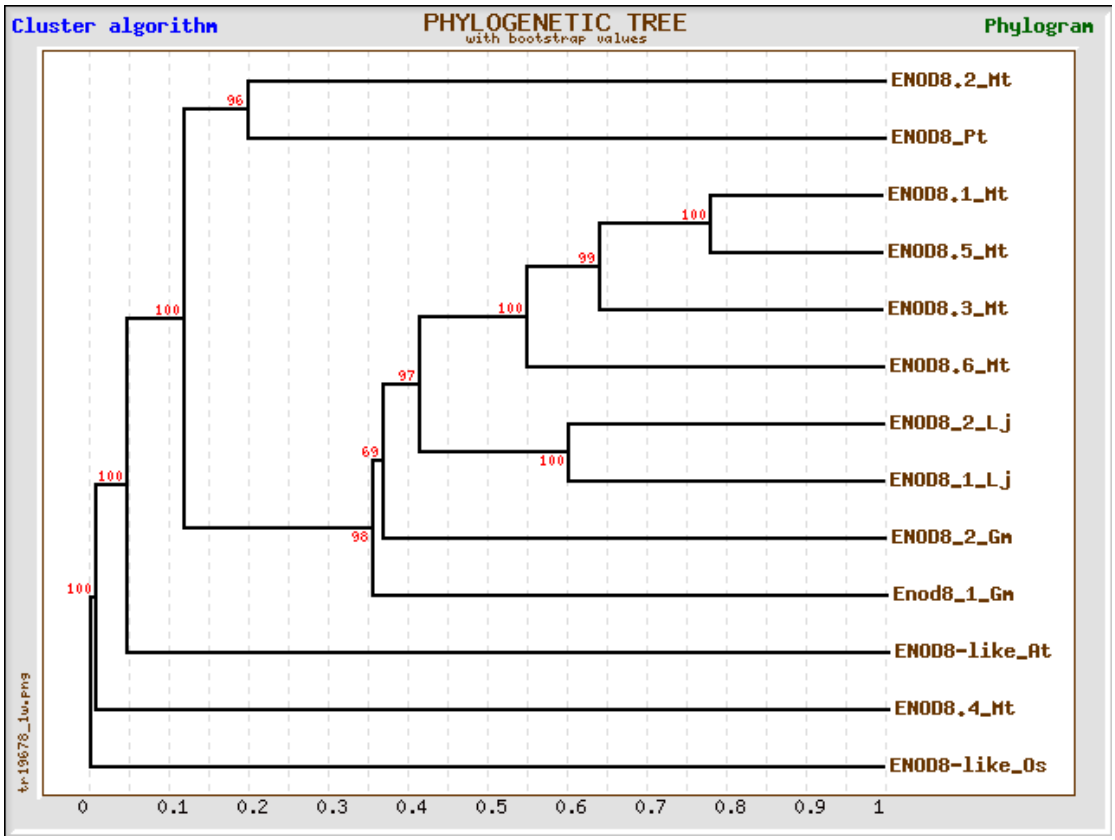


Figure 3.19 Phylogenetic tree of ENOD8 genes

Ka/Ks ratios between ENOD8 genes mainly ranged from 0.17 to 0.72, indicating that ENOD8 genes were mostly under purifying selection although there was one Ka/Ks

value greater than 1 and also indicating that their function is conserved even though that they are expressed in different organs or tissues.

The analysis of 1000 bp upstream the above nodulins which could be located on the *M. truncatula* genome also found that the two nodule-specific motifs CTCCT and AAAGAT near or in the promoter regions.

3.2.2 Genes expressed in nodules tend to cluster on *M. truncatula* chromosomes

When all the nodule-specific TCs were compared with the *M. truncatula* genomic sequence to find their locations using BlastN, the location of the genes for 96 TCs were found. When all 8 pseudomolecules were searched against all the *M. truncatula* TC and singleton ESTs sequences using BlastN, 47 out of 96 TCs were found on different chromosomes clustered with other nodule-specific TCs or ESTs expressed in nodule libraries and located within 50 kb of each other (Figure 3.20).

As seen below in figure 3.17, the neighboring co-expressed genes are oriented in three alternative combinations: parallel transcription (→→ or ←←), divergent transcription (← →), or convergent transcription (→←). The clustered genes in parallel direction are more common than in convergent or divergent direction. The members of gene clusters in nodulation may be expressed in different developmental stages but most of the neighboring genes are expressed in the same stage.

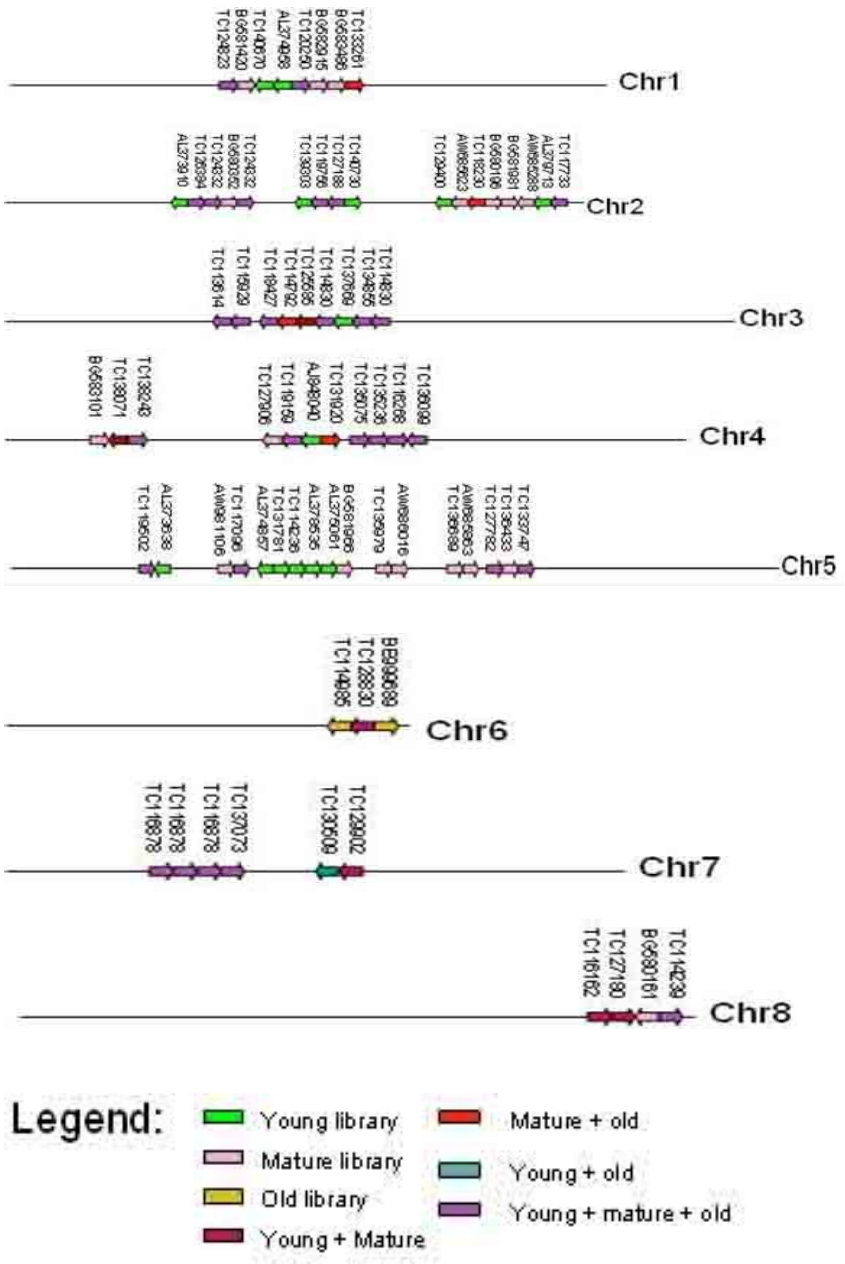


Figure 3.20 The clusters or colocalization of nodule-specific genes in *M.truncatula*

In prokaryotes, genes functioning in the same pathway often are clustered into operons that then are transcribed into a single polycistronic mRNA. In eukaryotes, operons are very rare and only are found in nematodes and trypanosomes, although these operons function differently from those in prokaryotes (Blumenthal et al, 1998; Blumenthal et

al. 2002). Gene transcription in eukaryotes is controlled by trans-acting factors that do not require the co-transcribed genes to be located near each other (Niehrs et al. 1999). However, previous studies showed that gene order in eukaryotic genomes is not completely random and that genes with similar expression patterns tend to be clustered together (Bortoluzzi et al. 1998, Lercher et al. 2002, and Birnbaum et al. 2003). In *Drosphila melanogaster*, about 20% of genes are organized into clusters and range from 10 to 30 genes within up to 200kb (Spellman et al. 2002). In the mouse genome, both housekeeping and immunogenic genes are clustered (Williams and Hurst, 2002), and in human genome, clusters of muscle-specific genes (Bortoluzzi et al. 1998), highly expressed genes (Caron et al. 2001) and housekeeping genes (Lercher et al. 2002) have been found. In plants, clustered genes in root development (Birnbaum et al. 2003) and mitochondrial function (Elo et al. 2003) have been identified in *Arabidopsis thaliana.* For example, Lee et al. (Lee et al. 2003) found that genes in a pathway are in closer proximity than would be expected by chance in five sequenced eukaryotic genomes. Why do coexpressed genes or genes in the same pathway tend to be clustered when colocalization in the genome is not generally necessary because the transcription factor system is sufficient for coregulation of widely dispersed genes in eukaryotes? There are some possible explanations to these apparent contradictory observations. For example, the clustering of functionally related genes may result from recent tandem duplications and has nothing to do with aiding coregulation of gene expression as is the case with 7 copies of nodule-specific calmodulin-like protein genes that are clustered within 40kb

on chromosome 3 (TC118427, TC114792, TC125585, 2 copies of TC114830, TC137669, and TC134855). However, since in most cases, clustered genes in the same pathway as seen in nodulation do not have sequence similarity, they likely are not the result of recent tandem duplication. Since the eukaryotic genome needs to fold tightly to fit into the nucleus and energy must be expended to unfold regions of DNA when gene transcription occurs, keeping functionally related genes in close proximity, even if not adjacent, could reduce the amount of energy required to unfold larger regions of the genome during transcription of numerous genes involved in a single pathway (Lee et al. 2003). Therefore, clustering must be advantageous as it often is preserved by nature selection. Another possibility is that the close proximity of genes in a pathway or a biological process might lead to sharing of cis-regulatory elements such as enhancers. Interestingly, it was found in *Arabidopsis thaliana* that gene pairs with divergent (← →) or parallel (→→ or ←←) orientation have a higher degree of coexpression than those genes with convergent (→←) orientation (Williams et al. 2004), which indicates the possibility of share cis-regulating elements.

4. Conclusions

4.1 The *Medicago truncatula* genomic sequence and predicted features

About 255 Mbp of the euchromatic regions of the Medicago genome have been sequenced and analyzed, revealing that the genome encodes at least 50,540 putative protein-encoding genes and 11% repetitive elements. The gene density of the *M. truncatula* is similar to that in *A. thaliana* and it is almost twice as much as that in *O. sativa*, *P. tricorcarpa,* and *L. japonicus* and three times as much as that in *G. max.* Approximately 50% of the predicted genes had a high identity match with a plant EST or TC, which since it is less than that in *O. sativa* and in *A. thaliana* (both ~60%), indicates approximately 80% of the *M .truncatula* genes have been captured by the Medicago genome sequencing up to date and only about 50 Mbp more euchromatic regions still need to be sequenced. The *M .truncatula* proteins on average have a shorter amino acid sequence, likely due to a larger than normal number of small (<99 amino acid peptides, the lowest exon number, and the second lowest gene size in average among the *M. truncatula*, *L. japonicus*, *G. max*, *P. trichocarpa*, *O. sativa*, and *A.thaliana* genome. Nearly 40% of the predicted *M .truncatula* genes are intronless and about 55 % of them are expressed. A comparison of the average intron size and exon size among the above six organisms showed that the average exon size is quite conserved whereas the intron size differs, suggesting natural selection is acting on the

exon size and keep it conserved. Fourteen out of 17 miRNA families are conserved in the six plants, suggesting a core set of miRNAs are required for regulating the expression of similar plant genes. The comparison of the predicted proteins in *M .truncatula* with the predicted proteins in the other five plants revealed that the proteins are more conserved in the legumes but least conserved in rice, the monocot.

The comparison between GO annotation results among *M.truncatula*, *L. japonicus*, *G. max*, *P. trichocarpa*, *O. sativa*, and *A.thaliana* genomes revealed that all the six genomes have similar percentage of each of the major functional domains. The comparison of the top 40 Interpro domain hits in *M. truncatula* with the corresponding domains in the other five plants also indicated that most of the overrepresenting domains are overrepresenting in all the six genomes. Therefore, we can conclude that in angiosperms, they all have a fixed percentage of major domains functioning in basic biological processes whether they are dicots or monocots. However, they also may contain species-specific domains functioning in species-specific biological processes such as late nodulin domain in *M. truncatula*.

4.2 Nodule-specific genes

The *in silico* analysis of the Medicago Gene Index 9.0 revealed that 191 genes only are expressed at nodules, 100 of which were found similar to sequences from known genes in GenBank. The analysis of 50 nodule-specific cysteine-rich peptides (NCR) showed

that they have a conserved signal peptide, a conserved cysteine motif, and a highly divergent remaining sequence. These NCR genes are clustered or dispersed on the *M .truncatula* genome, suggesting that they most probably underwent tandem or segmental gene duplication. The Ka/Ks analysis of NCR genes indicated that some NCR genes underwent positive selection and some underwent purifying selection. NCR genes were thought to have evolved from antimicrobial defensins to avoid infections by other soil microorganisms during nodule formation or alternatively they act as signal molecules assuring communication between plant cells or between plant cells and rhizobial bacteria (Mergaert et al. 2003). Therefore it is very likely that many NCR genes were under positive selection to adapt to the environment with wide-spectrum soil microorganisms and rather than prevent symbiotic relationships with selected microbes that the plant encouraged them. The maintaining of the nonsynonymous mutations of the NCR genes makes it possible that the legumes recognize the changing non-symbiotic microbes in the environment and prevent the nodules from invading by them. Two motifs (AAAGAT and CTCCT) are highly conserved in NCR regulatory regions and they mostly are located at the proximal promoter regions or even in the promoter regions. Thus, since the NCR introns are highly conserved, it may be that they act as an enhancer and determine nodule-specific NCR gene expression in combination with the conserved upstream cis-acting motifs.

All three glycine-rich genes in *M .truncatula* are located on chromosome 2. The close proximity and highly conserved sequence with a deletion between GRP2 and GRP3

indicate that they arose from tandem gene duplication followed by either a deletion or an insertion. GRP1 is highly diverged from GRP2 and GRP3 and the phylogenetic tree showed that GRP1 belong to a different clade from GRP2 and GRP3, however, the conservation of the signal peptide, the glycine-motif, and the 200bp upstream DNA sequence suggest that they share a common ancestor. The Ka/Ks analysis indicated that positive selection played an important role during GRP gene evolution. As GRP genes found in *Medicago sativa* (Kevei et al. 2002), GRP genes in *M. truncatula* also are expressed in different nodule developmental stages, suggesting GRPs might play distinct, nonredundant roles during nodule development.

Leghemoglobin (Lb) proteins that are found in nodules of legumes transport oxygen to nitrogen-fixing endosymbiotic bacteria (Trevaskis et al. 1997). Nonsymbiotic hemoglobins that are found in legumes and nonlegumes are believed to be the ancestor of Lbs (Arredondo-Peter et al. 1998). The analysis of Ka/Ks among Lb and Hb shows that plant hemoglobins underwent purifying selection, suggesting that natural selection conserves the important function of plant hemoglobins. Since NCR and GRP genes, and Lb genes have nodule-specific motifs (CTCCT and AAAGAT), almost all the nodule-specific genes that could be located on the *M .truncatula* chromosomes contain these two nodule-specific motifs or either one of them, an observation suggesting that expressions of nodule-specific genes likely are co-regulated.

The analysis of the genome organization of all nodule-specific genes showed that about 50% of them are clustered with each other or that the corresponding gene is expressed

in nodules. However, since most of the nodule-specific gene clusters do not have a high degree of sequence similarity, it may be that keeping functionally related genes in close proximity could be advantageous and thus the clusters are preserved by natural selection.

Finally, the study of the nodule-specific genes indicated that some of them evolved through duplication and modification of plant defense genes, as for example observed with the defensin genes. The phylogenetic tree of defensin and NCR genes revealed that after the defensin gene duplication, some genes still remained defensins such as the five medicago defensins in the tree, while the other genes mutated such that they now seem to function in symbiosis.

References

Adrian B. (2007). "Perceptions of epigenetics". Nature 447: 396-398.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J.Mol.Biol. 215:403-10.

Andersen JW, Story L, Sieling B, Chen WJ, Petro MS, Story J (1984) Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. Am J Clin Nutr 40: 1146-1155.

Arredondo-Peter R, Hargrove MS, Moran JF, Sarath G., Klucas RV (1998) Plant hemoglobin. Plant Physiol.118: 1121-1125.

Avery OT, Macleod CM, Mcarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J.Exp. Med. 98: 451-460.

Backus JW, Smith HC (1992) Three distinct RNA sequence elements are required for efficient apolipoprotein B (apoB) RNA editing in vitro. Nucl. Acids Res. 20: 6007-6014.

Bankier AT, Weston KM, Barrell BG. (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Meth. Enzymol. 155, 51-93.

Barciszewska MZ, Erdmann VA, Barciszewski J. (1994) A new type of RNA editing. 5S ribosomal DNA transcripts are edited to mature 5S sRNA. Biochem Mol Biol Int.34 (3): 437-448.

Becker JD, Bauer P, Poirier S, Ratet P, Kondorosi A (1997) MsEnod12A expression is linked to meristematic activity during development of indeterminate and determinate nodules and roots. Mol. Plant-Microbe Interact. 10:39-49.

Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant R-genes. Science 292: 2281-2285.

Birnbaum K, Shasha D. E, Wang JY., et al. (2003) A gene expression map of the Arabidopsis root. Science 302: 1956-1960.

Birnboim HC, Doly J (1979) A rapid alkaline extraction procedure for screening

recombinant plasmid DNA. Nucleic Acides Res. 7: 1513-1523.

Blanc G and Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667-1678.

Blumenthal TD, Evans CD. Link CD, et al. (2002) A global analysis of Caenorhabditis elegans operons. Nature 417:851-854.

Blumenthal T (1998) Gene clusters and polycistronic transcription in eukaryotes. Bioessays 20: 480-487.

Bontems F, Roumestand C, Gilquin B, Menez A, Toma F (1991) Refined structure of charybdotoxin: common motifs in scorpion toxins nd insect defensins. Science 254: 1521-1523.

Bortoluzzi S, Rampoldi L, Simionati B, et al. (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res. 8:817-825.

Bodenteich A., Chissoe S., Wang YF., and Roe BA. (1994) Shotgun cloning as the strategy of choice to generate templates for high throughput.

Brewin NJ. et al. (1998) Tissue and cell invasion by Rhizobium: the structure and development of infection threads and symbiosomes. In The Rhizobiacea, the molecular biology of model plant associated bacteria, Spaink Hp, Kondorosi A, Hooykaas PJJ (eds). Kluver Academic Publishers: Dordretch; 417-429.

Broekaert WF, Terras FRG, Cammue BPA, Osborn RW (1995) Plant defensins: novel antimicrobial peptides as components of the host defense system. Plant physiol. 108: 1353-1358.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Cannon SB, Crow JA, Heuer ML, Wang X, et al. (2005) Databases and information integration for the Medicago truncatula genome and transcriptome. Plant physiol.1: 38-46.

Cannon SB, Sterck L, Rombauts S, Sato S, et al (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proc Natl Acad Sci USA 99: 13627-13632.
Caron H., van Schaik B., van der Mee M., et al. (2001) The human transcriptome map:

clustering of highly expressed genes in chromosomal domains. Science 291: 1289-1292. Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia, 6(6):201-209

Charon C, Sousa C, Crespi M, Kondorosi A (1999) Alternation of enod 40 expression modifies Medicago truncatula root nodule development induced by Sinorhizobium meliloti. Plant Cell 11:1953-1966.

Cheng Z, Dong F, Langdo T, et al. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14, 1691–-1704

Chissoe, SL, Wang, YF, Clifton, SL, Ma, N, Sun, HJ, Lobsinger, JS, Kenton, SM, White, JD, and Roe, BA (1991) Strategies for rapid and acute DNA sequencing methods: A comparative to methods in enzymology 3: 55-65.

Choi HK, Mun JH, Kim DJ, et al. (2004) Estimating genome conservation between crop and model legume species. Proc. Natl. Acad. Sci. U.S.A 101: 15289-15294.

Conesa A, Götz S, Miguel J, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

Cook D.R., (1999) Medicago truncatula - a model in the making! Curr. Opin. Plant Bio.2: 301-304.

Coque L, Neogi P, Pislariu C, Wilson KA, Catalano C, Avadhani M, Sherrier DJ, Dicksterin R. (2008) Transcription of ENOD8 in Medicago truncatula nodules directs ENOD8 esterase to developing and mature symbiosomes. Mol. Plant Microbe Interact. 21: 404-410.

Crespi M, Galvez S (2000) Molecular mechanisms in root nodule development. J Plant Growth Regul 19: 155-166.

Crick FHC (1958) On protein synthesis. Symp. Soc. Exp. Biol. XII, 139-163.

Crick F. (1970) Central Dogma of Molecular Biology. Nature 227: 561-563.

DeAngelis MM, Wang DG., and Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Research 23 (22): 4742-4743

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27 (23): 4636-4641.

Deschamps S, Meyer J, Chatterjee G, Wang H, et al. (2003) The mouse Ifi200 gene cluster: genomic sequence, analysis, and comparison with the human HIN-20 gene cluster. Genomics 82: 34-46.

Dixon RA, Achnine L, Kota P, Liu CJ, Reddy MSS, Wang LJ (2002) The phenylpropanoid pathway and plant defense: a genomics perspective. Mol Plant Pathol 3: 371-390.

Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol 131: 900-910.

Duret L (2000) tRNA gene number and codon usage in the C.elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Gene. 16: 287-289.

Eckardt N.A. (2001) Everything in its place: conservation of gene order among distantly related plant species. The Plant Cell 13: 723-725

Eckardt N.A. (2001) The new biology: Genomics fosters a “systems approach” and collaborations between academic, government, and industry scientists. The Plant Cell 13: 725-732.

Eddy S.R. and Durbin R. (1994) RNA sequence analysis using covariance models. Nucleic Acid Res.22 (11): 2079-2088.

Eddy S.R. (2001) Non-coding RNA genes and the modern RNA world.

Elo A., .Lyznik A., Gonzalez DO, et al. (2003) Nulear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the Arabidopsis genome. Plant cell 15: 1619-1631.

Endre G, Kereszt A, Kevei Z, Mihacea S, Kalo P, Kiss GB (2002) A receptor kinase gene regulating symbiotic nodule development. Nature 417: 962-966.

Ewing B, Hillier L, WendI MC, Green P (1998a) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175-185.

Fang YW, Hirsch AM (1998) Studying early nodulin gene ENOD40 expression and induction by nodulation factor and cytokinin in transgenic alfafa. Plant Physiol. 116: 53-

68.

Fedorova M. van de Mortel J, Matsumoto PA, et al. (2002) Genome-wide identification of nodule-specific transcripts in the model legume Medicago truncatula. Plant Physiol. 130: 519-537.

Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25: 351-360.

Fichant G.A and Burks C (1991 Identifying potential tRNA genes in genomic DNA sequences. J. Mol. Biol.220: 659–671.

Fitch DHA, et al (1991) Duplication of the gamma-goblin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc. Nat'l Acad.Sci. USA 88: 7396-7400

Flavell RB. (1986) Repetitive DNA and chromosome evolution in plants. Philosophical Transactions of the royal society of London. Series B, Biological Sciences 312, No. 1154, the evolution of DNA sequences: 227-242.

Flavell AJ, Pearce SR, Kumar A (1994) Plant transposable elements and the genome. Curr. Opin. Genet.Dev. 4: 838-844.

Froy O, Gurevitz M (1998) Membrane potential modulators: a thread of scarlet from plants to humans. FASEB J 12: 1793-1796.

Gamas P, de Billy F, Truchet G. (1998) Symbiosis-Specific Expression of Two Medicago truncatula Nodulin Genes, MtN1 and MtN13, Encoding Products Homologous to Plant Defense Proteins. Mol. Plant-Microbe Interact. 11: 393-403.

Gepts P, Beavis WD, Brummer EC, et al. (2005) Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. Plant Physiol. 137:1228-1235.

Gianfranceschi L, Tarchini R, Komjanc M, Gessler C (1998) Simple sequence repeats for the genetic analysis of apple.Theoretical and Applied Genetics 96 (8): 1069-1076.

Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet 16: 45-68.

Gordon D, Desmarais C, Green P (2001) Automated finishing with autofinish Genome

Res.11: 614-625.

Gotoh, O (1982), An improved algorithm for matching biological sequences,
J. Mol. Biol., 162, 705-708.

Grusak M.A (2002) Phytochemicals in plants: genomics-assisted plant improvement for nutritional and health benefits. Curr Opin Biotechnol 13: 508-511.

Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. Genome Res. 8: 195-202.

Götz S, García-Gómez JM, Terol J, Williams TD, Nueda MJ, Robles M, Talón M, Dopazo J and Conesa A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite**.** Nucleic Acids Res. 36(10): 3420–3435.

Graham PH, Vance CP (2003) Legumes. Importance and constraints to greater use. Plant Physiol.131: 872-877.

Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. U.S.A. 94: 4168-4173.

Gresshoff PM (2003) Post-genomic insights into plant nodulation symbiosis. Genome Biology 4: 201

Gualtieri G., Kulikova O, Limpens E, et al (2002) Microsynteny between pea and Medicago truncatula in the SYM2 region. Plant Mol. Biol. 50: 225-235.

Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. Biochem Biophys Acta 561: 167-183.

Higo K, Ugawa Y, M. Iwamoto M, and Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research Vol.27 No.1 pp. 297-300

Hua A, Roe B.A. (2003) Exgap: a visualization of shotgun sequencing assembly results.

Huguet T, et al. Genetic mapping of symbiosis-related genes in Medicago truncatula. http://www.intl-pag.org/8/abstracts/pag8040.html

International Human Genome Sequencing Consortium. (2001). Initial sequencing and

analysis of the human genome. Nature 409: 860-921.

IRGSP (2005). The map-based sequence of the rice genome Nature 436: 793-800

Jackson SA, Rokhsar D, Stacey G, Shoemaker RC, Schmutz J, Grimwood J (2006) Toward a reference sequence of the soybean genome: a multiagency effort. Crop Sci 46: S55-S61.

Jacob HJ, Lindpaintner K, Lincoln SE, et al (1991) Genetic mapping of a gene causing hypertensive rat. Cell 67: 213-224.

Javie T. (2006) Whole genome assembly using paired end reads in E.coli, B.licheniformis, and S. cerevisiae. Diagnostics (Roche)

Jannick DB, Henrik N, Gunnar VH, Brunak S. (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 340:783-795, 2004.

Jones J, Field JK, Risk JM (2002) A comparative guide to gene prediction tools for the bioinformatics amateur. Inter. J Onto 20: 697-705.

Jörg D. Becker1, Leonilde M. Moreira2, Dieter Kapp3, S. Christian Frosch3, Alfred 3 and Andreas M. Perlick3

Jørgensen J-E, Stougaard J, Marcker KA (1991) A two component nodule-specific enhancer in the soybean N23 gene promoter. The Plant Cell. 3: 819-827

Kevei Z, Vinardell JM, Kiss GB, Kondorosi A, Kondorosi E (2002) Glycine-rich proteins encoded by a nodule-specific gene family are implicated in different stages of symbiotic nodule development in Medicago Spp. Molecular Plant-Microbe interactions 9: 922-931.

Korbel J.O, Urban A.E. Affourtit J.P., et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420-426.

Kubis S, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Annals of Botany 82: 45-55.

Lalli E, Ohe K, Latorre E, Bianchi ME, Sassone-Corsi P (2003) Sexy splicing: regulatory interplays governing sex determination from Drosophila to mammals. J. of

Cell Science 116: 441-445.

Larkin MA, Blackshields G, Brown NP, et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics application note 23: 2947-2948.

Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during tertiary. Syst. Biol. 54: 575-594.

Lee RC, Feinbaum RL, Ambros V (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75: 843-854.

Lerouge RP, Roche P, Faucher C, et al. (1990) Symbiotic host-specificity of Rhizobium meliloti is determined by a sulphated and acylated glucosamine oligosaccharide signal. Nature 344: 781-784.

Leung H, Hettel GP, Cantrell RP (2002) International rice research institute: roles and challenges as we enter the genomics era. Trends in Plant Science 7: 139-141.

Lee JM. And Sonnhammer LL. (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13: 875-882.

Lercher, MJ., Urrutia A.O., and Hurst L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31: 180-183.

Lin JY, Jacobus BH, SanMiguel P, et al (2005) Pericentromeric regions of soybean chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 170: 1221-1230

Liu Q, Feng Y, Zhao XA, Dong H, and Xue, Q. (2004) Synonymous codon usage bias in Oryza sativa. Plant Sci. 167: 101-105.

Lonergan, K. M., and Gray, M. W. (1993) Editing of Transfer RNAs in Acanthamoeba castellanii Mitochondria. Science 259: 812-816.

Lowe T.M, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25:955-64.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72: 595-605.

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26(4): 1107-1115.

Madar Z, Stark AH (2002) New legume sources as therapeutic agents. Br J Nutr 88: S287-S292.

Margulies M, Egholm M, Altman WE, et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Mathis R., Grosjean C, de Billy F, Huguet T, Gamas H. (1999) The Early Nodulin Gene MtN6 Is a Novel Marker for Events Preceding Infection of Medicago truncatula Roots by Sinorhizobium meliloti. Mol. Plant-Microbe Interact. 12: 544-555.

Maxam A. & Gilbert, W. (1977). A new method of sequencing DNA.Proceedings of the National Academy of Sciences, USA, 74, 560-4.

Mayer K, Schüller C, Wambutt R, et al (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature 402: 769-777.

Mayor C, Brudno M, Schwartz JR, et al (2000) Vista: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics Application Notes. 16: 1046-1047.

McClintock B. (1951) Chromosome organization and genic expression. Cold Spring Harbor Symposium on Quantitative Biology 16: 13-47.

Mergaert P, Nikovics K, Keleman Z, et al (2003) A novel family in Medicago truncatula consisting of more than 300 nodule-specific genes coding for small secreted polypeptides with conserved cysteine motifs. Plant Physiol. 132: 161-173.

Meselson M. and Stahl, F.W. (1958). "The Replication of DNA in Escherichia coli". PNAS 44: 671-82

Moreira LM, Kapp D, Frosch C, Puhler A, Perlick AM. (2001)The nodulin VfENOD18 is an ATP-binding protein in infected cells of Vicia faba L. nodules. Plant Molecular Biology 47: 749-759.

Mudge J, Huihuang Y, Denny RL, et al (2004) Soybean bacterial artificial chromosome contigs anchored with RFLPs: insights into genome duplication and gene clustering. Genome 47: 361-372.

Mudge J, Cannon SB, Kalo P, et al (2005) Highly syntenic regions in the genomes of soybean, Medicago truncatula and Arabidopsis thaliana. BMC Plant Biology 5: 15.

Nagaki K, Cheng ZK, Quyang S, et al (2004). Sequencing of a rice centromere uncovers active genes. Nature Genet. 36, 138–-145.

Niehrs C., Pollet N. (1999) Synexpression groups in eukaryotes. Nature 402: 483-487.

Oefner P. J., Hunicke-Smith, S. P., Chiang, L., Dietrich, F., Mulligan, J., and Davis, R. W. (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucl. Acids Res. 24(20): 3879-3886.

Ohno S. (1970). Evolution by gene duplication. Springer-Verlag SBN0-04-575015-7

Ovcharenko I., Pachter, L., Dubchak,I., Rubin, E. (2002) rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. Genome. Res. 12:832-839

Pan, H., Chissoe, S. L., Bodenteich, A., Wang, Z., Iyer, K., Clifton, S.W., Crabtree, J.S., and Roe, B. A. (1994) The complete nucleotide sequences of the SacBII Kan domain of the P1 and pAD10-SacBII cloning vector and three cosmid vectors: pTCF, svPHEP, and LAWRIST16. GATA 11 (5-6): 181-186.

Pavesi A., Conterio, F., Bolchi, A., Dieci, G., Ottonello, S. (1994) "Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions", Nucl. Acids Res., 22, 1247-1256.

Pertea M, Lin X, Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction . Nucleic Acids Res .29(5):1185-90.

Peters NK, Frost JW, Long SR (1986) A plant flavone, luteolin, induces expression of Rhizobium meliloti nodulation genes. Science 233: 977-980.

Quackenbush J, Cho J, Lee D, et al (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res

Rakesh C. Sharma and Robert T. Schimke, (1996) Preparation of Electro-competent E. coli Using Salt-free Growth Medium. Biotechniques 20: 42-44.

Ross EJ, Stone JM, Elowsky CG, Arredondo-Peter R, Klucas RV, Sarath G.RT (2004) Activation of the Oryza sativa non-symbiotic haemoglobin-2 promoter by the cytokinin-regulated transcription factor, ARR1.J Exp Bot. 55: 1721-1731.

Sachetto-Martins G., Franco LO, de Oliveira DE. (2000) Plant glycine-rich proteins: A family or just proteins with a common motif? Biochim. Biophys. Acta 1492: 1-14.

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees Mol. Biol. Evol. 4: 406-425.

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosphila genomic DNA. Genome Research 10 (4): 516-531.

Salanoubat M, Lemcke K, Rieger M, et al (2000) Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. Nature 408:820-822.

Samac DA and Graham MA (2007) Recent advances in legume-microbe interactions: Recognition, defense response, and symbiosis from a genomic perspective. Plant Physiol.144: 582-587.

Sambrook J, Fritsch EF, and Maniatis T (1989) in Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, NY, Vol. 1, 2, 3.

Sandal NN, Bojsen K, Marcker KA (1987) A small family of nodule specific genes from soybean. Nucleic Acids Res. 15: 1507-1519.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors.Proceedings of the National Academy of Sciences, USA, 74, 5463-7

Sasaki T, Matsumoto T, Yamamoto K, et al (2002) The genome sequence and structure of rice chromosome1. Nature 420: 312-316

Sato S, Nakamura E, Kaneko T, et al (2008) Genome structure of the legume *Lotus japonicus*. DNA research: 1-13.

Schauser L, Roussis A, Stiller J, Stougaard J (1999) A plant regulator controlling development of symbiotic root nodules. Nature 402:191-195.

Scheres B, van Engelen F, van der Knaap E, van de Wiel C, van Kammen A, Bisseling T, (1990) Sequential induction of nodulin gene expression in the developing pea nodul. Plant Cell 2:687-700.

Schlueter JA, Dixon P, Granger C, et al (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47: 868-876.

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

Schopfer CR, Nasrallah ME, Nasrallah JB (1999) The male determinant of self-incompatibility in Brassica. Science 286: 1697-1700.

Schröder, G, Frühling M, Pühler A, Perlick AM. (1997) The temporal and spatial transcription pattern in root nodules of Vicia faba nodulin genes encoding glycine-rich proteins. Plant Mol. Biol. 33: 113-123.

Schwartz-Sommer Z, Leclercq L, Göbel E, Saedler H (1987) Cin4, an insert altering the structure of the A1 gene in Zea mays, exhibits properties of nonviral retrotransposons. EMBO J. 13: 3837-3880.

Shiu SH, Bleecker AB (2001) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proc. Natl Acal Sci USA 98: 10763-10768.

Shoemaker RC, Polzin K, Labate J, et al (1996) Genome duplication in soybean (Glycine subenus soja). Genetics 144: 329-338.

Silverstein KAT, Graham MA, Paape TD, VandenBosch KA. (2005) Genome organization of more than 300 defensin-like genes in Arabidopsis. Plant Physiol 138: 600-610.

Smit A and Green P (1999) Repeatmasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195-197.

Spellman, PT, and Rubin GM. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J.Biol. 1:5.

Stekel D, Git Y, Falciani F (2000) The comparison of gene expression from multiple cDNA libraries. Genome Research 10: 2055-2061.

Stougaard J, Jorgensen JE, Christensen T, et al (1990) Interdependence and nodule specificity of cis-acting regulatory elements in the soybean leghemoglobin lbc3 and N23 gene promoters. Mol. Gen. Genet. 220: 353-360.

Stracke S, Kistner C, Yoshida S, et al (2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. Nature 417: 959-962.

Studier FW (1973) Analysis of bacteriophage T7 early RNAs and proteins on slab gels. J. Mol. Biol. 79:237-248.

Szittya G., Moxon S, Santos DM, et al (2008) High-throughput sequencing of Medicago truncatula short RNAs identifies eight new miRNA families. BMC Genomics 9:593-601.

Taylor JS. & Raes, J. (2004) "Duplication and Divergence: The Evolution of New Genes and Old Ideas" Annual Review of Genetics 9: 615-643.

The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.

The Gene Ontology Consortium, (2000) Gene Ontology: tool for the unification of biology. Nat. Genet. 25: 25-29.

Theologis A, Ecker JR, Palm CJ, et al (2000) Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. Nature 408: 816-820.

Thompson WF, et al (1980) Sequence organization in pea and mung bean DNA and a model for genome evolution. In Fourth John Symposium, pp31-45.

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acid Res 22: 4673-4680.

Thoquet P, Ghérardi M, Journet EP, et al (2002) The molecular genetic linkage map of the model legume Medicago truncatula: a essential tool for comparative legume genomics and the isolation of agronomically important genes. BMC Plant Biology 2: 1-

Town CD (2006) Annotating the genome of Medicago truncatula. Curr. Opin. Plant Biol 9: 122-127.

Tuskan GA, DiFazio S, Jansson S, et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596-1604.

van Kammen A (1985) Replication of plant virus RNA. Microbiol Sci. 2:170-174.

Vanin EF (1985) Processed pseudogenes, characteristic and evolution. Annu. Rev. Genet. 19: 253-272.

Vanoosthuyse V, Miege C, Dumans C, Cock M (2001) Two large Arabidopsis thaliana gene families are homologous to the Brassica gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response. Plant Mol Biol 16: 17-34.

Vasse J, de Billy F, Camut S, Truchet G (1990) Correlation between ultrastructural differentiation of bacteroids and nitrogen fixation in alfalfa nodules. J Bacteriol 172: 4295-4306.

Verma DPS, Kazazian V, Zogbi v, Bal AK (1978) Isolation and characterization of the membrane envelope enclosing the bacteroids in soybean root nodules. J.Cell Biol. 78: 919-936.

Walling JG, Shoemaker R, Young N, et al (2006) Chromosome level homeology in paleopolyploid soybean (Glycine max) revealed through integration of genetic and chromosome maps. Genetics 172: 1893-1900.

Watson JD & Crick FH (1953) A structure for deoxyribose nucleic acid. Nature 171: 737-738.

Williams EJB and Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res. 14: 1060-1067.

Williams EJ. and Hurst LD (2002) Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J. Mol. Evol. 54: 511-518.

Wright DA, Ke N, Smalle J, et al (1996) Multiple non-LTR retrotransposons in the genome of Arabidopsis thaliana. Genetics 142: 569-578.

Wu J, Yamagata H, Hayashi-Tsugane M, et al (2004). Composition and structure of the centromeric region of rice chromosome 8. Plant Cell 16, 967–-976.

Yan H, Mudge J, Kim DJ, et al (2003) Estimates of conserved microsynteny among the genomes of Glycin max, Medicago truncatula and Arabidopsis thaliana. Theor. Appl. Genet. 106:1256-1265.

Zasloff M (2002) Antimicrobial peptides of multicellular organisms. Nature 415: 389-395.

Zaug AT, Cech TR (1986) The intervening sequence RNA of Tetrahymena is an enzyme Science 231(4737):470-475.

Zhang J (2003). "Evolution by gene duplication: an update." Trends in Ecology & Evolution 18(6): 292-298.

Zhang Y, Huang YC, Zhang L, et al. (2004) Structural features of the rice chromosome 4 centromere. Nucleic Acids Res. 32, 2023–-2030.

Zhu HY, Cannon SB, Young ND, Cook DR (2002) Phylogeny and genomic organization of the TIR and non-TIR NBS-LRR resistance gene family in Medicago truncatula. MPMI 5: 529-539.33.

Zhu HY, Kim DJ, Baek JM, et al (2003) Syntenic relationships between Medicago truncatula and Arabidopsis reveal extensive divergence of genome organization. Plant Physiol. 131: 1018-1026.

.