Chapter 9

Transcriptomic Studies in Non-Model Plants: Case of *Pisum sativum* L. and *Medicago lupulina* L.

Olga A. Kulaeva, Alexey M. Afonin, Aleksandr I. Zhernakov, Igor A. Tikhonovich and Vladimir A. Zhukov

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.69057

Abstract

Transcriptomics is a dynamically developing branch of biology highly important for geneticists and molecular ecologists alike. A large number of studies concerning differential gene expression, mapping of genes and quantitative trait loci (QTL), analysis of genotyping variations and so on has been conducted recently on several non-model plants using next-generation sequencing techniques. One example of non-model legumes is garden pea (*Pisum sativum* L.), a valuable pulse crop capable of forming nitrogen-fixing nodules and arbuscular mycorrhiza. Adaptation of standardised RNA-seq approaches and data analysis developed for model plants to *P. sativum* should facilitate both studying of pea molecular genetics and breeding of new cultivars possessing agriculturally important traits. Another non-model legume is black medick *Medicago lupulina* L. (a close relative of model legume plant barrel medick, *Medicago truncatula* Gaertn.), for which unique genetic lines almost obligatory dependent on arbuscular mycorrhiza symbiosis formation have been obtained. Such lines show promise as the perfect model for studying the genetic bases of arbuscular mycorrhiza development. In this chapter, we give a brief description of the current developments in the field of garden pea and black medick transcriptomics. Our aim is to provide a quick start guide to the non-expert researchers for next-generation sequencing (NGS)-based transcriptome analysis.

Keywords: transcriptomics, RNA-seq, non-model legume plants, nitrogen-fixing symbiosis, arbuscular mycorrhiza, *Pisum sativum* L., *Medicago lupulina* L.

1. Introduction

Transcriptome is defined as the sum of all the messenger RNA molecules expressed from the genes of an organism, tissue, or a cell. Transcriptome analysis is a powerful method for

plant biology research since studying expressed genes facilitates investigation into plant development, responses to environmental stresses, plant-microbe interactions and so on. Transcriptomic analysis of model organisms, such as the classical object of plant genetics, *Arabidopsis thaliana* (L.) Heyhn., with available full-genome sequence enables researchers to conduct more precise measurements of gene expression level, including alternative splicing and epigenetic modifications studies, in order to reveal the molecular mechanisms involved in specific biological processes [1]. Undoubtedly, many aspects of plant biology, for example, economically important traits such as specific immunity, pathogen resistance and symbiotic efficiency contributing to high crop productivity, cannot be studied with the use of model plants only, making the investigation of non-model plants a necessity.

The rapid decrease of per-base sequencing cost coupled with unprecedented development rates of computational biology practices opened the field of transcriptomics for in-depth investigation of non-model plants [1]. In the last few years, a large number of studies concerning differential gene expression, mapping of genes and quantitative trait loci (QTLs), analysis of genotyping variations and so on using next-generation sequencing (NGS) techniques has been conducted on several non-model plants including legumes (members of family Fabaceae) [2–4].

The leguminous plants (chickpea (*Cicer arietinum* L.), pea (*Pisum sativum* L.) and lentil (*Lens culinaris* Medik.)) were among the earliest domesticated plant species [5] and are to this day an integral part of agricultural systems [6]. These and other members of the Fabaceae family are essential for economics as a food, fodder and oil source [3]. A significant feature of most legume species is their capability of forming mutualistic symbioses with soil microorganisms. Root-nodule symbiosis, the association of the legumes with nodule bacteria collectively called rhizobia, provides the plant with fixed atmospheric nitrogen [7]. This fact makes the legume-rhizobial inter-organismal system an essential component of natural and agricultural ecosystems [8]. Arbuscular-mycorrhizal (AM) symbiosis (association with arbuscular mycorrhizal fungi), inherent to over 80% of land plants including most of legumes [9], facilitates water and mineral (especially phosphorous) uptake of the plant and consequently the nutritional value of the crop. Legumes are also capable of forming symbioses with endophytic plant growth promoting bacteria also contributing to plant productivity [10, 11].

In the early 1990s, two legume species—*Medicago truncatula* Gaertn. and *Lotus japonicus* (Regel.) K. Larsen—were introduced as model objects for studying plant genetics of symbiotic nitrogen fixation and AM development [12–14]. Both species have small diploid genomes (approx. 500 Mb) [15] and are self-pollinators with short generation time able to produce hundreds to thousands of seeds per plant. Intensive studies of genetics resulted in high-quality annotated genomes for both *L. japonicus* and *M. truncatula*, accumulation of gene expression microarray datasets and development of several tools and repositories combining the diverse genetic, genomic and transcriptomic data in these model species (the *Medicago* Gene Expression Atlas [16, 17], the *Medicago* genome database [18], the *Lotus* Base information portal [19], etc.).

During the last decade, rapid development of sequencing and bioinformatics technologies significantly improved the state of genomics in non-model legumes. In the past few years, genomes of important legumes, such as *Glycine max* (L.) Merr. [20], *Phaseolus vulgaris* L. and *Trifolium*

pratense L. [21], were sequenced and are currently available at Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html) and in the integrative bioinformatic platform Legume IP providing information about gene and protein sequences, gene models and annotations, syntenic regions, protein families and phylogenetic trees [22].

Despite all the recent research progress, most of the agriculturally important legumes were considered 'orphan' crops for a long time as separated from the intense genomic studies due to large genomes, and their agricultural significance mainly in developing countries lacking funds for large-scale 'omics' studies [3]. Most genome and transcriptome analysis tools were developed for particular model objects [23] and can generally be used for studying 'orphan' species [24, 25], although careful fine-tuning may be necessary for successful deployment of said tools in non-model organisms (see **Figure 1**). With the cost of genome assemblies remaining prohibitively high, researchers are forced to work with only transcriptome data, making the analysis strategy all the more important.

It is worth noting that one of the most challenging steps of transcriptome analysis pipelines is correct transcript annotation. The simplest approach giving a sufficiently accurate result is BLAST search against annotated sequences of other species. The development of transcriptome annotation pipelines, for example, Trinotate [26], has more or less taken the burden of transcriptome

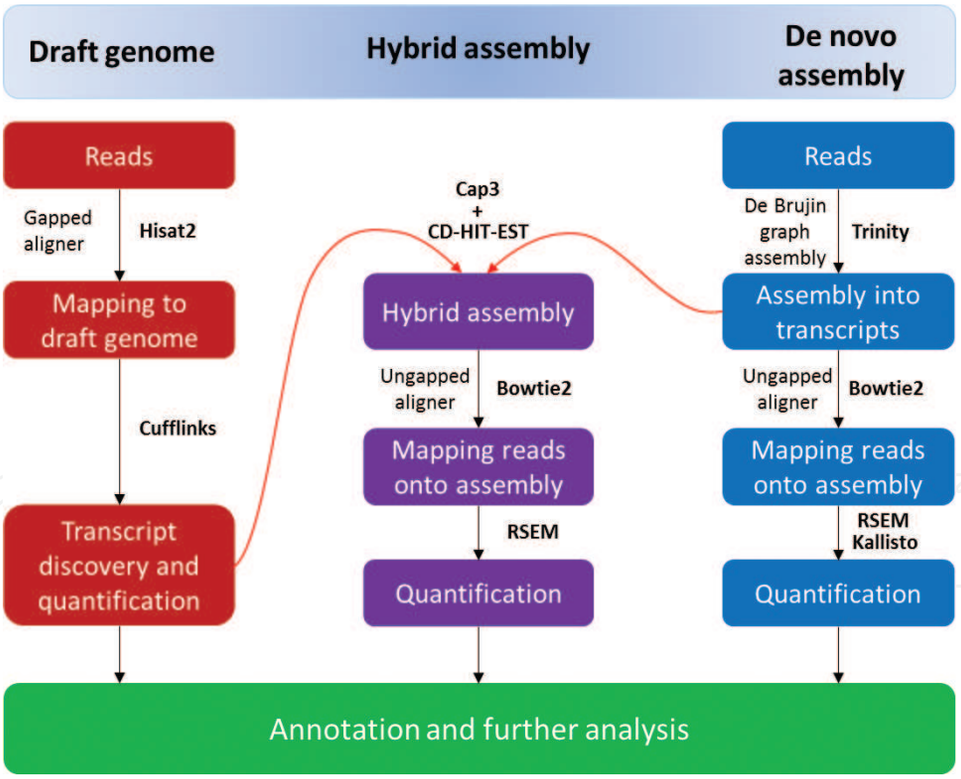

Figure 1. Pipelines of transcriptome assembly in non-model plants (based on the information from Refs. [23, 24].) Three strategies for RNA-seq analysis. (A) Using a draft genome. Novel transcript discovery, quantification and functional annotation. (B) De novo transcriptome assembly with no reference. For quantification, reads are mapped back to the novel reference transcriptome followed by the functional annotation of the novel transcripts as in (A). (C) Combination of the two methods. Transcriptomes are first assembled using methods (A) and (B) then merged using CD-HIT-EST and cap3. Transcripts are then annotated as in (B).

annotation off of the researcher. Trinotate combines the output of a number of annotation tools into an integrated database simplifying the following deeper analysis of acquired data.

One example of an 'orphan' legume is garden pea (*Pisum sativum* L.), a valuable pulse crop capable of forming both nitrogen-fixing symbiosis and arbuscular mycorrhiza. Global production of green pea in 2014 was 17.4 million tons, harvested from 2.3 million hectares, with an additional 11.2 million tons of dried pea from 6.9 million hectares [6]. The genome of the species is considered to be about 4300 Mb with high percentage of repetitive sequences [27]. Adaptation of RNA-seq data analysis approaches standardised for model plants to *P. sativum* should facilitate both studying of pea molecular genetics and breeding of new cultivars possessing agriculturally important traits.

Black medick (*Medicago lupulina* L.), a close relative of a model legume plant barrel medick (*M. truncatula* Gaertn.), is another example of an important (but almost not studied in terms of genetics) non-model legume. It is valuable as a pasture legume component in complex grass mixtures and can also be used as an intermediate culture in crop rotation and as green manure. Black medick is characterised by high protein, vitamin and mineral content, long growing season and ability for improving soil fertility due to nitrogen fixation, therefore being a perfect lawn plant [28]. Black medick is a very promising object for studying AM functioning and development, since a unique genetic line of *M. lupulina* obligatory dependent on arbuscular mycorrhiza symbiosis formation has been selected from the spring landrace population VIK-32 of *M. lupulina* var. *vulgaris* Koch originating from Kazakhstan [28, 29]. Plants of the line MlS-1 (for *Medicago lupulina* Spring) [28] demonstrate dwarfism when grown in the soil with low Pi (inorganic phosphorus) level in the absence of the AM fungi inoculation but can grow normally when inoculated with AM fungus. Therefore, MlS-1 line is considered highly effective in AM symbiosis formation (as inoculation by fungi dramatically heightens the plant biomass). Apparently, MlS-1 line is only capable of using the symbiotrophic way of phosphorus uptake from the soil, supposedly due to yet unidentified mutation(s) and, consequently, can serve as a model object for the investigation of arbuscular-mycorrhizal symbiosis. For instance, this line is suitable for mutagenesis aimed at selection of mutants with defects in arbuscular mycorrhiza development, since plants carrying mutations in genes related to AM formation can be easily identified by visual examination as demonstrating dwarfism under inoculation with AM fungi [29].

High level of genome synteny, similarity of gene sequences and developmental processes provide the opportunity to use the vast amounts of data accumulated on *M. truncatula* in genetics, genomic and transcriptomics of these non-model legumes *M. lupulina* and *P. sativum*. In this chapter, we give a brief description of the current achievements in the field of transcriptomics of non-model legumes black medick (*M. lupulina*) and garden pea (*P. sativum*).

2. Transcriptome assembly studies

2.1. P. sativum transcriptomics

The genome of *P. sativum* is as of yet not assembled due to its comparatively large size and numerous repeats, greatly reducing the number of research methods available. Pea transcriptome,

unlike genome, is closer in size to transcriptomes of other legumes, including model plant *M. truncatula*, making it more susceptible to analysis. Due to the existence of tissue-specific gene expression, different plant tissues possess unique sets of transcripts, making the choice of tissue samples important for further research. Furthermore, transcriptome assemblies from distinct plant organs should be used as reference for analysis of tissue-specific processes. A high-quality transcriptome assembly with full tissue representation is therefore crucial for studies associated with gene interactions (differential gene expression, see section 3), gene polymorphism studies and proteome analysis.

In the last 5 years, several pea transcriptome assemblies of distinct organs and tissues were presented by different workgroups. The first publication of pea transcriptome sequencing and assembly was made by Franssen et al. [30]. Total of 20 libraries from flowers, leaves, cotyledons, epicotyls and hypocotyls and etiolated and light-treated etiolated seedlings were sequenced using the Roche 454 sequencing platform. Several iterations of de novo assembly and merging yielded 81,449 unigenes. Sudheesh et al. [31] sequenced transcriptomes from different parts (leaf, stipule, stem, tendril tissues from multiple nodes, root-tip tissues, flowers, stamens, pistils, immature pods, immature seeds and nodules) of two pea cultivars (Parafield and Kaspa) differing in both seed and plant morphological characteristics. Read assembly for separate cultivars yielded 126,335 and 145,730 contigs, respectively, with 87% showing significant expression levels in both cultivars. Later on, Liu et al. sequenced samples from pea seeds harvested at the stage of 10 and 25 days after pollination and assembled 77,273 unigenes [32].

Several transcriptome assembly sets were generated for Single Nucleotide Polymorphism (SNP) marker development and genetic mapping in pea (see section 4). Duarte et al. [33] sequenced libraries from eight pea cultivars (six spring sown, one winter sown field pea, one fodder pea cultivar) with Roche 454 technology. A total of 3,826,797 reads were assembled into 68,850 contigs by MIRA transcriptome assembler [34]. Sindhu et al. sequenced 3′-anchored libraries of eight diverse pea accessions (six *P. sativum* cultivars (CDC Bronco, Alfetta, Cooper, CDC Striker, Nitouche and Orb) and two wild accessions P651 (*P. fulvum*), PI 358610 (*P. sativum* ssp. *abyssinicum*)) with Roche 454 technology, generating 4,008,648 reads in total. De novo assembly was performed for 520,797 reads from the CDC Bronco by MIRA, resulting in a set of 29,725 reference contigs representing a significant proportion of the 3′ end of genes in pea [35].

Since analysis of inter organismal genetic network between pea and rhizobia is a poorly developed field, assembly of a high-quality transcriptome provided researchers with the much-needed data on nodule-specific transcripts. Transcriptomes of pea nodules and root tips were obtained by Zhukov et al. [36]. Transcriptome sequencing using the Illumina Genome Analyzer IIx platform (Illumina Inc.) generated 52,021,865 reads from the 'Nodules' library and 17,684,604 reads from the 'Root Tips' library, yielding 58,397 and 37,287 contigs assembled de novo by Trinity, respectively [37]. A total of 13,000 nodule-specific contigs were annotated by alignment to known plant protein-coding sequences and by Gene Ontology search. Of these, 581 sequences were found to possess full Coding DNA Sequence (CDSs) and could thus be considered novel nodule-specific transcripts of pea. Further investigation of those transcripts can potentially lead to the discovery of key regulators of nodule symbiosis, such as identification of pea gene homologous to *Nodulation signaling pathway 1 (NSP1*) gene of *M. truncatula* [38]. In this study, pea gene *Sym34* was shown to be homologous to the *M. truncatula NSP1* gene,

based on preliminary stop codons detected in an open reading frame of *NSP1* homologous sequence in two *sym34* allelic mutants (RisNod1 and RisNod23) and full co-segregation of the alleles of the hypothetical pea *Nsp1* gene with the nodulation phenotype in F_2 generation.

Alves-Carvalho et al. [39] sequenced transcriptomes of roots, nodules, shoots, leaves, flowers, seeds, tendrils and pods harvested at different developmental stages of pea cultivar 'Caméor'. Sequencing of 20 cDNA libraries produced one billion reads. After de novo assembly and several steps of redundancy reduction, 46,099 contigs were obtained. The main objective of their study was to obtain the most complete transcriptome and to filter out all the artefacts and chimeric contigs so a rigorous filtration pipeline was developed and implemented. The accumulated transcriptome data was used for the development of the Pea RNA-Seq gene atlas containing expression profiles of thousands of genes in several organs, including symbiotic nodules. It is worth noting that the pipeline used in this work filtered out a large proportion of short protein-coding transcripts, including a number of NCR peptide-coding transcripts [40], making the Pea RNA-Seq gene atlas less useful than tissue-specific transcriptomes in some cases.

Pea RNA-Seq gene atlas is also lacking information regarding mycorrhiza-specific transcripts. Genetic framework of mycorrhizal symbiosis is as of yet not fully understood in either model or non-model legumes [38]. In order to discover symbiotically active genes both in plant roots and arbuscular-mycorrhizal fungus, a transcriptome of Frisson pea cultivar roots colonised by *Rhizophagus irregularis* isolate BEG144 was assembled by our workgroup. Sequencing was performed on an Illumina HiSeq2000 sequencing platform yielding 120 million pair end reads. In order to separate the transcriptomes of two organisms present in the samples, all the reads were mapped using the HISAT2 mapper [41] to the genome of *R. irregularis* [42]. Over 5 million successfully mapped reads were assembled by Trinity with default parameters yielding 30,000 transcripts, in good correlation with 28,000 of known genes for the fungus [42, 43].

All the transcripts not mapped to the *R. irregularis* genome were then assembled with the Trinity pipeline with standard assembly parameters and quality trimming parameters. This resulted in more than 200,000 contigs, of which more than 100,000 were similar to genes of pea and other plants of the Fabaceae family.

An assessment of transcriptome assembly and annotation completeness with single-copy orthologs for all available pea transcriptomes was carried out using BUSCO V.2 software with OrthoDB v9.1 'embryophyta' base as a reference [44]. The lowest number of present groups in the transcriptome published by Franssen et al. [30] named 'Franssen' is due to low transcriptome coverage. High number of missing groups in 'Kaspa', 'Parafield' and 'SGE' assemblies are most likely the result of limited tissue representation (see **Figure 2**). Deep sequencing of mycorrhized roots yielded similar results in regard to transcriptome completeness as a combined transcriptome from 20 tissues, indicative of assembly of low-copy transcripts due to high transcriptome coverage.

2.2. *M. lupulina* transcriptomics

M. lupulina is a plant of the Fabaceae family, a close relative to the *M. truncatula*, for which a unique genetic line MlS-1 characterised by obligate mycotrophic lifestyle was obtained [28].

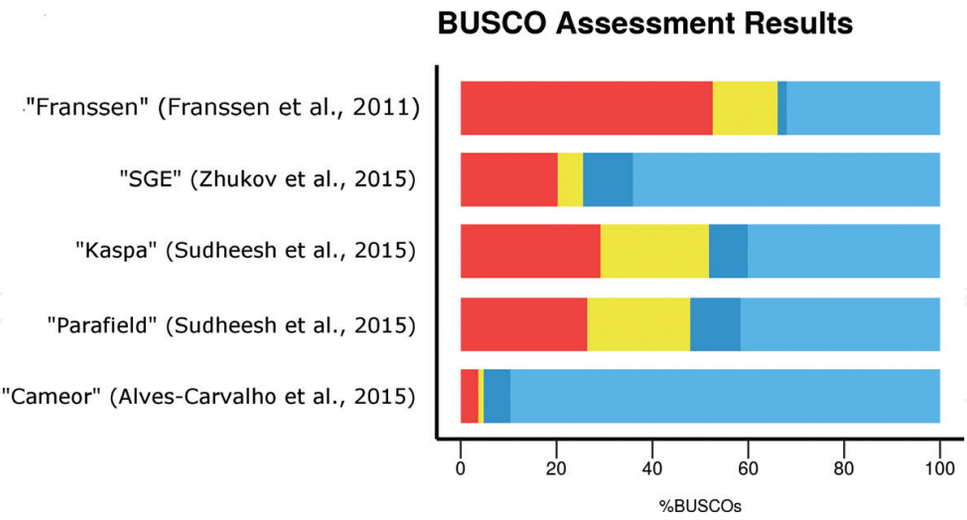

Figure 2. The results of BUSCO analysis of pea transcriptomes. Light-blue: complete and single-copy genes; dark-blue: complete and duplicated genes; yellow: fragmented genes; red: missing genes.

This line may potentially be extremely useful as a model for investigation of genetic foundations of mycorrhizal symbiosis. *M. lupulina* is a novel object for genomic studies, so to kick-start its analysis the transcriptome of the mycorrhized roots of *M. lupulina* was sequenced using the Illumina 2500 platform. Plants of MlS-1 line were grown in soil under inoculation with *R. irregularis* strain RCAM00320, followed by total RNA extraction from the mycorrhized root system and appropriate preparation of cDNA libraries for Illumina sequencing. Using Trinity assembly pipeline, 41 million paired reads were assembled yielding over 138,000 contigs, of which 19,022 showed resemblance to genes of *R. irregularis*. Further analysis revealed over 70,000 contigs similar to known genes of *M. truncatula*. The assembled transcriptome can be used as reference for differential gene expression analysis.

3. Differential gene expression (DGE)

Analysis of alterations in gene expression between conditions or genotypes is the most significant part of transcriptomic data analysis. The differences in expression levels can help determine the important genes and elucidate the processes taking place in the investigated samples.

Extensive analysis of gene expression can be carried out by microarray analysis or RNA sequencing technology. Microarray technology requires prior knowledge of gene sequences and is more suitable for objects with available genome sequence. In the case of model object *M. truncatula,* combination of microarray data resulted in development of atlas of gene expression profiles (*Medicago truncatula* Gene Expression Atlas (MtGEA)) (https://mtgea.noble.org/v3/). MtGEA contains information about gene expression in roots, nodules, stems, petioles, leaves, vegetative buds, flowers, seeds, pods and is potentially helpful for studying other legumes. Despite the fact that pea genome is not sequenced yet, several studies of pea gene expression have been carried

out by microarray technology. Analysis of gene expression during *Mycosphaerella pinodes* infection was carried out using a microarray [45] containing 16,470 different 70-mer oligonucleotides from *M. truncatula* and only 25 did not show a detectable signal [46]. In another study, microarray transcriptome profiling based on known pea Expressed Sequence Tags (ESTs) revealed altered expression of genes associated with programmed cell death, oxidative stress and protein ubiquitylation during seed aging [47].

In spite of many advantages of microarrays, this technique is not effective for quantification of transcript splice variants and, furthermore, cannot provide information about novel genes not included in the array. The development of NGS technology made analysis of full transcriptome gene expression possible. To date, there were several studies of pea gene expression based on RNA-seq technology. Comparative analysis of transcriptional control of pea seed development conducted by RNA-seq revealed significant differences in gene expression between vegetable and grain pea. Genes associated with sugar and starch biosynthesis were significantly activated during seed maturation. Analysis of differential expression of these genes revealed a negative correlation between soluble sugar and starch flux in vegetable and grain pea seeds [32]. Alves-Carvalho et al. [39] developed the Pea RNA-Seq gene atlas containing expression profiles of thousands of genes in different pea tissues harvested at distinct developmental stages [48].

Although RNA-seq technology is indispensable for exhaustive transcriptome studies, it is not the most cost-efficient tool for gene expression analysis due to substantial sequencing depth required for rare transcript detection. There are RNA-seq modifications, for example, Massive Analysis of cDNA Ends (MACE) developed by GenXPro GmbH (Frankfurt am Main, Germany) (http://genxpro.net/) that increase the sequencing depth (number of reads per-transcript) by sequencing only a 50–500 bp fragment (adjacent to the 5′ or 3′-end of the transcript, dependent on the version) [49]. As each read originates from a distinct copy of mRNA, MACE technology is free of duplications and similar artefacts, leading to much more accurate transcript quantification. Even though MACE data cannot be used to distinguish expression of splice-variants of genes, it can be successfully applied in a number of scenarios even with species not possessing a high-quality transcriptome.

In our opinion, 5′MACE is a technology possessing potential for simultaneous analysis of gene expression in prokaryotic and eukaryotic organisms; therefore, this technology is practically tailor-made for the analysis of plant-microbe interaction, particularly for studying the process of root nodule development in the plants of the Fabaceae family.

One of the many challenges in analysing the onset of nodule symbiosis is the small amounts of tissue available. Enclosed environments of symbiotic compartments complicate direct measurements. Implementation of 5′MACE technology made it possible to analyse the gene expression patterns of both organisms simultaneously in a developing nodule and at a fraction of the cost of a full RNA-seq study.

In our group, 5′MACE was implemented in a study investigating the expression changes in pea nodules caused by a mutation in the *Sym31* gene with unknown function. This gene is responsible for the unique fix$^-$ mutant phenotype (non-nitrogen-fixing nodules) with halted bacteroid development [50]. Two plant genotypes Sprint-2Fix$^-$ (carrying a mutation in the

Sym31 gene) and parental wild-type line Sprint-2 were inoculated with an efficient *Rhizobium leguminosarum* bv. *viciae* RCAM1026 [51]. All the obtained reads were sequentially mapped to the RCAM1026 genome (about 8% mapped reads), then to the pea transcriptome assembly from Alves-Carvalho et al. [39] (about 60% mapped reads) resulting in two sets of differential transcriptome data. The transcript quantification was carried out using the edgeR package [52]. Differentially expressed genes were then visualised on a metabolic map using KOBAS 2.0 annotation server [53]. Analysis resulted in the discovery of a coordinated shift in sulphur metabolism in both organisms. These preliminary data show the great potential of the 5′MACE technology in furthering our understanding of inter-organismal gene regulatory networks in plant-microbe interactions.

4. Transcript-based markers and their usage

The application of NGS for massive genetic polymorphism discovery is widely used due to being much more labour and time efficient than previously used methods such as microarray hybridisation [54] or denaturing high-performance liquid chromatography (HPLC) [55]. Originally, the main challenge in using NGS methods for massive polymorphism screening was obtaining sequences of a particular genomic locus for multiple lines due to complexity of plant genomes and the relatively low productivity of the first-generation NGS-sequencing platforms, leading to the development of several methods for sequencing optimisation.

For example, Restriction site Associated DNA-sequencing method (RAD-Seq) consists of genome cleavage and selection of fragments of appropriate size flanked by specific restriction sites (as with RFLP and AFLP analyses) [56]. RAD-Seq yields fragments distributed randomly over a genome and is suitable for discovering indels (insertion-deletion polymorphisms), SNVs (single nucleotide variations) and microsatellites simple sequence repeats (SSR). Using RAD-Seq, Boutet et al. [57] discovered a total of 419,024 SNVs between at least two of the four pea lines analysed in their work. Pea genetic map constructed by genotyping a subset of 64,754 SNVs on a subpopulation of 48 RILs (recombinant inbred lines) was collinear with previous pea consensus maps and therefore with the *M. truncatula* genome. Yang et al. [58] using Illumina HiSeq 2500 platform uncovered 8899 putative SSR-containing sequences. Reliable amplifications of detectable polymorphic fragments among 24 genotypes of pea were obtained for about a half of randomly selected SSR, 820 in total.

Another way of data complexity reduction is transcriptome sequencing. It makes the discovery of polymorphic sites in open reading frames (ORFs) and 5′- and 3′-untranslated regions (UTR) of a gene possible. Moreover, polymorphic sites associated with individual genes may have special meaning for evolutionary studies and QTL analyses. Even though the transcriptome sequencing omits introns and intergenic regions, it can successfully be used for SSR site detection.

Several polymorphism-screening studies aimed on SNVs and SSR sites discovering in transcriptomic data were performed on pea (see **Table 1**). SNVs detection may be executed by mapping NGS reads to an existing reference transcriptome assembly [59] or by de novo assembly of those reads [33, 35, 60]. In the case of existing assembly, the additional data complexity

Year	Plant material	Platform, technique	Number of putative discovered SNVs	Number of putative discovered SSR-sites	Number of created and mapped markers	References
2013	Parafield, Yarrum, Kaspa, 96–286	454 Roche, GS-FLX	36,188	2932	705	Leonforte et al. [60]
2014	*Six spring sown*: Lumina, Hardy, Panache, Rocket, Kayanne, Terese *One winter sown*: Cherokee *One fodder*: Champagne	Roche 454, GS-FLX	35,455	2397	1340	Duarte et al. [33]
2014	*Pisum sativum*: CDC Bronco, Alfetta, Cooper, CDC Striker, Nitouche, Orb. *P. fulvum*: P651 *P. sativum* ssp. *abyssinicum*: PI 358610	Roche 454, Titanium	over 20,000	406	1536	Sindhu et al. [35]
2017	SGE = JI3023 Finale = JI2678 Frisson = JI2491 NGB1238 = JI0073 Sparkle = JI0427 Sprint-2 = JI2612	Illumina HiSeq 2000, MACE	34,711	-	-	Zhernakov et al. [59]

Table 1. Studies aimed at gene polymorphism detection in pea (*Pisum sativum* L.) using transcriptome NGS-sequencing.

reduction is achievable by limiting sequenced mRNA regions. Since UTRs are generally more polymorphic than ORFs using sequences from the 3′ and 5′ mRNA, ends in SNV analysis should yield comparable results to those obtained with RNA-seq. 3′MACE protocol for cDNA-libraries preparation was used by Zhernakov et al. [59] to discover SNVs distinguishing six pea lines. Mapping MACE reads to the reference nodule transcriptome assembly of the pea line SGE [36] resulted in characterisation of over 34,000 polymorphic sites in more than 9700 contigs. Several of these SNVs were located within recognition sites of restriction endonucleases which allowed the design of co-dominant Cleaved Amplified Polymorphic Sequences (CAPS) markers for the particular transcript.

SNVs are markers of choice now due to their abundance and the availability of high-throughput screening techniques. SNV genotyping systems are now available, varying in the number of samples and markers to be genotyped, such as GoldenGate® and Infinium from Illumina Inc., SNPStream from Beckman Coulter and GeneChip from Affymetrix [61]. Illumina GoldenGate® oligonucleotide pool assay (OPA) designed for transcriptome-discovered SNVs was used for pea salinity tolerance QTLs search [60].

As the pea genome is not sequenced yet, the genetic linkage maps are still relevant, since determination of loci responsible for target traits requires their fine mapping and subsequent

search for candidate genes in the already sequenced genome of the model legume plant *M. truncatula*. Transcriptome-discovered SNVs and high-throughput genotyping systems made the construction of several highly saturated genetic maps of pea possible (see **Table 1**) [33, 35, 60].

5. Conclusion

Next-generation sequencing techniques make the analysis of differential gene expression and molecular marker development by transcriptome sequencing possible even in species lacking genomic information. Further development of sequencing and bioinformatics should substantially promote the investigation into genetics of non-model plants. It is worth noting that numerous traits like effectiveness of symbioses development [62] or specific resistance to pathogens can only be studied in each particular cultivated plant species, most having limited genomic data available. In addition, the decline in biodiversity makes the investigation of unique secondary metabolites inherent to non-model medicinal plants a pressing matter.

Leguminous plants capable of improving the soil quality due to the formation of the mutualistic symbioses with nodule bacteria and arbuscular mycorrhizal fungi are an integral part of agricultural systems. The genetics of most crop legumes lags behind that of model plants, and some are even considered 'orphan' crops, separated from the intense genomic studies due to a number of factors. Fortunately, the similarity of genome organisation, or 'genome synteny', characteristic for most related species, can help 'translate' the genomic data from the model legumes to their pulse crop relatives [63].

Using RNA-seq technologies for de novo transcriptome assembly provides opportunities for finding novel genes and isoforms in non-model species and investigation of their differential expression. Comparison to genomes and transcriptomes of closely related species can help determine the level of evolutionary distance between the two species and discover possible evolutionary pressures shaping contemporary species. Technologies for determining gene expression levels using transcript ends (like 3′ and 5′ MACE) can be used to conduct large-scale gene expression studies on a smaller budget. 5′ MACE, a technology for simultaneous analysis of prokaryotic and eukaryotic transcript abundancies, is particularly useful for studying plant-bacteria interactions. Using transcriptome-sequencing data in genetic marker development streamlines the construction of high-quality genomic maps, crucial for routine gene identification tasks as well as potentially for refining genome assemblies for non-model organisms. All the methods are useful in investigation of the unique phenotypes not present in the model plants, for example, *M. lupulina* MlS-1 genetic line, uniquely dependent on the AM formation. Adaptation of standardised RNA-seq approaches and data analysis developed for model plants to an important crop culture *P. sativum* should facilitate the breeding of new cultivars that meet the requirements of the present-day agriculture and possess the complex of beneficial traits, including increased efficiency of interactions with nodule bacteria and arbuscular-mycorrhizal fungi.

Acknowledgements

The work was supported by Russian Foundation for Basic Research (grant # 16-34-60132 for O.A. Kulaeva), by grant of the President of the Russian Federation (project NSh-6759.2016.4 for A.M. Afonin), by Russian Science Foundation (grant # 14-24-00135 for I.A. Tikhonovich and grant # 16-16-00118 for A.I. Zhernakov and V.A. Zhukov). The authors thank Dr A.P. Yurkov (ARRIAM, St. Petersburg, Russia) for providing the *Medicago lupulina* L. MlS-1 line and A.S. Sulima (ARRIAM, St. Petersburg, Russia) for critical reading of the manuscript.

Author details

Olga A. Kulaeva[1], Alexey M. Afonin[1], Aleksandr I. Zhernakov[1], Igor A. Tikhonovich[1,2] and Vladimir A. Zhukov[1]*

*Address all correspondence to: zhukoff01@yahoo.com

1 All-Russia Research Institute for Agricultural Microbiology, Saint Petersburg, Russia

2 Saint Petersburg State University, Saint Petersburg, Russia

References

[1] Dong Z, Chen Y. Transcriptomics: Advances and approaches. Science China Life Sciences. 2013;**56**:960-967. DOI: 10.1007/s11427-013-4557-2

[2] Kumawat G, Gupta S, Ratnaparkhe MB, Maranna S, Satpute GK. QTLomics in soybean: A way forward for translational genomics and breeding. Frontiers in Plant Science. 2016;**7**:1852. DOI: 10.3389/fpls.2016.01852

[3] Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR. Orphan legume crops enter the genomics era! Current Opinion in Plant Biology. 2009;**12**:202-210. DOI: 10.1016/j.pbi.2008.12.004

[4] Xiao M, Zhang Y, Chen X, Lee EJ, Barber CJ, Chakrabarty R, et al. Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. Journal of Biotechnology. 2013;**166**:122-134. DOI: 10.1016/j.jbiotec.2013.04.004

[5] Zohary D, Hopf M, Weiss E. Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin. Oxford University Press Inc., New York; 2012

[6] Food and Agriculture Organization of the United Nations (FAOSTAT) [Internet]. 2017. Available from: http://www.fao.org/faostat/en/

[7] Sprent JI, Nodulation in legumes. Royal Botanic Gardens. Kew: Royal Botanic Gardens; 2001.

[8] Provorov N, Tikhonovich I. Genetic resources for improving nitrogen fixation in legume-rhizobia symbiosis. Genetic Resources and Crop Evolution. 2003;**50**:89-99. DOI: 10.1023/A:1022957429160

[9] Smith SE, Read DJ. Mycorrhizal symbiosis. Academic press; 2008

[10] Elvira-Recuenco M, Van Vuurde J. Natural incidence of endophytic bacteria in pea cultivars under field conditions. Canadian Journal of Microbiology. 2000;**46**:1036-1041

[11] Mishra PK, Mishra S, Selvakumar G, Bisht J, Kundu S, Gupta HS. Coinoculation of Bacillus thuringeinsis-KR1 with Rhizobium leguminosarum enhances plant growth and nodulation of pea (*Pisum sativum* L.) and lentil (*Lens culinaris* L.). World Journal of Microbiology and Biotechnology. 2009;**25**:753-761. DOI: 10.1186/1471-2164-8-427

[12] Barker DG, Bianchi S, Blondon F, Dattée Y, Duc G, Essad S, et al. *Medicago truncatula*, a model plant for studying the molecular genetics of theRhizobium-legume symbiosis. Plant Molecular Biology Reporter. 1990;**8**:40-49

[13] Cook DR. *Medicago truncatula*—a model in the making!: Commentary. Current Opinion in Plant Biology. 1999;**2**:301-304

[14] Stougaard J. Genetics and genomics of root symbiosis. Current Opinion in Plant Biology. 2001;**4**:328-335

[15] Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, et al. Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus.* Plant Physiology. 2005;**137**:1174-1181. DOI: 10.1104/pp.104.057034

[16] Benedito VA, Torres-Jerez I, Murray JD, Andriankaja A, Allen S, Kakar K, et al. A gene expression atlas of the model legume *Medicago truncatula.* The Plant Journal. 2008;**55**:504-513. DOI: 10.1111/j.1365-313X.2008.03519.x

[17] He J, Benedito VA, Wang M, Murray JD, Zhao PX, Tang Y, et al. The *Medicago truncatula* gene expression atlas web server. BMC Bioinformatics. 2009;**10**:441. DOI: 10.1186/1471-2105-10-441

[18] Krishnakumar V, Kim M, Rosen BD, Karamycheva S, Bidwell SL, Tang H, et al. MTGD: The *Medicago truncatula* genome database. Plant and Cell Physiology. 2014;**56**:pcu179. DOI: 10.1093/pcp/pcu179

[19] Mun T, Bachmann A, Gupta V, Stougaard J, Andersen SU. Lotus base: An integrated information portal for the model legume *Lotus japonicus.* Scientific Reports. 2016;**6**:39447. DOI: 10.1038/srep39447

[20] Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;**463**:178-183. DOI: 10.1038/nature08670

[21] De Vega JJ, Ayling S, Hegarty M, Kudrna D, Goicoechea JL, Ergon Å, et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Scientific Reports. 2015;**5**:17394. DOI: 10.1038/srep17394

[22] Li J, Dai X, Zhuang Z, Zhao PX. LegumeIP 2.0—a platform for the study of gene function and genome evolution in legumes. Nucleic Acids Research. 2016;**44**:D1189-D1194. DOI: 10.1093/nar/gkv1237

[23] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, Mcpherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;**17**:13. DOI: 10.1186/s13059-016-0881-8

[24] Marchant A, Mougel F, Almeida C, Jacquin-Joly E, Costa J, Harry M. De novo transcriptome assembly for a non-model species, the blood-sucking bug *Triatoma brasiliensis*, a vector of Chagas disease. Genetica. 2015;**143**:225-239. DOI: 10.1007/s10709-014-9790-5

[25] Garg R, Jain M. RNA-Seq for Transcriptome Analysis in Non-model Plants. In: Rose RJ, editor. *Legume Genomics: Methods and Protocols.* Totowa, NJ: Humana Press; 2013. pp. 43-58

[26] Trinotate: Transcriptome Functional Annotation and Analysis [Internet]. Available from: https://trinotate.github.io/

[27] Macas J, Neumann P, Navrátilová A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula.* BMC Genomics. 2007;**8**:427. DOI: 10.1186/1471-2164-8-427

[28] Yurkov A, Jacobi L, Gapeeva N, Stepanova G, Shishova M. Development of arbuscular mycorrhiza in highly responsive and mycotrophic host plant-black medick (*Medicago lupulina* L.). Russian Journal of Developmental Biology. 2015;**46**:263-275. DOI: 10.1134/S1062360415050082

[29] Yurkov AP, Jacobi LM. Selection of mycorrhizal mutants in black medic (*Medicago lupulina*) [in Russian]. Natural and Technical Sciences. 2011;**6**:127-134

[30] Franssen SU, Shrestha RP, Bräutigam A, Bornberg-Bauer E, Weber AP. Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. BMC Genomics. 2011;**12**:227. DOI: 10.1186/1471-2164-12-227

[31] Sudheesh S, Sawbridge TI, Cogan NO, Kennedy P, Forster JW, Kaur S. De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. BMC Genomics. 2015;**16**:611. DOI: 10.1186/s12864-015-1815-7

[32] Liu N, Zhang G, Xu S, Mao W, Hu Q, Gong Y. Comparative transcriptomic analyses of vegetable and grain pea (*Pisum sativum* L.) seed development. Frontiers in Plant Science. 2015;**6**:1039. DOI: 10.3389/fpls.2015.01039

[33] Duarte J, Rivière N, Baranger A, Aubert G, Burstin J, Cornet L, et al. Transcriptome sequencing for high throughput SNP development and genetic mapping in pea. BMC Genomics. 2014;**15**:126. DOI: 10.1186/1471-2164-15-126

[34] MIRA—Sequence Assembler and Sequence Mapping for Whole Genome Shotgun and EST/RNASeq Sequencing Data [Internet]. Available from: https://sourceforge.net/projects/mira-assembler/

[35] Sindhu A, Ramsay L, Sanderson LA, Stonehouse R, Li R, Condie J, et al. Gene-based SNP discovery and genetic mapping in pea. Theoretical and Applied Genetics. 2014;**127**:2225-2241. DOI: 10.1007/s00122-014-2375-y

[36] Zhukov VA, Zhernakov AI, Kulaeva OA, Ershov NI, Borisov AY, Tikhonovich IA. De novo assembly of the pea (*Pisum sativum* L.) nodule transcriptome. International Journal of Genomics. 2015;**2015**:695947. DOI: 10.1155/2015/695947

[37] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology. 2011;**29**:644. DOI: 10.1038/nbt.1883

[38] Shtark O, Kumari S, Singh R, Sulima A, Akhtemova G, Zhukov V, et al. Advances and prospects for development of multi-component microbial inoculant for legumes. Legume Perspectives. 2015;**8**:40-44. DOI: 10.13140/RG.2.1.1634.0247

[39] Alves-Carvalho S, Aubert G, Carrère S, Cruaud C, Brochot AL, Jacquin F, et al. Full-length de novo assembly of RNA-seq data in pea (Pisum sativum L.) provides a gene expression atlas and gives insights into root nodulation in this species. The Plant Journal. 2015;**84**:1-19. DOI: 10.1186/s13059-016-0881-8

[40] Tikhonovich IA, Kliukova MS, Kulaeva OA, Zhernakov AI, Zhukov VA. The Process of Bacteroid Differentiation in Pea (*Pisum sativum* L.) is Controlled by Symbiotic Genes that Regulate the Expression of the NCR Gene Family. In: Book of Abstracts 12th European Nitrogen Fixation Conference; 25-28 August 2016; Budapest, Hungary; 2016. p. 232

[41] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols. 2016;**11**:1650-1667. DOI: 10.1038/nprot.2016.095

[42] Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, et al. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proceedings of the National Academy of Sciences. 2013;**110**:20117-20122. DOI: 10.1073/pnas.1313452110

[43] INRA GlomusDB, the Glomus Intraradices Genome Database Version 2.0. [Internet]. Available from: http://mycor.nancy.inra.fr/IMGC/GlomusGenome/index3.html

[44] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;**31**:3210-3212. DOI: 10.1093/bioinformatics/btv351

[45] Küster H, Hohnjec N, Krajinski F, El Yahyaoui F, Manthey K, Gouzy J, et al. Construction and validation of cDNA-based Mt6k-RIT macro- and microarrays to explore root endosymbioses in the model legume Medicago truncatula. Journal of Biotechnology. 2004;**108**:95-113. DOI: 10.1016/j.jbiotec.2003.11.011

[46] Fondevilla S, Küster H, Krajinski F, Cubero JI, Rubiales D. Identification of genes differentially expressed in a resistant reaction to *Mycosphaerella pinodes* in pea using microarray technology. BMC Genomics. 2011;**12**:28. DOI: 10.1186/1471-2164-12-28

[47] Chen H, Osuna D, Colville L, Lorenzo O, Graeber K, Kuester H, et al. Transcriptome-wide mapping of pea seed ageing reveals a pivotal role for genes related to oxidative stress and programmed cell death. PLoS One. 2013;**8**:e78471. DOI: 10.1371/journal.pone.0078471

[48] The Pea RNA-Seq Gene Atlas [Internet]. 2015. Available from: http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi

[49] Zawada AM, Rogacev KS, Müller S, Rotter B, Winter P, Fliser D, et al. Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. Epigenetics. 2014;**9**:161-172. DOI: 10.4161/epi.26931

[50] Borisov AY, Rozov S, Tsyganov V, Morzhina E, Lebsky V, Tikhonovich I. Sequential functioning of *Sym-13* and *Sym-31*, two genes affecting symbiosome development in root nodules of pea (*Pisum sativum* L.). Molecular and General Genetics MGG. 1997;**254**:592-598

[51] Afonin A, Sulima A, Zhernakov A, Zhukov V. Draft genome of the strain RCAM1026 *Rhizobium leguminosarum* bv. *viciae.* Genomics Data. 2016;**11**:85-86. DOI: 10.1016/j.gdata.2016.12.003

[52] Robinson MD, Mccarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;**26**:139-140. DOI: 10.1093/bioinformatics/btp616

[53] Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research. 2011;**39**:W316-W322. DOI: 10.1093/nar/gkr483

[54] Beaudet L, Bédard J, Breton B, Mercuri RJ, Budarf ML. Homogeneous assays for single-nucleotide polymorphism typing using AlphaScreen. Genome Research. 2001;**11**:600-608. DOI: 10.1101/gr.1725501

[55] Xiao W, Oefner PJ. Denaturing high-performance liquid chromatography: A review. Human Mutation. 2001;**17**:439-474. DOI: 10.1002/humu.1130

[56] Davey JW, Blaxter ML. RAD-Seq: Next-generation population genetics. Briefings in Functional Genomics. 2010;**9**:416-423. DOI: 10.1093/bfgp/elq031

[57] Boutet G, Carvalho SA, Falque M, Peterlongo P, Lhuillier E, Bouchez O, et al. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics. 2016;**17**:121. DOI: 10.1186/s12864-016-2447-2

[58] Yang T, Fang L, Zhang X, Hu J, Bao S, Hao J, et al. High-throughput development of SSR markers from pea (*Pisum sativum* L.) based on next generation sequencing of a purified chinese commercial variety. PLoS One. 2015;**10**:e0139775. DOI: 10.1371/journal.pone.0139775

[59] Zhernakov A, Rotter B, Winter P, Borisov A, Tikhonovich I, Zhukov V. Massive analysis of cDNA ends (MACE) for transcript-based marker design in pea (*Pisum sativum* L.). Genomics Data. 2017;**11**:75-76. DOI: 10.1016/j.gdata.2016.12.004

[60] Leonforte A, Sudheesh S, Cogan NO, Salisbury PA, Nicolas ME, Materne M, et al. SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum* L.). BMC Plant Biology. 2013;**13**:161

[61] Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu C, Lejeune-Hénaut I, et al. Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. BMC Genomics. 2010;**11**:468. DOI: 10.1186/1471-2164-11-468

[62] Shtark OY, Borisov AY, Zhukov VA, Tikhonovich IA. Mutually beneficial legume symbioses with soil microbes and their potential for plant production. Symbiosis. 2012;**58**:51-62. DOI: 10.1007/s13199-013-0226-2

[63] Young ND, Udvardi M. Translating *Medicago truncatula* genomics to crop legumes. Current Opinion in Plant Biology. 2009;**12**:193-201. DOI: 10.1016/j.pbi.2008.11.005

IntechOpen